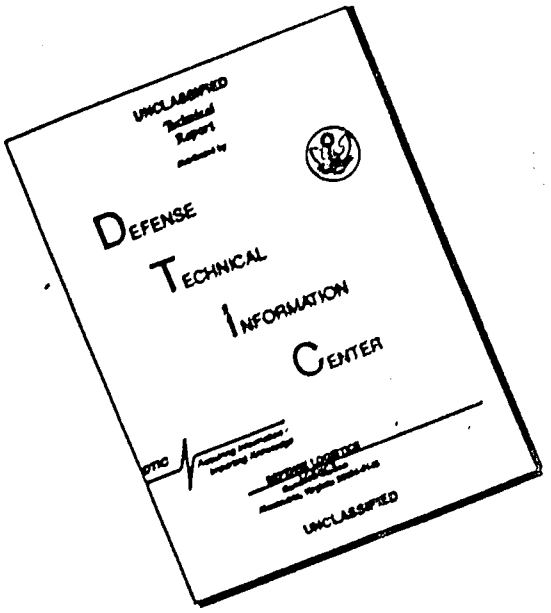
ADA-130703

BEST AVAILABLE COPY

# DISCLAIMER NOTICE




THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

# COMPONENT PART NOTICE

THIS PAPER IS A COMPONENT PART OF THE FOLLOWING COMPILATION REPORT:

(TITLE): Proceedings of the Annual Conference of the Military Testing Association (23rd) held at Arlington, Virginia 25-30 October 1981. Volume 2.

(SOURCE): Army Research Inst. for the Behavioral and Social Sciences, Alexandria, VA

TO ORDER THE COMPLETE COMPILATION REPORT USE AD-A130 703.

THE COMPONENT PART IS PROVIDED HERE TO ALLOW USERS ACCESS TO INDIVIDUALLY AUTHORED SECTIONS OF PROCEEDINGS, ANNALS, SYMPOSIA, ETC. HOWEVER, THE COMPONENT SHOULD BE CONSIDERED WITHIN THE CONTEXT OF THE OVERALL COMPILATION REPORT AND NOT AS A STAND-ALONE TECHNICAL REPORT.

THE FOLLOWING COMPONENT PART NUMBERS COMPRISE THE COMPILATION REPORT:

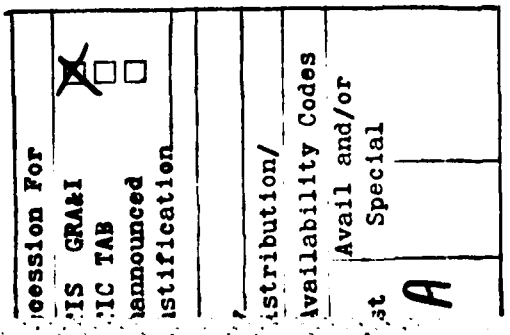
| AD#: | TITLE: |
|---|---|
| AD-P001 357 | CODAP (Comprehensive Occupational Data Analysis Programs). Analytical Tool for Position Classification? |
| AD-P001 358 | An Empirical Investigation of Firts Term Attrition. |
| AD-P001 359 | The Use of Computer Graphics for Practical Problem Solving. |
| AD-P001 360 | Biographical Correlates of a Cognitive Abilities Test for Firefighter Selection. |
| AD-P001 361 | The Measurement of Army Battalion Performance. |
| AD-P001 362 | The Commander's Unit Analysis Profile (CUAP). |
| AD-P001 363 | CODAP (Comprehensive Occupational Data Analysis Programs): Some New Techniques to Improve Job Type Idenfification and Definition. |
| AD-P001 364 | Effects of Psychological Differentiation and Cognitive Consistence on Students' Course Evaluations. |
| AD-P001 365 | Predictors of Success in Basic Enlisted Submarine School. |
| AD-P001 366 | Aviation Training Task Proficiency: A Probabilistic Approach. |
| AD-P001 367 | Mobilization Planning & Control: A Model for the Selective Service System. |
| AD-P001 368 | Human Aptitude Ability Assessment Techniques for System Designers. |
| AD-P001 369 | Race Ihfluence on Peer Ratings in ROTC Training Platoons. |
| AD-P001 370 | The Navy Personnel Accessioning System. |
| AD-P001 371 | Human Factors Evaluation of Division Air Defense Gun Systems. |
| AD-P001 372 | Identifying Common Duties among Naval Skilled Trades. |
| AD-P001 373 | A Performance Management System for the U.S. Coast Guard: Strategy for Development and Implementation. |
| AD-P001 374 | Pilot Research for Validation of ASVAB and Enlistment Standards against Performance on the Job. |

## COMPONENT PART NOTICE (CON'T)

| AD#: | TITLE: |
|---|---|
| AD-P001 375 | An Examination of the Group Differences Aspect of the Construct Validity of the Organizational Assessment Package. |
| AD-P001 376 | Attrition Casualty, Explanation, and Level of Analysis. |
| AD-P001 377 | An Evaluation of the Air Force Airman Retraining Program. |
| AD-P001 378 | A Rationale for Designing and Managing Technical Training Programs. |
| AD-P001 379 | Computer Assisted Task Selection in US Army SQT (Skill Qualification Test) Development. |
| AD-P001 380 | Interrater Reliability: The Development of an Automated Analysis Tool. |
| AD-P001 381 | An Innovative Approach to Data Capture in Automated Assessment. |
| AD-P001 382 | Simulation of Command Group Operations: An Evaluation Report. |
| AD-P001 383 | Shortening of Defense Language Aptitude Battery. |
| AD-P001 384 | Models of Human Information Processing: Implications for Trainer/Simulator Design. |
| AD-P001 385 | CHAPARRAL Training Subsystem Effectiveness Analysis (TSEA). |
| AD-P001 386 | The Personality Research Form (PRF) as a Prediction for Success in Pilot Training. |
| AD-P001 387 | Self-Assessment in Personnel Selection and Placement. |
| AD-P001 388 | An MTA (Military Testing Association) Publishing Challenge: A Textbook Concept and Prospects. |
| AD-P001 389 | The Officer Candidate Preparatory School: An Intense Program that Doubles as a Personnel Assessment Center. |
| AD-P001 390 | The Development and Application of Measures of Occupational Learning Difficulty. |
| AD-P001 391 | Validation of RCMP (Royal Canadian Mounted Police) Selection Procedures. |
| AD-P001 392 | A Systematic Approach to Training Program Evaluation. |
| AD-P001 393 | Training Evaluation: A Fragmatic Overview. |
| AD-P001 394 | Why and How to Insure that Items in the Same Pool are Appropriately Credentialed: Data Collection Strategies and Analytical Methods for Item Linking. |
| AD-P001 395 | Test Development Issues for Adaptive Testing. |
| AD-P001 396 | Foundations for the Mathematical Notion of Information in Item Response Theory and Robust Ability Estimation. |
| AD-P001 397 | Validity Considerations for Adaptive Testing Systems. |
| AD-P001 398 | Individual Test Patterns: Are We Correcting for Guessing in the Wrong Direction? |
| AD-P001 399 | Legal and Political Considerations in Large-Scale Adaptive Testing. |
| AD-P001 400 | Supervisor Ratings as Criteria for Skill Qualification Tests. |
| AD-P001 401 | Success and Failure in Skill Qualification Testing: Troop Views. |
| AD-P001 402 | Skill Qualification Test Feedback: Timeliness Matters. |

COMPONENT PART NOTICE (Cont'd)

| AD#: | TITLE: |
|---|---|
| AD-P001 403 | Predicting Skill Qualification Test Item Difficulty from Judgments. |
| AD-P001 404 | Measuring and Improving the Readability of Military Documents. |
| AD-P001 405 | Computer Aids for Authoring Tests. |
| AD-P001 406 | A Literacy Task Inventory for Identifying Literacy Skill Levels of Jobs. |
| AD-P001 407 | The Air Force Job Performance Appraisal System. |
| AD-P001 408 | Development of the Job Reading Test (JRT). |
| AD-P001 409 | A Theory and Model of Item Readability. |
| AD-P001 410 | Soldier Reading Ability: The Advocacy Point of View. |
| AD-P001 411 | What are the Literacy Components of Job Proficiency? An Objective. Opinionated Commentary. |
| AD-P001 412 | Job-Related Measurement of Reading Ability. |
| AD-P001 413 | Military Compensation: An Overview of the Cash Pay Structure. Challenges for the Future. |
| AD-P001 414 | The Need for a Theory of Military Compensation. |
| AD-P001 415 | Nonpecuniary Rewards and Military Compensation. |
| AD-P001 416 | The Victorian Legacy: A Social Historical Analysis of Attitudes Toward Women in the Canadian Forces. |
| AD-P001 417 | Social Change and the Participation of Women in the American Military. |
| AD-P001 418 | Supervisors' Attitudes Toward Women and the Performance Appraisals given to Men and Women in the Canadian Forces. |
| AD-P001 419 | Psychological Screening for Weapons Use Suitability: A Formal Decision Model. |
| AD-P001 420 | Psychological Screening for Weapons Use: A Clinical Validation of Measures. |
| AD-P001 421 | Assessing the Impact of the Army's Organizational Effectiveness (OE) Program Model, Methodology, and Illustrative Cases. |
| AD-P001 422 | Problems in the Measurement of Major Change in Specialized Organization. |
| AD-P001 423 | Machine Instruction and Tests. |
| AD-P001 424 | Criterion-Referencing of Performance by Latent-Trait Scaling. |
| AD-P001 425 | Issues in Designing and Validating Alternative Performance Indicators. |
| AD-P001 426 | Subpopulation Analyses of Current Youth Aptitudes. |
| AD-P001 427 | Aptitude Testing in DoD and the Profile of American Youth Study. |
| AD-P001 428 | Military and Civilian Test Score Trends (1950 - 1980). |
| AD-P001 429 | Student and Course Evaluation at the Combined Arms and Services Staff School. |
| AD-P001 430 | Using the Systems Approach to Develop the GAS3 (Combined Arms and Services Staff School) Curriculum. |



COMPONENT PART NOTICE (Cont'd)

| AD#: | TITLE: |
|---|---|
| AD-P001 431 | Creating a New Curriculum to Train Army Staff Officers. |
| AD-P001 432 | The Combined Arms and Services Staff School from the Perspective of an Author and Instructor. |
| AD-P001 433 | Kalman Filtering Techniques Applied to Assignment Systems. |
| AD-P001 434 | Decision Aids for Personnel Actions |

# PROCEEDINGS

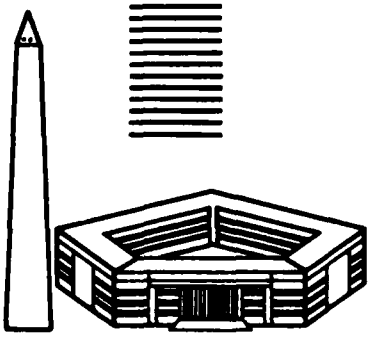
## 23rd Annual Conference of the

# MILITARY TESTING ASSOCIATION

**Volume II**

**Coordinated by the**

**U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES**



**ARLINGTON, VIRGINIA 25-30 OCTOBER 1981**

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

83 05

PROCEEDINGS

23RD ANNUAL CONFERENCE

of the

MILITARY TESTING ASSOCIATION

coordinated by

US ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

ALEXANDRIA, VIRGINIA

VOLUME II

Pentagon City - Quality Inn

Arlington, Virginia

25 - 30 October 1981

DISTRIBUTION STATEMENT
Approved for public release;
Distribution Unlimited

TABLE OF CONTENTS


| | Page |
|---|---|
| 1981 MTA Conference | 1 |
| Program | 2 |
| Contributed Papers | 10 |
| Author Index | 11 |
| Keynote Address | 15 |
| Paper Presentations | 71 |
| Panel Discussions | 1287 |
| MTA Steering Committee | 1692 |
| Minutes of the 1981 Steering Committee Meeting | 1693 |
| By-Laws | 1697 |
| Names and Addresses of Registrants | 1703 |



| Accession For | |
|---|---|
| NTIS GRA&I | ☒ |
| DTIC TAB | ☐ |
| Unannounced | ☐ |
| Justification | |
| By Per Ltr. on File | |
| Distribution/ | |
| Availability Codes | |
| Dist | Avail and/or Special |
| A | |

DTIC COPY INSPECTED

# AUTHOR INDEX


| | Paper Presentations | Page |
|---|---|---|
| VOL I - | Adams, Jerome and Hicks, Jack M. | 73 |
| | Adams, Jerome; Richards, John; and Fullerton, Terry | 85 |
| | Affourtit, Thomas D. | 95 |
| | Ansbro, T.M. and Hayes, W.A. | 103 |
| | Ansbro, T.M. and Hayes, W.A. | 115 |
| | Ballentine, Rodger D. and Cunningham, J.W. | 125 |
| | Beaton, Albert E. and Barone, John L. | 135 |
| | Beel, C. Derek | 149 |
| | Berkowitz, Melissa | 159 |
| | Birke, Werner | 171 |
| | Bobko, Douglas J. | 181 |
| | Braun, Frank B. | 191 |
| | Braun, Henry I. and Jones, Douglas H. | 201 |
| | Bridges, Claude F. | 211 |
| | Brittain, Clay V. and Brittain, Mary M. | 225 |
| | Butler, Richard P.; Bridges, Claude F.; and Houston, John W. | 239 |
| | Butler, Walter G. and Cooper, Ronald B. | 249 |
| | Camm, William B. | 259 |
| | Cicchinelli, Louis F. and Harmon, Kenneth R. | 269 |
| | Claeys, Willem; De Boeck, Paul; and Bohrer, Arnold | 289 |
| | Cormier, Stephen M. | 299 |
| | Croll, Paul R. | 309 |
| | Diane, Cynthia C. | 321 |
| | Dickinson, Richard W. | 329 |
| | Dodd, B.T. and Nelson, HMS | 341 |
| | Dohme, John A. | 345 |
| | Doorley, Richard D. | 355 |
| | Dow, Andrew N. | 367 |
| | Driskill, Walter E.; Keeth, James B.; and Gentner, Frank C. | 373 |
| | Dyer, Frederick; Schroeder, James E.; Czerny, Paul; Gillotti, Daniel P.; and Youngling, Edward W. | 379 |
| | Ekwall, Ralph W. | 385 |
| | Elig, Timothy W.; Gade, Paul A.; and Eaton, Newell Kent | 395 |
| | Finstuen, Kenn and Weaver, Charles N. | 409 |
| | Freda, Jon S.; Hall, Eugene, R.; and Ford, Larry H. | 419 |
| | Frederickson, Edward W. | 429 |
| | Frey, Robert L. | 439 |
| | Geddie, James C. | 449 |
| | Gilbert, Arthur C.F. and Tulloh, Nancy E. | 459 |
| | Gilbert, Arthur C.F. and Tulloh, Nancy E. | 471 |
| | Goodgame, Doug | 485 |
| | Gorman, Steven | 493 |
| | Gould, R. Bruce | 503 |
| | Greenup, H. William and Levin, Steven | 511 |
| | Hagman, Joseph D. | 523 |
| | Hall, Eugene R. and Freda, Jon S. | 535 |
| | Heaton, C.L. | 547 |
| | Hicks, Jack M. | 557 |

Hirsh, Hannah Rothstein 563
Hobson, Edward N. 573
Horton, D.S. 581
Hosmer, Clark L. 595
Johnson, James H; McGinley, Kathleen; Byrnes, Elizabeth; Godin, Steve W.; and Bloomquist, Michael 603
Kass, Richard A; Weltin, Mary; Seeley, Leonard; and Wing, Hilda 613
Katznelson, Judah 629
Kemery Edward; Pleban, Robert; Williams, Beverly; and Dyer, Frederick 635
Kenaston, Alicia 643
Kerner, Susan E. and O'Mara, Francis E. 653
Kimmel, Melvin J. and O'Mara, Francis E. 669
Knox, Allen N. 679
Koch, Christopher G., Englert, Judith A; Vestewig, Richard E.; and Larson, John T. 687
Lessard, Pierre 701
Lidderdale, I.G. 709
Lipscomb, Suzanne M. 719
Loo, Robert 739
Lyons, Thomas J. 751
Mann, Walter 761
Mann, Walter G. 769
Mathews, John J. and Roach, Bennie W. 777
Mays, Pamela V. and Dyer, Frederick 789
McKenzie, Robert C. 801
Meissner, Adelheid and Puzicha, Klaus J. 821
Miles, John L., Jr. 833
Mitchell, Jimmy L.; Keeth, James B.; and Wiekhorst, Linda A. 839
Modrick, John A.; Plocher, T.A.: Hutcheson, J.D.; and Chambers, R.M. 853
VOL II - Newton, Jean 861
Nogami, Glenda Y. 871
Oelrich, Fritz and Bennett, Bruce 885
Olson, Darlene M. 899
O'Mara, Francis E. 913
Palmer, R.L. 929
Phalen, William J. 939
Pigeon, E. Richard 957
Potter, Earl H., III 971
Rankin, William C. and McDaniel, William C. 981
Roll, Charles Robert, Jr. and Berger, B. Michael 993
Rossmeissl, Paul G.; Kostyla, Stanley J.; and Baker, James D. 1001
Rumsey, Michael G. 1013
Sands, W.A. 1025
Sarli, Gary G. and Carter, Richard J. 1031
Sharon, Amiel T. 1039
Sharp, Bradford P. and Allen, Nicholas H. 1049
Shields, Joyce; Hanser, Lawrence; Williams, Edward; and Popelka, Beverly 1063
Short, Lawrence O. 1089
Siebold, Guy L. 1099
Skinner, Mary J. 1109
Smith, Brandon B. 1121
Smith, Robert M. 1131
Staley, Michael R. and Weissmuller, Johnny J. 1141
Taylor, John F. 1151

Taylor, John F. and Begland, Robert R. 1163
Thain, John W. 1173
Thomason, Spencer C. 1179
Tubbs, John D.; Deason, Paul J.; Everett, James E.; and Hansen, Alan D. 1189
Vandyke, G.A. 1203
van Rijn, Paul 1215
Waldkoetter, Raymond O. 1227
Washbush, John B. 1235
Weeks, Joseph L. 1249
Wevrick, L. and Hung, C.K. 1259
Witmer, Bob G. and Kristiansen, D.M. 1269
Yard, Gilbert F. 1279

Panel Discussions

Bejar, Isaac I., (Chair); Dorans, Neil J.; and Dwyer, Carol A. 1289

Bejar, Isaac I., (Chair); Jones, Douglas H.; Rock, Donald A., Wainer, Howard; Waters, Brian; and Lee, Gus C. 1313

Bolin, Stanley F., (Chair); Eastman, Robert; Sova, James; Harman, Joan; Kessler, John; and Macpherson, Douglas 1355

Duncan, Eric R., (Chair); Kincaid, Peter J.; Braby, Richard, Wulfeck, Wallace H., II; Hooke, Lydia R; Sticht, Thomas G.; Thompson, Nancy A.; Cowan, Douglas K.; and Guerrieri, John A. 1391

Duncan, Eric R., (Chair); Claudy, John J.; Fischl, M.A.; Kern, Richard P.; Wisher, Robert A.; and Payne, Sandra A. 1421

Hale, Linda Pappas, (Chair); Hale, Thomas; and Oglobin, Peter 1469

Kinzer, Nora Scott, (Chair); Forestell, Diane G.; Segal, David R.; Segal, Mady Wechsler; Simpson, Suzanne P.; and Edwards, Henry 1497

Krug, Samuel E., (Chair); Behrens, Gary; Palese, Robert; Williams, John; and Winn, Frank 1529

Oliver, Laurel W., (Chair); James, U.S.; McCorcle, M.D.; and Mietus, J.R. 1551

O'Neil, Harold F., (Chair); Baker, Eva L.; Choppin, Bruce; and Quellmalz, Edys S. 1573

Sellman, Wayne S. (Chair); Eitelberg, Mark J.; Laurence, Janice; and Waters, Brian K. 1617

Stillman, Jon, (Chair); Anderson, Mike; DeFrain, Dennis; and Ekwall, Ralph 1647

Ward, Joe H., Jr., (Chair); Dumas, Neil S.; and Kroeker, Leonard P. 1687

AD P001357

# CODAP: ANALYTICAL TOOL FOR POSITION CLASSIFICATION?

Jean Newton, Occupational Specialist
Office of Personnel Management
Standards Development Center
Washington, D. C. 20415

## Summary

The U. S. Office of Personnel Management (OPM) experimented to determine the economy and feasibility of using Comprehensive Occupational Data Analysis Programs (CODAP) in developing position classification standards for civilian occupations. Classification standards are structured using a nine-factor evaluation methodology. They must achieve equal pay for substantially equal work, not only within the occupation, but across occupational and organizational lines.

The experimental study included a nationwide survey of about 2000 positions in the GS-203, Personnel Clerical and Assistance Series, with CODAP clustering analysis to identify subgroups which require separate treatment in occupational analysis. Through this method, we identified duty profiles for various work situations and evaluated the magnitude of the subgroups in recommending a restructuring of subgroups within the occupation. More traditional methods of factfinding such as personal observations and on-site interviews were required to explain reasons why one kind of work is more difficult than another in arriving at grades for the occupation.

## INTRODUCTION

Our question was: Can we use computers and automated job analysis in developing classification standards for Federal civilian occupations?

Some time ago, a task force at the Office of Personnel Management (OPM) recommended streamlining the collection and analysis of occupational data in OPM studies--using multipurpose questionnaires and computer processing. From a theoretical standpoint, the Standards Development Center (SDC) staff had reservations about the potential value of automated job analysis for SDC occupational studies. However, in the interests of keeping up with the State-of-the-Art and simplifying our occupational analysis and documentation procedures, we decided to experiment to evaluate the possibilities more objectively.

We selected CODAP for the experiment because it has been used successfully by other Governmental offices in occupational analysis for training programs and test development, and the computer programs were already developed. The experimentation was planned in conjunction with our ongoing schedule of occupational studies with production being given foremost consideration.

## BACKGROUND

### Position Classification Standards

Position classification standards are published by the Standards Development Center and applied by agency classification specialists throughout the U. S. in determining the appropriate titles, series, and pay grades for civilian positions. These standards must achieve "equal pay for substantially equal work" across occupational lines and agency lines, as required by law (Chapter 51, Title 5, U. S. C.)

In developing classification standards, we need to do several kinds of analysis:

--Series Analysis: What are the boundaries of the occupation? Which jobs should be included; which should be excluded?

--Level Analysis: What characteristics distinguish between levels of work? What makes one kind of work more difficult than another?

--FES Analysis: How should the nine factors of the Factor Evaluation System be described? Distinguishing characteristics identified in the level analysis must be considered, and subdivided or splintered, as appropriate, into levels of the nine FES evaluation factors. How this information is distributed under the various factors can affect grade levels of jobs because of the FES weighting scheme, e.g., Factor 1, Knowledge Required by the Position, has an average weight of 40% of the total classification of a job, while Factor 4, Complexity, has an average weight of 10%

Considerable judgment goes into tailoring factfinding to meet the needs of a standards project. We research background information in our project files or libraries; interview people in the occupation; analyze training programs and guidelines governing the work; review work samples; observe work in progress; conduct panels of subject-matter experts, etc.

During the factfinding, the occupational specialist traditionally examines duties and tasks as "clues," continually comparing and probing until s/he personally arrives at logical reasons for classification differences and similarities in various kinds and levels of work. A serious disadvantage in this methodology is that key information is embedded in the mind of the occupational specialist. Another specialist cannot usually assume a study in progress without conducting additional factfinding, thus prolonging the project.

### The Occupation Studied

We selected the Personnel Clerical and Assistance Series, GS-203, for this study because it was on our priority projects listing; the work was understandable to people who would review our reports of the experiment; and tasks could be listed readily in a questionnaire.

Positions in this occupation have, as a common base, knowledge of some aspect of civilian personnel rules and procedures. The 1966 classification standard authorized twelve different titles:

| | |
|---|---|
| Personnel Clerk | Personnel Assistant |
| Classification Clerk | Classification Assistant |
| Employee Development Clerk | Employee Development Assistant |
| Empoloyee Relations Clerk | Employee Relations Assistant |
| Labor Relations Clerk | Labor Relations Assistant |
| Staffing Clerk | Staffing Assistant |

Sixty Federal agencies have positions in different kinds of work settings such as civilian personnel offices, administrative offices, examining offices, training centers, and Federal job information centers. The total occupation has about 10,000 positions.

Clerical titles overlap with assistant titles at grade GS-6, and assistant titles overlap with trainee specialist titles at grades GS-5 and GS-7. The Central Personnel Data File (CPDF) gives the number of jobs at each grade, but not the number of clerks and assistants or the number of jobs in each specialization. So we felt that this occupation was a good candidate for CODAP analysis, particularly the clustering of jobs according to the tasks performed.

## CODAP SURVEY PROCEDURES

The CODAP survey included developing the occupational questionnaire and answer sheet, identifying the survey sample, administering survey forms, and processing the data collected.

### The Questionnaire

Some information for the questionnaire was obtained from position descriptions, adjudicated appeals, training courses, regulations, and procedural guides. To obtain the kind of detail needed for task statements, we also interviewed employees and supervisors at their work sites.

In the hope of being able to identify levels of difficulty and specialized kinds of work in analyzing CODAP reports, we organized the 538 task statements in a special way:

--Twenty duty headings covered work requiring specialized personnel knowledges. To the extent possible, task statements under these headings were written to bring out different levels of work. For example, instead of writing a task statement such as "Compose job announcements," we used two statements:

1. "Compose job announcements for recurring vacancies," and
2. "Compose job announcements for one-of-a-kind vacancies."

In this example, the first task is easier because recurring vacancies tend to be for clerical or low grade jobs which have readily understandable kinds of work; previous announcements can be modified with a few changes. The second task requires developing new material usually for higher grade jobs which have more difficult kinds of work and more complicated kinds of qualification requirements.

--Six duty/function headings were added to round out the occupational data however, they do not involve specialized personnel knowledges, i.e., records maintenance, meeting arrangements, coding, and typing. Where possible tasks under these headings were written more broadly to keep the questionnaire as short as possible, e.g., "Assign serial numbers to items such as applications, announcements, PD's, employee suggestions, etc.

A complicating factor in organizing the tasks was that CODAP processing limits the number of duty headings to twenty-six. Because of the diversity in kinds of work, we had to synthesize tasks of unrelated jobs under some of the headings, e.g., the heading "Federal Pay Explanations and Procedures" ranges from explaining when pay days are (which could be performed in a variety of jobs) to participating in wage surveys (usually a function of the classification and wage office).

Employees were asked to rate their tasks on a relative time-spent scale, and supervisors were asked to rate tasks supervised on a relative-difficulty scale.

The Survey Sample

Our initial survey sample had about 2700 employees and 350 supervisors:

--Ten agencies identified approximately one-third of their employees and 15% of their supervisors on the basis of representativeness. Most of these were scheduled for group administration of the survey forms.

--A 10% random sample from the Central Personnel Data File provided positions in an additional 41 agencies to which we mailed the survey forms.

One large agency declined to participate because of the workload required.

The good returned sample had about 1900 questionnaires. The primary reasons for the reduced sample was a high rate of "no shows" for OPM-sponsored group sessions and failure to return mailed forms.

Administration

Fifty-seven group sessions were conducted nationwide. In 21 of these, agency representatives volunteered to conduct group sessions for their own employees rather than have them travel away from installations.

A potentially serious problem arose when a large military command advised that approval must be obtained from a union headquarters which had nationwide recognition before they could participate. Apparently, the union viewed the questionnaire as a vehicle for going directly to employees, thus circumventing union participation. There were also union concerns about:

--The adequacy of the task statements; and
--The possibility that if employees did not fill out the forms properly, their grades would be adversely affected.

In resolving the issue, our Labor Relations Officer explained that the project leader would personally be conducting on-site interviews to supplement and verify data obtained in the survey.

## CODAP Processing

OPM's computer is not adapted to do CODAP clustering. We are grateful to the Navy Occupational Data Analysis Center for coming to our rescue, scanning the answer sheets and producing CODAP reports using their standard programs.

In identifying clustered groups to analyze, we considered the percentage of commonality within the group and the number of jobs. 40% commonality is generally considered to be significant. For classification, we need to describe discrete work situations so that one is readily discernible from another. Therefore, to the extent that sufficient jobs were provided, we used a 50% or higher degree of commonality within the group as our criterion in selecting 11 groups for further analysis.

Despite our care in writing and organizing task statements, positions within the groups were at several grade levels. In standard studies, we do not assume that because a kind of work is now performed at a current grade level that that is the correct grade level. However, in order to analyze levels of work in the CODAP groups, it was necessary to obtain CODAP job descriptions by current grades.

# ANALYSES FOR GS-203 STANDARDS

Decisions in the occupational analysis involved a blending of knowledges gained from a variety of sources. During the interim between administering the survey and developing the standards, we interviewed an additional 100 employees and their respective supervisors.

CODAP methodology was most helpful in determining the specializations and coverage of the standard. Information gained in interviews and on-site observations assisted in explaining or confirming data in CODAP reports and furnished insights for the more detailed level analyses and FES factor analysis.

## Series Analyses

In the process of determining the coverage of the standards and specializations, we looked at CODAP duty profiles by current classifications and then by the CODAP clusters:

--Current Classifications We developed a bird's eye view of the occupation using CODAP job descriptions for current titles. By selecting cut-off points for total members performing and percentage of time spent, we developed summary duty profiles which give information such as:

- o Profiles of major duties in each specialization,
- o The number of jobs in each specialization and their relationship to the total occupation.
- o The degree of nonspecialized work performed in relation to specialized work.
- o Work which is common to two or more titles, e.g., health and life insurance procedures.

--CODAP Clustered Groups

- The largest clustered groups were clerks effecting official personnel actions and maintaining master personnel records. We knew the work existed, of course, but not its magnitude in relation to other kinds of clerical work. We recommended a new specialization called, "Personnel Processing Clerk."

  To develop information for the boundaries of the Personnel Processing Clerk specialization, we compared the duty profiles of typical full performance jobs at grade GS-5 in three different CODAP groups (See Table 1). We selected the "purest" example of personnel processing work as a control group to determine whether the other two groups were sufficiently similar for classification purposes. Employees in the other groups spent more time in staffing support work, but close examination of the kind of tasks actually performed showed them to be work that is easily learned, e.g., arranging for physical examinations and sending out routine appointment letters. We, therefore, provided for the performance of incidental staffing tasks in the new specialization.

- We found the duty profiles of CODAP groups of agency staffing jobs, OPM examining jobs, and OPM job information jobs to be sufficiently similar for grouping in the Staffing specialization. Employees in these jobs spend considerable time providing information about how to apply for Federal employment or promotion and reviewing or evaluating qualifications of candidates. In a closer look at the tasks performed under the duty headings, we were also able to see similarities in the clerical jobs between the groups and similarities in the assistant jobs.

- The CODAP grouping of federal employee benefits and incentive awards kinds of jobs did not yield clear-cut assistant work. However, we were able to find several bona fide assistants in this specialty during factfinding in the field.

- The other CODAP groups formed individually and then together, presumably because of significant typing, recordkeeping, and other nonspecialized duties. These groups included employee development, labor relations, classification, and OPM certification clerks and employees who operate remote control computer terminals. This grouping supported on-site factfinding information that there are no common career ladders or common guidelines across agency lines to support all of the current specializations authorized for clerks. In fact many of these jobs are established to provide the principal clerical support or secretarial service for an office. We recommended a more general title of Personnel Clerk for positions in these groups if they clearly should be included in the GS-203 series.

Our final recommendation was to reduce the number of specialized clerical titles from six to three; we retained specialized titles of all assistants because they perform routine technical work of personnel specialist occupations and must have KSA's related to those occupations.

## Level Analysis

After identifying the specializations, we looked for progressive levels of difficulty in the work which would later be matched to grades. CODAP provided clues, but not reasons. We compared the average time spent on significant individual tasks by grade level within the CODAP groups. It usually follows that, when the amount of time spent on a particular task progressively increases at higher grades, data related to that task might be useful as classification criteria in distinguishing between grades. Conversely, when a task is performed a similar amount of time at all grades, it usually has no value for grade-determining purposes.

Data from other factfinding sources were used to explain the reasons or characteristics that make some kinds of work more difficult than others. To do this, we needed to know how the work is done, the kind of guidelines used, problems in applying the guidelines, the degree of judgment required in decision-making, etc.

## FES Analysis

As with the level analysis, CODAP procedures did not provide the kinds of detailed information needed to develop FES classification criteria in the nine-factor format. Again, we relied heavily upon information gained through observations and interviews in arriving at conclusions.

# CONCLUSIONS

Based on our experience in using CODAP as an analytical tool for the Personnel Clerical and Assistance Series, GS-203, we conclude that:

## Potential Value in Standards Studies

CODAP might be of considerable value in standard studies for occupations which have mixed kinds of work, variations in the makeup of jobs, and a large number of employers, for example, the Administrative Officer Series, GS-341; Support Services Administration Series, GS-342; and Management Analysis Series, GS-343. The initial problem in these studies is to identify the different kinds or categories of jobs for further analysis, coverage of the occupation, and positions to be excluded. While an occupation may seem too complex for the human brain to cut through the maze, CODAP readily provides a diagram of the occupation and duty profiles of kinds of jobs. The number of jobs in each category is also considered in deciding the extent and nature of classification treatment needed. These procedures, however, do not negate the need to do additional factfinding such as on-site interviews and observations.

The face validity of CODAP reports might also be useful in studies when there is controversy such as conflicts between interest groups regarding distinctions between occupations and pressures for higher grades.

## Potential Value To Other Organizations

CODAP would have potential spinoff value for agencies if they were able to use the reports pertaining to their own employees for internal projects such as assessing training needs.

## Economy

The addition of CODAP methodology in our classification standards studies, would not be cost effective. In this particular project, the survey added 1 and 1/2 years to the project at an estimated cost of $50,000 to OPM and another $50,000 to agencies for salaries of participants and CODAP processing. We doubt that these costs could be pared below a total of $50,000 with a smaller sample size and distribution of forms by mail.

In summary, the use of CODAP in classification standard studies does not negate the need for on-site interviews; and, for the usual study, the same information derived from CODAP can be obtained through interviews and other traditional factfinding methods. In view of the additional costs to the project and disruption of agency offices, we cannot recommend its standard usage as an analytical tool in SDC projects. An exception might be an extremely complex or highly controversial project; in which case, consideration should be given to contracting out the printing, distribution, and processing aspects of the survey.

## REFERENCES


Abbe, Elizabeth N., A Methodological Strategy for Identifying Similarities Among Jobs, Office of Personnel Management, Washington, D. C., Sept 1980

Archer, Wayne B. and Fruchter, Dorothy T., The Construction, Review, and Administration of Air Force Job Inventories, Lackland AFB, Texas, Aug 1963

Driskill, Walter, "Comprehensive Occupational Data Analysis Programs," presentation at conference of the Classification and Compensation Society, Washington, D. C., July 28, 1978

Gandy, Jay A. and Maier, William, Utah Clerical Linkup Study: Comparison of Federal and State Jobs, Office of Personnel Management, Washington, D. C. 1979

Lecznar, William B., Three Methods for Estimating Difficulty of Job Tasks, Lackland AFB, Texas, July 1971

Mann, Walter G., Jr., A Multipurpose Analysis of Clerical and Administrative Positions in the U. S. Virgin Islands, Office of Personnel Management, Washington, D. C., September 1980

Mead, Donald F., Development of an Equation for Evaluating Job Difficulty, Lackland AFB, Texas, November 1970

Melching, William H., and Border, Sidney D., Procedures for Constructing and Using Task Inventories, Ohio State University, Columbus, Ohio, March 1973

Newton, Jean, "Factfinding for CSC Standards," Civil Service Journal, 19-1, July-Sept. 1978, pp 12-13

Newton, Jean, Computer-Assisted Occupational Analysis for Classification and Qualification Standards, Office of Personnel Management, Washington, D. C. 1981

Prien, Erich, "The Function of Job Analysis in Content Validation," Personnel Psychology, 1977, 30, pp. 167-173

Schwartz, Donald J., "A Job Sampling Approach to Merit System Examining," Personnel Psychology, 1977, 30, pp. 175-185

TABLE No. 1

COMPARISON OF DUTY PROFILES

Personnel Processing Clerical Work -- Current GS-5 Positions

| DUTY/ FUNCTION | Group 1119 (N= 51) | | Group 811 (N=77) | | Group 745 (N=62) | |
|---|---|---|---|---|---|---|
| | % Time Spent | % Members Performing | % Time Spent | %Members Performing | % Time Spent | % Members Performing |
| A. Job Info | 1 | 59 | 3 | 96 | 1 | 100 |
| B. Tests | | 2 | | 17 | 1 | 55 |
| C. Qual Rev | 2 | 65 | 4 | 94 | 8 | 98 |
| D. Registers | | 8 | | 55 | 7 | 88 |
| E. Classification | 1 | 47 | 1 | 75 | 2 | 79 |
| F. SF-52's | 11 | 100 | 10 | 100 | 8 | 100 |
| G. Agency Staffing | 4 | 90 | 7 | 100 | 9 | 100 |
| H. Processing | 12 | 100 | 10 | 100 | 7 | 100 |
| I. Suspense Dates | 11 | 100 | 8 | 100 | 6 | 100 |
| J. Leave | 4 | 92 | 5 | 99 | 4 | 100 |
| K. Health/Life Ins | 19 | 100 | 16 | 100 | 10 | 98 |
| L. Retirement Proc. | 3 | 88 | 6 | 98 | 5 | 92 |
| M. Pay Procedures | 8 | 100 | 7 | 100 | 7 | 47 |
| N. Other Svs | 1 | 74 | 2 | 93 | 2 | 95 |
| O. Grievances | | 13 | | 36 | | 65 |
| P. Inc. Awards | | 10 | | 43 | 1 | 60 |
| Q. Employee Dev. | | 6 | | 8 | 1 | 34 |
| R. Labor Relations | | 2 | | 9 | | 20 |
| S. EEO | | | | 6 | | 32 |
| T. General Writing | 1 | 33 | 1 | 62 | 2 | 71 |
| U. Records Maint. | 4 | 98 | 7 | 100 | 7 | 100 |
| V. Meeting Arrang. | | 10 | | 16 | 1 | 71 |
| W. Computer Proc. | 8 | 82 | 6 | 92 | 3 | 77 |
| X. Typ. for Others | 1 | 47 | 1 | 74 | 2 | 92 |
| Y. Typ. Own Work | 1 | 55 | 1 | 83 | 2 | 96 |
| Z. Leading | 2 | 57 | 2 | 79 | 2 | 85 |
| Average Similarity Within Group = | 61.1 | | 56.4 | | 54.8 | |

Typical Personnel Processing Work

Interface of Staffing Clerical and Personnel Processing Work

Incidental Support Staffing Tasks

Clerical Work Not Requiring Personnel Knowledges

Note: Numbers are rounded; less than .5% dropped. Formula for significance: 5% average amount of time spent and 80% of group performing

AD P001358

Nogami, Glenda Y., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Tues. A.M.)

## An Empirical Investigation of First Term Attrition

The research reported in this paper represents an attempt to identify "high risk" attrition factors in first term enlistees. In contrast to previous studies, this resaerch looks at a comparison between soldiers still in the Army, early dischargees, and soldiers completing their tour (ETS). The variables being investigated include soldier demographics, job characteristics, work environment, attractiveness of civilian opportunities, quality of location of assignment, and soldier gender.

# An Empirical Investigation of First Term Attrition

Glenda Y. Nogami

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

## INTRODUCTION

First term enlisted attrition costs the Army hundreds of millions of dollars each year. In a 1980 Report to Congress, the General Accounting Office (GAO) determined that enlisted attrition from the 1974 to 1977 Cohort Years cost the government 5.2 billion dollars. These costs include the recruiting, training, and out-processing of early discharges, replacement recruiting, training and in-processing, in addition to veteran's and medical benefits to qualified early discharges (attritees). In addition to the high monetary costs, attrition also puts stress into the Army system. Personnel turbulence may affect unit readiness and unit effectiveness. Attrition creates a "domino effect" affecting recruiting (USAREC) and training (TRADOC) - i.e., more recruits and training for these new recruits is needed.

Although attrition is costly - as evidenced by both monetary and system stress, a certain amount of attrition is to be expected and welcome. No system of recruiting is perfect, so some "weeding out" of unqualified recruits is needed. What the optimum level of attrition should be has not been determined. What has been determined, however, is that the present attrition rate and attrition costs are too high.

When reviewing the literature on attrition, (c.f., Sinaiko, 1977; Goodstadt, Yedlin, 1979; Mobley, Hand, Baker and Meglino, 1979; Martin, 1977; Wiskoff, Atwater, Houle and Sinaiko, 1980; etc.) several things became clear. Differences in rates of attrition have been found between male and female soldiers (Ross & Nogami, 1981; Fox, 1979; Martin, 1979; Addington, 1979; and Tolk, 1978), between educational groups (Fox, 1979; Guthrie, Lakota, and Matlock, 1978; Manning & Ingraham, 1981), between female traditional and non-traditional MOS (Tolk, 1978; Ross & Nogami, 1981) and between stations of assignment (Whittenburg & Dahlinger, 1978).

Beyond these categorizations little research has been done. There have not been any answers to questions such as: How does male attrition differ from female attrition? What is it about female traditional MOS which effects attrition? What are the differences between stations of assignment which affects attrition? How do the various demographic and biographic characteristics interact to affect attrition?

---

The views expressed in this paper are those of the author and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

To systemically answer the above concerns, we at ARI have undertaken a four year programmatic research effort. We are now in the beginning of our second year. In this paper, I would like to present two studies--one, an analysis of the DMDC 1976 Cohort Data Base and, two, preliminary results from an on-going field research project.

1976 Enlisted Cohort Data Base

The objective of this research effort was to determine whether attrition rates differed by (a) soldier gender, (b) soldier characteristics, (c) type of discharge action, and (d) the traditionality of the Military Occupational Speciality (MOS). To achieve this objective, we analyzed the Defense Manpower Data Center (DMDC) fiscal year 1976 Enlisted Cohort Data Base.

To have comparable numbers for analysis, the ARI data base included all FY 1976 nonprior service 3-4 year female enlistees, and 10% of all FY 1976 nonprior service, 3-4 year male enlistees. Male enlistees were selected on the basis of the last digit of their social security number.

Because soldier gender, soldier characteristics and MOS were of importance for this project, only those MOS with at least 10 females and 10 male soldiers was included. Consequently, all combat MOS were excluded from the data base. In addition, all non-white, non-black, "other" personnel were excluded (this included only 680 persons - too small for analysis by ethnic group and MOS). For more information on the data base see Ross & Nogami (1981).

Analyses: Two types of statistical tests were used to analyze the data: (1) multi-dimensional chi-square and (2) analysis of variance. The multi-dimensional chi-square analyses compared all individuals within a traditionality category with all other individuals in other traditionality categories. The analyses of variance compared MOS attrition rates within a category to other MOS attrition rates in other categories. As such, the analysis of variance techniques emphasized the dispersion of different MOS attrition rates.

Results and Conclusions: 1. The variables of race, education, and AFQT were included in the overall multi-dimensional chi-square analysis used to evaluate gender and traditionality. Education, race, and AFQT related strongly to attrition; higher rates of attrition occur for (1) the non-high school diploma graduates (including GEDs) than high school diploma graduates, (2) whites than blacks, and (3) Category III and IV than Category I and II. It is important to note that there were several two-way interactions in these data (shown in Table 1). These interactions indicate that differences in attrition rates related to one variable (such as education) are not constant when a second variable (such as gender) is considered. In the data the overall difference between male and female attrition rates

TABLE 1

ATTRITION RATES FOR
EDUCATION x GENDER, RACE x GENDER
AND RACE x EDUCATION

EDUCATION x GENDER

| | | MALE | FEMALE | |
|---|---|---|---|---|
| EDUCATION | HSDG | .23 (4,428) | .39 (11,177) | .35 (15,605 |
| | NHSDG | .50 (3,453) | .56 ( 1,385) | .52 ( 4,838 |
| | | .35 (7,881) | .41 (12,562) | |

RACE x GENDER

| | | MALE | FEMALE | |
|---|---|---|---|---|
| RACE | BLACK | .33 (2,237) | .28 (2,430) | .31 ( 4,667 |
| | WHITE | .35 (5,644) | .44 (10,132) | .41 (15,776 |
| | | .35 (7,881) | .41 (12,567) | |

RACE x EDUCATION

| | | HSDG | NHSDG | |
|---|---|---|---|---|
| RACE | BLACK | .26 (3,616) | .48 (1,051) | .31 (4,667) |
| | WHITE | .37 (11,989) | .53 (3,787) | .41 (15,776 |
| | | .35 (15,605) | .52 (4,838) | |

AFQT x GENDER

| | | MALE | FEMALE | |
|---|---|---|---|---|
| AFQT | (Cat. I, II) | .27 (2,285) | .40 (9,051) | .37 (11,336 |
| | (Cat. III, IV) | .38 (5,596) | .44 (3,511) | .40 (9,107) |
| | | .35 (7,881) | .41 (12,562) | |

for Education x Gender was .06 (.35 - .41). But this difference was not constant for males and females at the two education levels. Graduate men and women differed by .16 (.23 - .39) while non-graduates differed by only .06 (.50 - .56) (See Table 1).

2. Males and females attrited for different reasons. More females than males attrited due to family related causes and pregnancy (9% and 25% differences, respectively). Male attrition was higher for TDP (8%), EDP (5%), medical (6%) and adverse causes (15%). There was no difference between male and female attrition due to "other non-adverse causes" (See Table 2).

**TABLE 2**

**ATTRITION RATES FOR SEPARATION CATEGORIES**

| | TDP | EDP | MEDICAL | PREGNANCY | ADVERSE | OTHER NON-ADVERSE | FAMILY RELATED |
|---|---|---|---|---|---|---|---|
| Male | 33% | 25% | 13% | -- | 24% | 2% | 3% |
| Female | 25% | 20% | 7% | 25% | 9% | 2% | 12% |

3. The MOS job traditionality data were analyzed with multi-dimensional chi-square and an analysis of variance techniques. The multi-dimensional chi-square analysis showed that MOS job traditionality has a moderate effect on female attrition rates (See Table 3). Overall female attrition was lowest in the traditional female MOS category, intermediate in the less traditional, and highest in the non-traditional female MOS category. For males, traditionality of MOS categories appeared to have no effect. In contrast to the chi-square analysis, the analysis of variance for job traditionality and gender was non-significant even though the percent/proportion differences between males and females was as large as or larger than the differences in the chi-square analysis.

In addition, the differences between male and female attrition rates is not as simple as Addington (1979) would lead us to believe. Addington suggested that female attrit at a constantly higher rate than males over MOS. OUr data indicates that in may MOS, females do have ahigher attrition rate (e.g., 03c with 21.8% more female attrition; 71D with 28.4% more female attrition, etc.). However, there are MOS where males have a higher attrition rate than females (e.g. 91R, with 21.9% more male attrition; 91T, with 9.1% more male attrition) and MOS with similar attrition rates for males and females (e.g., 71G and 91D). In an effort to understand the dynamics which influence different MOS attrition rates, we have embarked on a field research project. (See Table 4)

TABLE 3

## Multi-dimensional Chi-Square Analysis

**Gender X Traditionality**
**$p < .0000$**

| | Male | Female | Difference |
|---|---|---|---|
| Traditional | .34[1] | .37 | +.03 |
| Less Traditional | .35 | .43 | +.08 |
| Non Traditional | .35 | .46 | +.11 |

## Analysis of Variance

**Gender X Traditionality**
**$p = .18$**

| Not significant | Male | Female | Difference |
|---|---|---|---|
| Traditional | .28 | .36 | +.08 |
| Less Traditional | .35 | .41 | +.06 |
| Non Traditional | .31 | .45 | +.14 |

[1] Proportion of attrition to non-attrition. Can be directly translated to percent attrition.

TABLE 4

COMPARISON OF MALE AND FEMALE ATTRITION FOR SELECTED MOS

TRADITIONAL MOS

| PMOS | ($\Sigma$ n)MALE | ($\Sigma$ n)FEMALE | %MALE ATTRITION | %FEMALE ATTRITION | DIFFERENCE |
|---|---|---|---|---|---|
| 03C | 11 | 15 | 18.2 | 40.0 | 21.8 |
| 05B | 129 | 67 | 40.3 | 47.8 | 7.5 |
| 26Q | 21 | 57 | 4.8 | 17.5 | 12.7 |
| 31N | 14 | 69 | 50.0 | 55.1 | 5.1 |
| 31V | 104 | 70 | 42.3 | 37.1 | - 5.2 |
| 32D | 18 | 29 | 27.8 | 37.9 | 10.1 |
| 71D | 19 | 95 | 10.5 | 38.9 | 28.4 |
| 71G | 18 | 29 | 27.8 | 27.6 | - 0.2 |
| 91D | 27 | 56 | 29.6 | 26.8 | - 2.8 |
| 91P | 10 | 49 | 10.0 | 36.7 | 26.7 |
| 91R | 14 | 96 | 50.0 | 28.1 | -21.9 |
| 91T | 11 | 11 | 27.3 | 18.2 | - 9.1 |

## Field Investigation of First Term Attrition

The objective of this research is to determine how attrition varies as a function of the characteristics of (a) the enlistee (i.e., demographics, reasons for elisting, morale); (b) the MOS or job (i.e., traditionality of MOS, job environment, and competing civilian opportunities); and (c) location of assignment (i.e., Continental United States (CONUS)-Europe (USAREUR), well-liked vs. disliked post, availability of recreational and service facilities).

The resulting research design incorporates 2 geographic locations (CONUS and USAREUR), 2 levels of quality of installation (desirable vs. undesirable), 4 levels of MOS traditionality (female traditional, less traditional, non-traditional and combat) and gender (male and female). (See Figure 1).

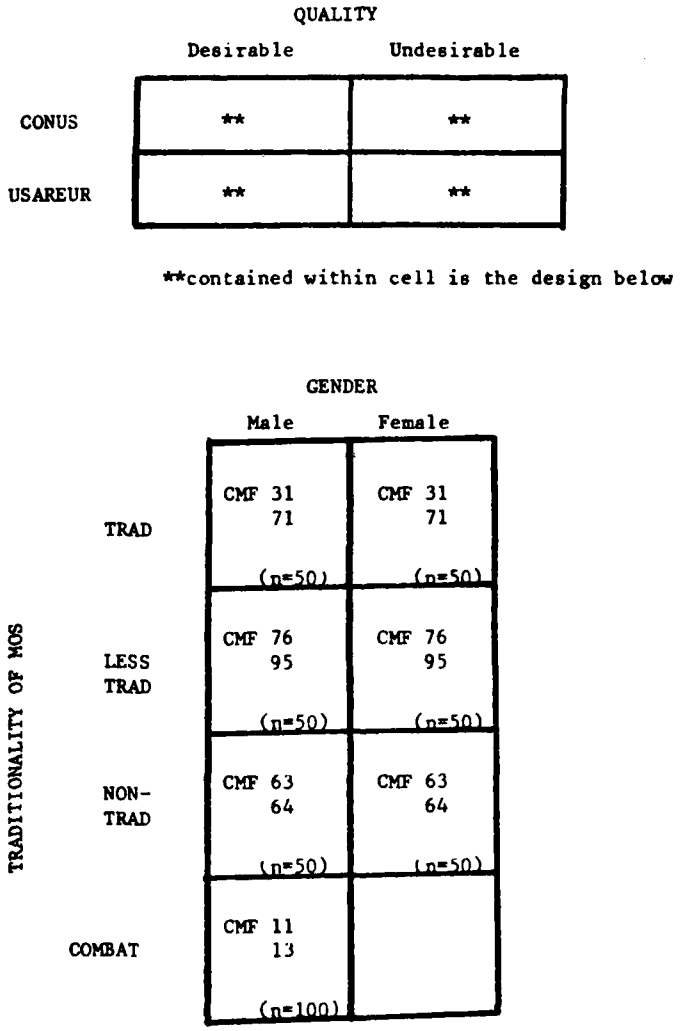



FIGURE 1. Research Design

Fifty first term enlisted males and fifty first term enlisted females in the traditional, less traditional an non-traditional MOS and one hundred first term enlisted males in the combat MOS were surveyed. In addition, one hundred first term enlisted early dischargees and one hundred soldiers completing their term of enlistment were surveyed to provide comparison groups.

To provide a leadership perspective to first term attrition, 25 non-commissioned officers (E6 and above) and 25 Company commanders were administered a leadership survey.

Results: Data collection in CONUS has been completed; but we are still waiting for 20 questionaires which have been in the mail for over 3 weeks. Data collection in USAREUR is nearing completion. Consequently, the results presented today will only cover the CONUS data, and these are preliminary findings.

Table 5 presents the total numbers of analyzable enlisted questionnaires in CONUS. The numbers of usable questionnaires from the early dischargees and soldiers completing their enlistment was too low to allow a full factorial design. Until all data is in, analysis of these may be misleading. Consequently, the data presented will only be on first term enlisted soldiers still in the Army.

The preliminary results presented below were selected from over 130 questions. They were selected on two criteria: (1) there must be two or more results in the same direction and (2) acceptable significant levels must have been demonstrated. The results are presented in five sections: (1) reasons for enlisting, (2) work environment, (3) off-duty environment, (4) MOS (Job) characteristics, (5) availability of facilities.

| | LOCATION | | | |
|---|---|---|---|---|
| | UNDESIRABLE | | DESIRABLE | |
| | M | F | M | F |
| TRAD | 46 | 34 | 41 | 48 |
| LESS TRAD | 44 | 47 | 35 | 34 |
| NON-TRAD | 83 | 43 | 63 | 25 |
| COMBAT | 113 | | 99 | |

TABLE 5

DEMOGRAPHICS: TRADITIONALITY OF CMF BY GENDER BY DESIRABILITY OF LOCATION

1. Reasons for Enlisting: Career opportunities are more important to females than to males as a reason for enlisting.

1. Career opportunities in the military look better than those in civilian life. (19) (p=0.028)

Males = 2.38
Females = 2.24

2. I could make more money outside the Army. (70) (p=0.00)

Males = 4.03
Females = 3.62

3. A person can get more of an even break as a civilian than as a soldier. (51) (p=0.083)

Males = 3.52
Females = 3.36

2. Work Environment: A more desirable location is related to more satisfactory work climate.

1. All in all, I am satisfied with the soldiers in my work group. (58) (p=0.009)

Desirable = 3.34
Undesirable = 3.07

2. All in all, I am satisfied with the Army. (59) (p=0.02)

Desirable = 2.84
Undesirable = 2.64

3. All in all, I am satisfied with my unit. (60) (p=0.031)

Desirable = 2.55
Undesirable = 2.34

3. Off-Duty Environment: A. The installation we labeled as desirable is confirmed by respondents.

1. I want a reassignment to another post. (41) (p=0.003)

Desirable = 3.41
Undesirable = 3.73

2. All in all, this is a good post for me to live on. (61) (p=0.000)

Desirable = 2.69
Undesirable = 2.21

**B. Females need more time off to take care of personal and family needs.**

1. From the time you arrived at this installation, how many days have you been sick and could not work? (6A) (p=0.000)

Males = 2.12
Females = 2.86

2. I have enough time off to take care of my personal and family needs. (38) (p=0.035)

Males = 2.53
Females = 2.32

4. MOS (Job) Characteristics: A. Females have more of a mismatch between their PMOS and duty MOS than males.

1. MOS: I am working in my Primary MOS. (119) (p=0.000)

Males = 1.34
Females = 1.61

2. I am working in the job areas for which I have been trained. (121) (p=0.001)

Males = 1.26
Females = 1.39

3. What is your PMOS? What is your DMOS? (8,10)

Average Percent Mismatch
Males = 13.28%
Females = 27.65%
(see Chart)

B. Males spend more duty time in traditional male jobs (outdoors), females spend more duty time in traditional female (desk) jobs.

1. Outdoors (76) (p=0.000)

Males = 3.19
Females = 2.54

2. Doing heavy labor (77) (p=0.000)

Males = 2.45
Females = 1.86

3. Dangerous work (78) (p=0.000)

Males = 2.13
Females = 1.63

4. Dirty-Muddy-Oily work (79) (p=0.000)

Males = 2.90
Females = 2.20

5. Ash and Trash (80) (p=0.030)

Males = 2.27
Females = 1.96

6. Indoors (81) (p=0.000)

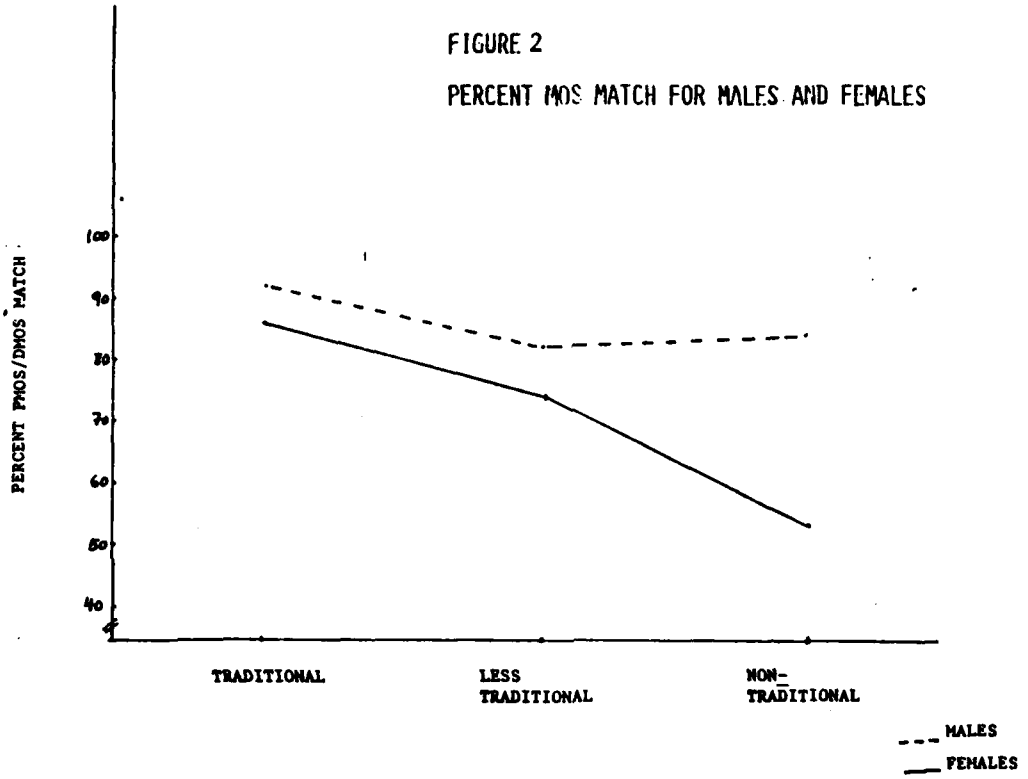
Males = 2.63
Females = 3.51

7. Doing paper work (82) (p=0.000)
Males = 2.48
Females = 3.45

8. Doing important work (83) (p=0.001)
Males = 3.15
Females = 3.61

FIGURE 2

PERCENT MOS MATCH FOR MALES AND FEMALES




5. Availability of Facilities: At both desirable and undesirable loc soldiers are satisified with about the same total number of on-post pl off-post facilities.

Satisfaction with Facilities and Services

| | Location | |
|---|---|---|
| | Undesirable | Desirable |
| On-Post | 7.57 | 6.025 |
| Off-Post | 3.93 | 5.06 |

References


Addington, J. Women's attrition: Establishing the relationship between high attrition and non-traditional skills. Information paper, DAPE-MPE-SS, 1979.

Comptroller General (General Accounting Office). Attrition in the military-An issue needing management attention. Report to the Congress. FPCD-80-10, 1980.

Fox, A. J. A comprehensive investigation of first term enlisted Army attrition. Draft report, Military Strength Programs Division, DCSPER, 1979.

Goodstadt, B. E. and Yedlin, N. C. A review of the state-of-the-art research on military attrition: Implication for policy and for future research and development. Final technical report. Washington, D. C.: Advanced Research Resources Organization, 1979.

Guthrie, N., Lakota, R., and Matlock, M. Voluntary release pilot program: Effects of attrition on general detail personnel. San Diego, California: NPRDC, 1978.

Manning, F. and Ingraham, L. Personnel attrition in the U.S. Army in Europe. Armed Forces and Society, 1981, 2, 256-270.

Martin, A. J. Trends in DOD First term attrition. In H. W. Sinaiko (Ed.), First term Enlisted Attrition - Vol. 1: Papers. ONR Technical Report 3, 1977.

Mobley, W., Hand, H., Baker, R., and Meglino, B. Conceptual and empirical analysis of military recruit training attrition. Columbia, South Carolina: University of South Carolina Center for Management and Organizational Research, 1979.

Ross, R. M. and Nogami, G. Y. The impact of MOS traditionality and soldier gender on first tour attrition. Draft ARI Technical report, 1981.

Sinaiko, H. W. (Ed.) First-term enlisted attrition - Volume 1: Papers. Washington, D. C.: ONR Technical report, 1977.

Tolk, A. B. The effects of MOS mismatch on females working in traditionally male and traditionally female MOSs. Final research report. Arlington, VA: Galler Associates, Inc., 1978.

Whittenburg, J. A. and Dahlinger, N. E. Optimum tour length in USAREUR: First term enlisted personnel. ARI draft technical report, USAREUR, 1978.

Wiskoff, M. F., Atwater, D. C., Houle, M. M., and Sinaiko, H. W. Enlisted first term attrition: Literature review and discussion. San Diego, California: NPRDC TR-80-, August, 1980.

AD P001359

Oelrich, Fritz & Bennett, Bruce, Selective Service System, Washington, DC. (Wed. P.M.)

## Revitalizing the Selective Service Local Board Structure through ADP Technology

The change in role for the Selective Service Local Board, from a "selecting" agency, to a "claims" agency whose primary function is to hear requests for judgmental reclassification from young men selected for induction, and the change in procedures from a State "draft call" to a National call process, led to new concepts in establishing the local board network. The presentation will briefly review the historical function of the local board network. The presentation will briefly review the historical function of the local board and contrast the pre-1975 system with post-1980 concepts.

During World War II, there was approximately one local board for each county in the United States. Present laws permit up to five counties to be represented by a single board. The presentation discusses the method of clustering actual or potential registrants by ZIP code to establish evenly distributed local boards with balanced workloads. A detailed discussion of the method, which relies on a computer plot of actual registrant ZIP codes by geographic coordinates, will follow.

Examples of the computer product will show how the clusters are formed, then how computer techniques "cross-level" populations to balance workloads. Overlaying the computer plots on state and county maps produces easily defined boundaries for each local board. Examples of "real-time" outputs will demonstrate how the method has been applied to establish the current local board system. Further discussion will focus on the flexibility of the system and its ability to respond to specific criteria.

PREVIOUS PAGE IS BLANK

# THE USE OF COMPUTER GRAPHICS FOR PRACTICAL PROBLEM SOLVING

by

Fritz Oelrich and Bruce Bennett

In early 1981, Selective Service was tasked with the job of redeveloping local draft boards to cover the United States. These boards needed to have easily identifiable boundaries, and also were to meet certain size and composition attributes. The problem faced by Selective Service was similar to a variety of boundary determination problems, including voter redistricting and school assignments. Selective Service's contractor--Science Applications, Incorporated--was able to employ a computer graphics approach to the problem, solving it with relative ease and at a much lower cost than would have otherwise been the case.

## Background

With the discontinuation of the draft in 1973, the Selective Service System was gradually phased down to a "deep standby" status in 1976. In that status, the central structure of the system was maintained at National Headquarters in Washington, but the local draft boards and area offices which would be needed in a draft no longer existed. As then constituted, the entire Selective Service System would have required months to be reconstituted before a draft could begin, making the status of the system, indeed, a "deep standby" no matter what the nature of the national emergency that would require a reinstitution of the draft.

In 1980, the draft registration was begun again by act of Congress, and a general revitalization program was also initiated. The goal of this program was to make it possible to provide the first inductees to training bases within two weeks of the beginning of a major mobilization. Clearly, then, local draft boards had to be reconstituted in order to meet and sustain these new mobilization requirements (though these boards would remain dormant until an induction was initiated). Further, Selective Service was required to operate a "uniform national call" with no inequity between different regions of the country. In order to meet both the equity and time requirements of an induction, Selective Service determined that the issuance of induction notices would have to occur from the national headquarters, but in doing so was faced with a dilemma: how should local draft board boundaries be drawn so that each induction candidate's assignment could be easily determined?

## II. APPROACHING THE PROBLEM

Traditionally, local draft board boundaries were determined by the states and usually conformed roughly to existing political boundaries. Legislation for the Selective Service System made only four requirements: (1) the basic geographic unit should be the county; (2) for counties with few registrants, up to five contiguous counties could be combined into a single board; (3) for counties with many registrants, as many boards as necessary could be created; and (4) the staffing of the board should reflect the race and national origin of the registrants living therein. Clearly, for those boards which covered one or more counties, simply knowing the county of residence for an induction candidate would allow Selective Service to assign him to the appropriate board. Unfortunately, many registrants do not know the name of the county in which they live, making even these assignments difficult at times. However, the real problem in determining assignments was in counties with two or more boards, since board boundaries were often arbitrary and assignments were extremely difficult to determine without a detailed map of the area and some knowledge of the area as well.

Within those counties with multiple boards, the traditional system also had two other problems. First, since board assignments were difficult to determine, it would be difficult to keep track of the number of registrants assigned to each board in a national registration system, as begun in 1980. Without knowing these numbers, Selective Service could not modify local board boundaries in order to equalize the workload of boards within a given county. More importantly, the use of political boundaries generally did not lead to the formation of boards with homogeniety in race and national origin, making it more difficult to meet the legal requirements in these areas.

### Defining Geographic Areas

What was needed was a system which could provide a simple, direct indication of where each registrant lives, and in turn could serve as the basis for defining board boundaries. After some contemplation, it became clear that zip codes provided the best and probably only such system. To begin with, each zip code is (nominally) assigned to a county by the Postal Service, and thus determining the county of each registrant from his zip code would be relatively easy. Further, since registration was to be done by the Postal Service, they could easily verify the zip code information submitted by registrants. Also, since registrants were required to specify their zip codes when registering, the use of zip codes to define board boundaries made it possible to project the workload of local boards and to determine local board assignments when preparing induction

notices at the national level. Finally, a wealth of demographic information is available by zip code, allowing the establishment (to the extent possible) of boards homogeneous in race and national origin, simplifying the requirements for choosing local board members. In short, zip codes clearly provided the basis needed for making local board assignments.

## In Search of a Technique

While zip codes provide an useful geographic description of this country, the division of the country into local boards required that each zip code be precisely located, and then some procedure had to be applied in order to group the zip codes into local boards. In essence, the requirements here were similar to those in a variety of related problems such as voter redistricting and the determination of school boundaries. In a more general sense, a whole range of assignment-type problems faces the same challenges: (1) determining the relative locations and attributes of constituent parts, (2) the lack of a simple algorithm for combining, matching, and/or clustering these parts, and (3) the necessity of showing graphically (in the form of a map) the nature of the assignments made.

The initial efforts to solve these challenges at Selective Service involved two pilot states: Maryland and Kentucky. In these states, detailed zip code maps were obtained for the large urban areas, and the other zip codes were located by city on state maps. Once the state zip code maps were completed, the division of the local boards proceeded. While these pilot efforts showed the feasibility of the concept, they also clearly showed a number of problems with this technique of processing. First, the process was extremely time consuming and required a detailed knowledge of the area studied. Thus, while Selective Service had hoped to present new local board boundaries to the state directors of Selective Service at a conference in April, 1980, by January we realized that this procedure would allow us to complete the work on only about half a dozen states by that time. Second, the zip code maps available were often incomplete and did not cover many of the suburban and near urban areas adequately. As a result, many zip codes were added after the initial assignments were found to be incomplete, and the assignments then had to be redone. Finally, the entire process was very tedious, especially because of the inadequacies in the data and the iterations required for completeness.

At that point, we discussed these problems with one our contractors, Science Applications, Incorporated (SAI). The SAI staff suggested using a computer as the medium on which to perform our assignments. In order to locate each zip code, the latitude and longitude of the zip code would be specified. This information could then be drawn on a computer screen for a county that required more than one local board, and that county could be divided by a person using some form of clustering

procedure on the screen. If successful, the indicated computer technique would allow all local board assignments to be made in a much shorter time period; our goal was to complete the entire country in time for our April conference.

## III. IMPLEMENTING THE TECHNIQUE

The actual division of zip codes into local draft boards involved two steps. First, since local boards are always associated with counties, the county for each zip code had to be determined, and the corresponding local board or boards identified. Fortunately, previous work by Selective Service had already determined the pairings between local boards and counties based upon registration estimates for each county; in some cases these assignments were eventually modified because the actual registration data varied somewhat from the estimates. For counties with only one local board, or for local boards with more than one county, this identification process was sufficient to determine the local board. For counties or groups of counties with more than one local board, the second step involved the division of the zip codes into what were generally geographically compact, roughly equal-size boards.

### Data Problems

In the process, we encountered a variety of data problems, making these two steps somewhat more difficult than expected. First, while there are about 38,000 valid zip codes, our initial data set contained approximately 50,000 zip codes, many with only one or two registrants. Comparing the Selective Service zip codes to a post office master file showed that many of the extra zip codes were clearly due to errors on the part of one or two registrants (for example, the transposition or deletion of digits), though in other cases some "invalid" zip codes involved dozens or hundreds or registrants, and thus more likely reflected formerly valid or newly created zip codes. The zip codes containing 5 or more registrants were identified and retained, and the others were dealt with by combining all such zip codes with the same three digit prefix into a single zip code with that prefix followed by "00" (since the Postal Service does not assign zip codes with zeros in both of the last two digits--22100, for example). Thus, if 22194 and 22177 were both identified as "bad" zip codes, they were eliminated and the number of registrants therein was combined to form the 22100 zip code. Our resulting list contains just over 37,000 zip codes (less than 38,000 since some of them are only for industrial or commercial use).

Our next problem involved determining the county and state in which each zip code is located. While several data bases that we obtained contained this information for many of the zip codes, quite a number of the zip codes where not contained in any other data base, and thus this information had to be determined by hand. In other cases, the county or state assignment was not consistent between data bases. In part, this problem is due to the fact that the area associated with many zip codes cuts across county, and in some cases, state boundaries. Even when a zip code was consistently shown in a single county, there were a few cases in which the indicated county was incorrect. Therefore, a variety of simple checks were performed on the data, and the final results where reviewed by individuals from each state in order to reduce the possibility of error.

Once these problems were resolved, a simple table look-up procedure was used to assign zip codes either directly to their local boards or to an aggregation of local boards which was then to be split using techniques discussed below. In a few cases, these assignments resulted in the creation of local boards in excess of our maximum size guidelines; these boards where either split along county lines in the case of multi-county boards, or reserved to be split along with other multi-board counties if only a single county was involved.

Our final problem was determining the location of each zip code in terms of latitude and longitude, which would enable us to graphically display the data. One of the data bases we obtained had this information for many of the zip codes, but when this data was plotted, it became apparent that there were some errors (for example, some zip codes were located in oceans or otherwise outside of the appropriate state or area). The location data was run through a number of checks, including comparing sample plots to the available zip code maps for major urban areas. For the hundreds of points for which no latitude and longitude values were available, maps were used to roughly determine these values. For the "00" suffix zip codes, which were aggregations of errant data, the registrants combined therein were located at the average location of all registrants associated with the appropriated three digit prefix.

## IV. DIVIDING COUNTIES INTO LOCAL BOARDS

Once many of the above data manipulations were completed, we were in a position to begin dividing the zip codes into local boards for those counties with more than one local board. Our guidelines for this activity were fairly simple: within each county, the boards should each (1) contain roughly the same number of registrants, (2) have no more than 1500 registrants per year group, and (3) be geographically compact.

While we could theoretically have used some kind of automated cluster analysis procedure for creating these groupings, we began creating boards directly on the computer, retaining the hope that some fairly obvious algorithms would surface which could be automated to produce at least an initial assignment for each board.

## The Equipment Used

This work was performed on SAI's Hewlett-Packard 1000 mini-computer system, utilizing custom designed software based on the system's graphics package. This system is small enough to fit rather comfortably in a normal office, and provides a variety of capabilities above and beyond the specific application being addressed here.

We began the development of board boundaries by displaying the number of registrants at each location on the graphics display, as shown in the accompanying picture. In each case, the number of local boards to be created was already determined, and an average number of registrants per board was calculated to provide a guideline in board creation. The program then requested that the operator define the first board. This was done by moving the graphics cursor into the neighborhood of the zip codes to be chosen; the machine then placed the cursor on each point in that neighborhood, asking whether it was to be included in this board. If it was to be included, the point was removed from the screen, and the cumulative number of registrants in this board was displayed. Once the first board was completely defined, the computer prompted for input of the second board, and so forth. The program recognized when the final board was reached, and combined all of the remaining points into the that board. At this point, the assignments made could then be stored, plotted, or modified at the operator's discretion.

Once the board assignments were determined for each county, a county portfolio was prepared. This portfolio included a list of zip codes for each board showing the name of each zip code, the number of registrants born in 1960 and 1961, and the racial and ethnic composition of the people living within that zip code. The registrant and racial and ethnic data were also summed and averaged for each local board. The portfolio also includes two maps of the county drawn on SAI's eight color plotter. The first map shows the number of 1961 birth-year registrants for each zip code, and colors these numbers corresponding to their local board assignment. The second map essentially replicates the first, but shows the last four digits of the corresponding zip code instead of the number of registrants.

The importance of this procedure becomes clear when one realizes how quickly the process was performed. While Selective Service had assumed that boundaries could be established for local boards in about six states over

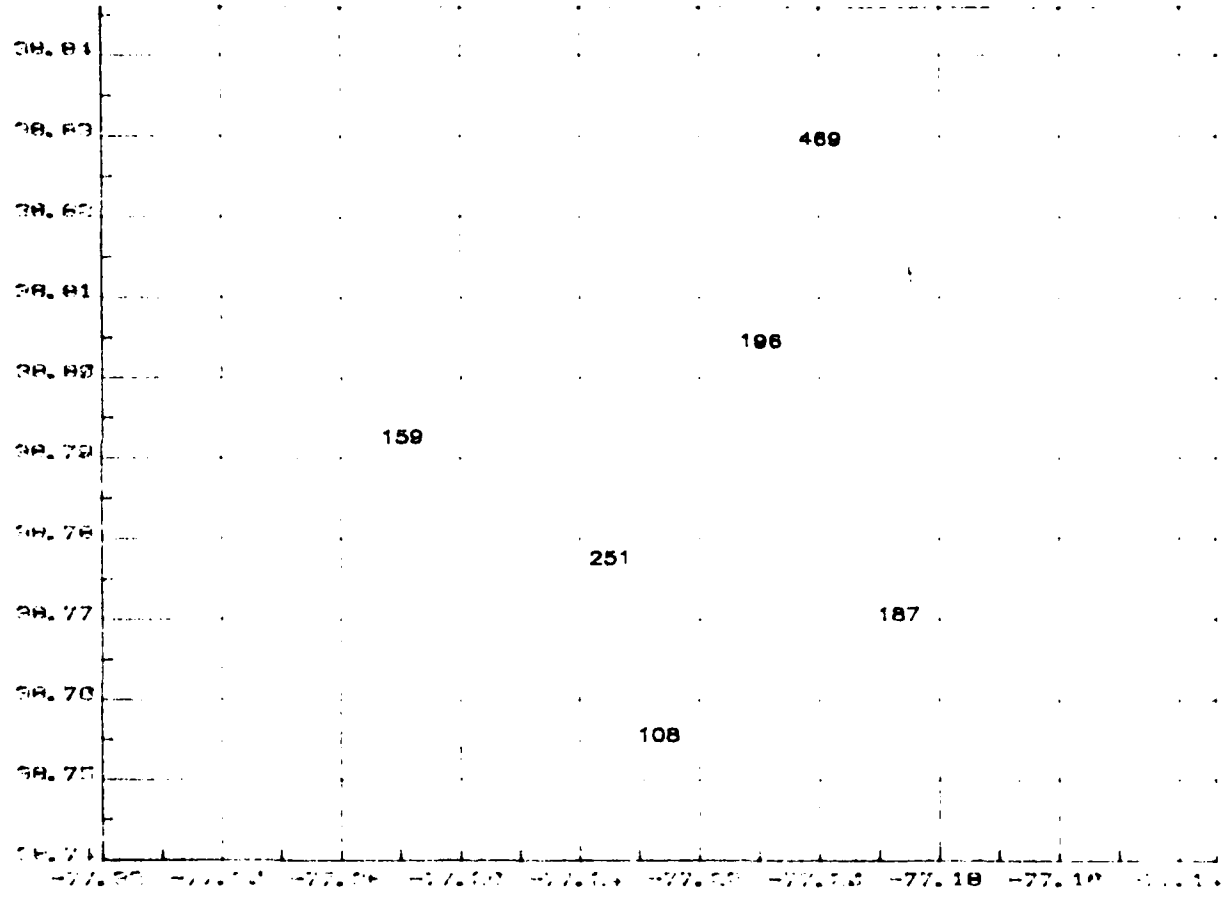
# OF BOARDS ?
0=MODIFY, −1=KEEP CURRENT BOARDS
−1000=KEEP BOARDS, BEGIN PLOTS
2

BOARD # ?
39

1370 TOTAL REGISTRANTS,
685 AVERAGE PER BOARD.

| | | | |
|---|---|---|---|
| 1 | 469 | 38.832 | −77.198 |
| 2 | 159 | 38.795 | −77.268 |
| 3 | 187 | 38.773 | −77.185 |
| 4 | 196 | 38.807 | −77.208 |
| 5 | 251 | 38.780 | −77.233 |
| 6 | 108 | 38.758 | −77.225 |

INPUT METHOD:
0 = POINT #s
1 = CURSOR



## LOCAL BOARD NO. 36

| ZIP | REGISTRANTS 1960 | REGISTRANTS 1961 | NAME | RACIAL DATA BLCK | RACIAL DATA OTHR | RACIAL DATA SPAN | FORN BORN |
|---|---|---|---|---|---|---|---|
| 22041 | 84 | 96 | FALLS CHURCH--BAILEYS CROS | 9% | 1% | 5% | 20% |
| 22042 | 215 | 213 | FALLS CHURCH--MOSBY | 3% | 1% | 3% | 13% |
| 22043 | 179 | 187 | FALLS CHURCH--PIMMIT | 3% | 1% | 2% | 12% |
| 22044 | 79 | 60 | FALLS CHURCH--SEVEN CORNER | 2% | 1% | 6% | 20% |
| 22046 | 119 | 97 | FALLS CHURCH | 6% | 1% | 3% | 13% |
| 22101 | 314 | 296 | MCLEAN--MC LEAN | 1% | 1% | 3% | 18% |
| 22102 | 89 | 89 | MCLEAN | 3% | 1% | 3% | 16% |
| 7 | 1079 | 1038 | | 4% | 1% | 3% | 16% |

## LOCAL BOARD NO. 37

| ZIP | REGISTRANTS 1960 | REGISTRANTS 1961 | NAME | RACIAL DATA BLCK | RACIAL DATA OTHR | RACIAL DATA SPAN | FORN BORN |
|---|---|---|---|---|---|---|---|
| 22000 | 13 | 10 | | | | | |
| 22020 | 52 | 62 | CENTREVILLE | 2% | 1% | 3% | 14% |
| 22024 | 27 | 31 | CLIFTON | 0% | 0% | 0% | 4% |
| 22030 | 393 | 385 | FAIRFAX | 3% | 0% | 2% | 9% |
| 22031 | 166 | 147 | FAIRFAX | 1% | 1% | 2% | 14% |
| 22032 | 202 | 269 | FAIRFAX | 2% | 0% | 2% | 11% |
| 22033 | 32 | 40 | FAIRFAX | | | | |
| 22034 | 1 | 2 | | | | | |
| 22039 | 46 | 53 | FAIRFAX STATION | 10% | 0% | 1% | 8% |
| 22100 | 10 | 13 | | | | | |
| 10 | 942 | 1012 | | 2% | 0% | 2% | 11% |

## LOCAL BOARD NO. 38

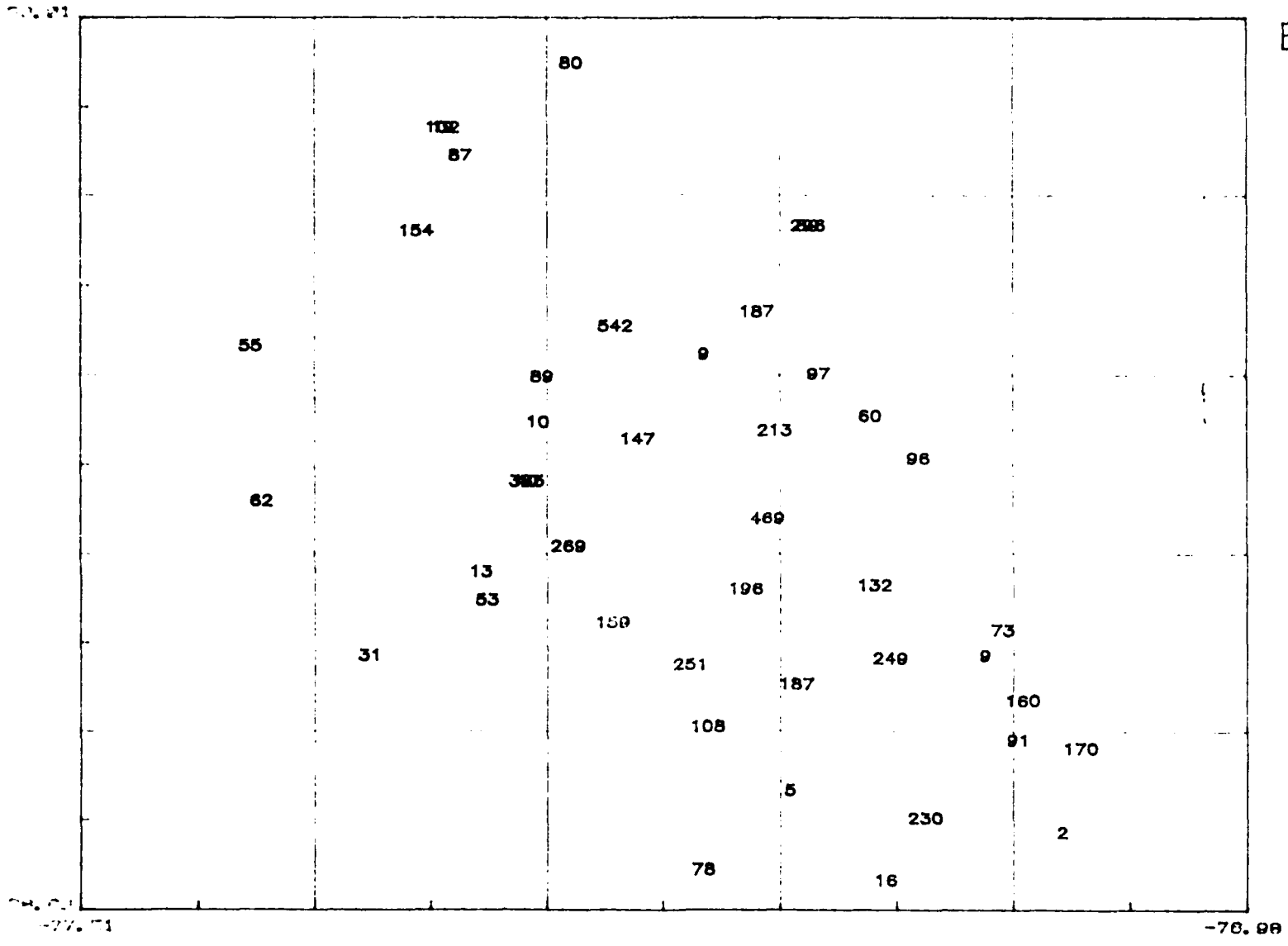
| ZIP | REGISTRANTS 1960 | REGISTRANTS 1961 | NAME | RACIAL DATA BLCK | RACIAL DATA OTHR | RACIAL DATA SPAN | FORN BORN |
|---|---|---|---|---|---|---|---|
| 22060 | 12 | 16 | FORT BELVOIR | 15% | 3% | 5% | 11% |
| 22079 | 65 | 78 | LORTON | 62% | 1% | 1% | 4% |
| 22121 | 0 | 2 | MOUNT VERNON | 21% | 0% | 2% | 9% |
| 22122 | 2 | 5 | NEWINGTON | 4% | 0% | 1% | 5% |
| 22300 | 9 | 9 | | | | | |
| 22303 | 77 | 73 | ALEXANDRIA--JEFFERSON MANO | 1% | 1% | 2% | 11% |
| 22306 | 186 | 160 | ALEXANDRIA--COMMUNITY | 5% | 1% | 2% | 12% |
| 22307 | 81 | 91 | ALEXANDRIA--BELLE VIEW | 4% | 1% | 3% | 13% |
| 22308 | 159 | 170 | ALEXANDRIA--WELLINGTON | 4% | 1% | 2% | 12% |
| 22309 | 225 | 230 | ALEXANDRIA--ENGLESIDE | 1% | 1% | 4% | 13% |
| 22310 | 248 | 249 | ALEXANDRIA--FRANCONIA | 1% | 1% | 1% | 10% |
| 22312 | 128 | 132 | ALEXANDRIA--LINCOLNIA | 2% | 1% | 4% | 15% |
| 12 | 1192 | 1215 | | 6% | 1% | 3% | 12% |

LOCAL BOARD NO. 39

| ZIP | REGISTRANTS 1960 | 1961 | NAME | RACIAL DATA BLCK | OTHR | SPAN | FORN BORN |
|---|---|---|---|---|---|---|---|
| 22003 | 470 | 469 | ANNANDALE | 1% | 1% | 3% | 14% |
| 22015 | 123 | 159 | BURKE | 1% | 0% | 2% | 10% |
| 22150 | 175 | 187 | SPRINGFIELD | 0% | 1% | 2% | 13% |
| 22151 | 196 | 196 | SPRINGFIELD--NORTH SPRINGF | 0% | 0% | 1% | 10% |
| 22152 | 248 | 251 | SPRINGFIELD--WEST SPRINGFI | 1% | 1% | 1% | 14% |
| 22153 | 95 | 108 | SPRINGFIELD--WEST SPRINGFI | 3% | 0% | 2% | 6% |
| 6 | 1307 | 1370 | | 1% | 1% | 2% | 12% |

LOCAL BOARD NO. 51

| ZIP | REGISTRANTS 1960 | 1961 | NAME | RACIAL DATA BLCK | OTHR | SPAN | FORN BORN |
|---|---|---|---|---|---|---|---|
| 22021 | 64 | 55 | CHANTILLY | 3% | 0% | 0% | 7% |
| 22027 | 7 | 9 | DUNN LORING | 1% | 0% | 1% | 13% |
| 22066 | 80 | 80 | GREAT FALLS--COLVIN RUN MI | 5% | 1% | 2% | 15% |
| 22070 | 119 | 102 | HERNDON | 3% | 0% | 0% | 7% |
| 22071 | 14 | 19 | HERNDON | 2% | 0% | 1% | 6% |
| 22090 | 81 | 87 | BETANA PARK--LAKE ANNE | 6% | 1% | 2% | 15% |
| 22091 | 145 | 154 | HERNDON--RESTON AREA 2 | 5% | 1% | 2% | 13% |
| 22124 | 69 | 89 | OAKTON | 4% | 1% | 1% | 10% |
| 22180 | 471 | 542 | VIENNA | 3% | 1% | 2% | 12% |
| 9 | 1050 | 1137 | | 4% | 1% | 2% | 11% |




# FAIRFAX, FAIRFAX CITY, ETC., VIRGINIA

## 1961 REGISTRANTS

approximately a four month period, an initial set of assignments on the computer required less than a month for the entire country after the initial data preparations. By the time of the Selective Service conference in April, 1981, we were able to make not only the initial assignments, but were also able to run two separate sets of updates and modifications based upon reviews of our early work.

## Modifying the Assignments

Over the course of our work, almost every set of assignments received some form of modification. Most of these changes were due to data problems, as discussed above. The software in the program was designed to facilitate these changes in a variety of ways. If only one or two zip codes needed to be changed, the boards involved were simply called onto the screen and the cursor was placed onto the zip codes to be shifted from one board to the other. If a county change was also required, the board change could also be made directly to the data base. In some cases, though, it was easiest to redraw the board boundaries within a county or area.

The flexibility with which changes can be made is one of the most important aspects of this system. Because some zip codes are changing continually and registration is an ongoing process, the board assignments are subject to constant review and change. The ability of the computer programs to make such changes easily will make it possible for the board assignments to be reflective of the underlying population from now on.

# V. LESSONS LEARNED

Perhaps the most important lesson learned from this project was the ease with which board assignments could be made using this procedure. In all, about 1,000 draft boards were created from multi-board counties, involving the assignment of about 20,000 individual zip codes to these boards, and yet the initial assignments required less than one man-month. Moreover, most of these assignments were made by individuals who were neither computer programmers nor experts in the area, but the interactive program was sufficiently user oriented to make the tasks relatively simple, allowing the user to feel comfortable about their work. Little training time was required to learn these procedures. Modifications to the initial assignments were performed using essentially the same procedures as the initial assignments, and were also easily learned and applied.

As part of this process, the results of the initial assignments were reviewed by people familiar with each area.

Most of the changes they suggested had to do with errors in the initial data (such as wrong county assignments or improper latitude and longitudes). Only in a few cases did their detailed knowledge of the local area lead to preferred board assignments significantly different from those prepared on the computer. Given that the computer did not contain information on roads, mountains, rivers, and other determinants of the actual distance between two points, this degree of agreement was reassuring.

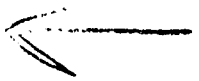
Throughout this entire activity, the decision rules being used by each person were reviewed in order to determine whether or not an initial automated assignment might be feasible. It was our conclusion that such a procedure would have to be very complicated because of the variations in geography and board requirements observed, and the output would, in any case, probably require fairly significant changes on the computer screen before the quality of the man-machine interaction could be duplicated. Thus, while some form of automated assignment probably could have been developed as a front-end to this process, it was our view that such a system would be very costly (both in terms of development and in terms of running costs) with relatively little benefit, and therefore any notion of developing such a procedure was discarded.

As the project progressed, we became aware of the fact that the individuals doing the assignments on the computer were learning a large amount about the areas with which they were dealing. They were able to recall many of the place names and general geographic layout of specific areas even though they had only spent a few minutes working with them. This observation suggests the value of having actual decision-makers make such assignments: the procedure is easy to learn and in the process they discover information about their area which they probably would not have otherwise known. In this sense, the computer gives a concrete, often unique picture of an area which can add to the perspective an individual has about the place being examined.

Perhaps the one weak aspect of our entire procedure was in not spending enough time training the reviewers individually to read and assess the materials that they were given. Initially, the reviewers were instructed in a group session without the materials in their possession. In almost all cases, someone familiar with materials then had to spend some time with each reviewer, walking through a single example based upon an area with which he/she was familiar. The entire process had become so simple and transparent to us that we underestimated the ability of people completely unfamiliar with this approach to appreciate what had been done. But once we had spent a few minutes walking through a single example, almost all reviewers found the materials understandable and quite useful.

## VI. SUMMARY

Due to close cooperation between Selective Service and our contractor, Science Apllications, Incorporated, we were able to develop boundaries for the local draft boards throughout the United States in an extremely powerful way. The computer graphics procedure in combination with the use of zip codes allowed us to meet the legal requirements for local boards much more easily than our previous system, and allowed us to perform this process in less than one-tenth of the time that it would have otherwise taken. The products of this system capture in interesting and useful ways information about the registrants that we had been unable to obtain before. By extension, these same procedures could easily be applied to a variety of boundary and assignment type problems, both in simplifying them and in providing much quicker solutions.



AD P001360

Olson, Darlene M., US Office of Personnel Management, Washington, DC. (Wed. P.M.)

## Biographical Correlates of a Cognitive Abilities Test for Firefighter Selection

This paper describes the results of biodata information collected with a 72 item biographical inventory, that was developed and administered to 476 firefighter applicants in the District of Columbia. Biograpical questions were developed from existing item pools and on the basis of rational relationships between proposed questions and successful job performance. The content of the questions addressed prior experiential activities, education, and interests of applicants. Since the long range validity of biodata is inconsistent and relatively little information is known about the adverse impact of biographical inventories, the data results are analyzed by race and responses to biodata questions are correlated with scores on a cognitive abilities battery. In addition, the intercorrelations of biographical variables were examined to determine whether a specific firefighter profile existed. Preliminary data results indicate that some biodata questions, such as education level, general job security, and the attractiveness of firefighter pay are stable across racial groups, while others such as experience in firefighting work, number of friends who are firefighters, and characteristics of the firefighter job like irregular shifts and working dangerous emergency conditions, showed large discrepancies between Black and White applicants.

# Biographical Correlates of a Cognitive Abilities Battery

Darlene M. Olson

U.S. Office of Personnel Management
Personnel Research and Development Center

This paper describes the results of the biographical information derived from a biodata questionnaire administered to 447 applicants for entry-level firefighter positions in a large metropolitan area. This information was collected to document the characteristics of the firefighter applicant sample, investigate the validity of biodata items as predictors and study the level of adverse impact of an empirically keyed biographical questionnaire (BQ).

## Rationale for the Development of a Biographical Questionnaire

In a comprehensive test development project conducted by Payne and van Rijn (1978) for the selection of entry-level firefighters a complex group of knowledges, skills, and abilities (KSAOs) was identified by a thorough task analysis of the job. When the job tasks were linked to the critical KSAOs by a panel of subject matter experts (SMEs), a group of cognitive, physical and affective abilities was identified. Some of the affective KSAOs judged to be important for the firefighter job were: 1) interest in working with people, 2) willingness to risk injury/death, 3) willingness to work rotating shifts, 4) interest in working with equipment, and 5) ability to work in high places and confined spaces.

Since the written Firefighter Selection Test (FST) assessed the cognitive abilities required to effectively perform the firefighter job, this experimental research was initiated to study whether an alternative selection instrument (BQ) could measure some of the affective components of this job. The biographical questionnaire format was chosen, because of its consistently high validity with a diverse group of criteria (Owens, 1971; Hsu, Darany & Nettles, 1978; and van Rijn, 1980).

Description of the Entry-Level Firefighter Applicant Sample. A total of 942 applicants took the entry-level FST, which was a cognitive abilities battery. Of the 942 applicants, 705 (74.84%) were Black, 200 (21.23%) were White, 12 (1.27%) were "Other" races and 25 (2.65%) did not respond to the race question. A breakdown of the total applicant sample by sex group indicates that 886 (94.1%) were male, 52 (5.5%) were female, and 4 (.4%) did not respond to the question.

As an experimental component of the FST administration approximately half of the applicants were randomly given a BQ and the remainder took a self-assessment instrument. A total of 447 applicants responded to the BQ.

Table 1 shows that the applicant sample, who took the BQ, consisted of 356 (79.6%) Blacks and 91 (20.4%) Whites. In the total BQ sample four Blacks and one White identified their national origin as Hispanic. The sex group composition of the total BQ sample was comprised of 95% males and 5% females. Table 1 demonstrates a comparable finding when race by sex group interactions are examined. For example, 94% of the Black and 99% of the White BQ samples are male. Consequently, the small numbers of females in the sample precluded meaningful comparisons of sex group differences.

Although Table 1 indicates that Black and White applicants had very similar profiles in terms of level of education, more Blacks tended to reside in the large metropolitan area to be served than Whites. In terms of the entry-level FST, Table 1 shows that there were statistically significant differences between the White and Black subgroups on all subtests of the selection instrument.

## Development of the Biographical Questionnaire

A total of 72 items were written or modified from the Catalog of Life History Items (Owens, Glennon, & Albright, 1966) and in accordance with the previously discussed results of the firefighter task analysis. Items were related to such areas as education, interests, hobbies, and work climate variables. Effort was expended to ask job-related questions and to eliminate any items that might be culturally biased, an invasion of privacy, or not legally justifiable for selection (e.g, marital status or religious preference). The items were both hard, which means they could be factually verified, and soft, which means they were more subjective and self-descriptive.

A scoring key for the BQ was developed utilizing the method described by England (1971). In accordance with this procedure, the 447 applicants were randomly assigned to two groups, a key developmental group with approximately two thirds of the sample (N= 298) and a hold-out group for the subsequent cross-validation, which contained approximately one third of the sample (N=149). A criterion measure, the FST predictor test, was used to divide the key developmental group into high and low performance groups. Each of these groups based on performance on the FST, included about 33% (N= 98) of the key developmental group.

In the key developmental group an analysis of the responses provided by the high and low performance groups to the 72 item BQ was conducted. Weights of 0, 1, or 2 were assigned to item alternatives on the basis of the percentage differences between the selection of each response by the two performance (high/low) groups. In order to assign the weights appropriately, "Strong's Tables of Net Weights for Differences in Percents" and the "Table of Assigned Weights Derived from the Net Weights" explained in England (1971) was followed. Results of the weighting procedure found that of the 72 total BQ items, 12 showed significant differences between the high and low performance groups. The BQs for applicants in the key developmental and hold-out groups were scored using the weights previously assigned to the 12 items, which differentiated the high from low performers on the criterion. Once a total score based on the differentiating items was computed, this score was correlated with the FST criterion. The correlation of the BQ total score for the devlopmental group with the FST was .56 and .49 for the cross-validation sample. Both of these correlations were significant at the $p < .01$ level. Hence, the BQ as presently empirically keyed appears to be a valid predictor of scores on the FST, a cognitive criterion.

## Results of the Data Analysis

Comparisons of Black and White Firefighter Applicant Profiles. The beginning of the BQ contained additional items designed to solicit information about the characteristics of individuals who apply to become firefighters. This information could help to determine for instance, to what extent persons who are familiar with the firefighter job, studied specific subjects in high school or worked a certain number of hours while in school, perform better on the FST, receive higher ratings on training school criteria and hence have a higher probability of becoming better firefighters. Table 2 shows the distributions of responses for the total biodata developmental group (distributions are comparable for total biodata sample and cross-validation group) and for Black and White subgroups separately on these variables. The asterisks in the "Total" column indicate the variables for which significant Black-White differences were found. For example, more Blacks tended to think about firefighting and decide to become firefighters after the 12th. grade than Whites. In addition, Blacks tended to have less volunteer/paid experience as firefighters and knew less friends/relatives who were firefighters than Whites. While in high school Blacks participated less in clubs/organizations and did less part-time work than Whites.

In order to examine Black-White differences on BQ items that related to recreation, hobbies, interests and work climate, Table 3 was constructed. Statistically significant results from this data indicate that Whites tended to perform landscaping, carpentry and plumbing activities more often than Blacks. With respect to recreation activities, Blacks participated in individual/team sports, and coached/officiated sports events more often than Whites. Also Blacks tended to participate more often than Whites in such activities as cooking, volunteering (school, hotline, Big Brother) and attending cultural activities (plays/concerts). The work climate variables shown in Table 3 demonstrate that many of the attributes of the firefighter job such as doing physical work in dangerous emergency situations and the performance of duties in high, dark and confined places were rated as more attractive by Whites than Blacks.

Intercorrelations of Selected Background Variables. Table 4 depicts the intercorrelations among designated "work attractiveness" biodata variables. Although the majority of the biodata items did not intercorrelate, some of the more significant work climate correlations are: a) job security correlates highest with pay (.64), b) working closely with people is related to a willingness to serve the community (.77) and the opportunity to learn new things (.75), and c) although the challenge of dangerous work was related to a willingness to work under emergency conditions (.74), in high places (.64), dark/smokey places (.50), and confined (.52) places; willingness to risk injury/death (.29) was not as strongly related to challenges in the firefighter job.

Biodata Variables as Predictors. Table 5 shows the correlations between selected BQ variables and the FST criterion. Race, education, exposure to/ and knowledge of the firefighter job, opportunity to serve the community, and work with people and equipment are significantly ($p < .01$) related to total test performance. A negative relationship between performance on the FST and residence (-.42), high school rank (-.21) and participation in volunteer activities (ranged from -.12 to -.22) was found. These observed intercorrelations among previously delineated variables and total FST score were also consistently observed across the six subtests of the criterion.

Table 5 also documents two interesting findings. First, the kinds of high school subjects studied and reading, which are both primarily cognitive activities, did not correlate with the cognitive criterion. Second, some job-related variables like risks, shifts, heights, dark and closets are not related to performance on the FST, even though they would have a high probability of correlating with job proficiency measures. From these results it appears that a major weakness of this research is the totally cognitive nature of the criterion.

Adverse Impact Estimates for the Scored Biographical Questionnaire. Adverse impact refers to a disparity in the rates with which minority or women applicants are selected for employment. In general, a selection rate for any racial group which is less than four-fifths (80%) of the rate for the group with the highest selection rate is evidence of adverse impact. Table 6 shows the potential adverse impact against Black applicants if the passing point (cut score) for the scored BQ was set at the mean score for the developmental group on the BQ, and if the cut score was set one standard deviation below the average BQ score. It is apparent from the data in Table 6 that a cut score at the mean ($X = 10$) of the developmental group on the scored BQ results in adverse impact for Blacks. When the cut score is lowered one standard deviation adverse impact is diminished, but not eliminated.

## Summary

Although some biodata variables were very stable across racial groups of entry-level firefighter applicants, others showed large discrepancies between White and Black samples. Major conclusions about the race differences in the biodata analysis suggest that Black applicants knew less about the firefighter occupation, had less firefighting experience and decided to pursue a career in firefighting much later than Whites. Perhaps racial group differences in applicant firefighters could be reduced if outreach programs, that provided career information on the fire service, were targeted toward minority groups. In addition, research has shown that specific BQ items with large amounts of adverse impact can be deleted from the BQ to eliminate the observed Black-White differences in the BQ score. However, since the BQ is empirically keyed in terms of items that differentiate high vs. low performers on the criterion, removal of even a small number of items could have a substantial negative impact on validity.

Although biodata has been shown to have substantial validity against a broad spectrum of criteria, the BQ has been criticized for being situation-specific and for being incapable of generalizing to other occupations or populations. Results of this particular research have shown that it is possible to apply an empirical scoring methodology to a BQ and cross-validate the scoring key on a comparable hold-out group. Although this BQ was a valid predictor of performance on the FST, a cognitive criterion, it should be emphasized that this BQ is valid only for that criterion. Whether this BQ could effectively predict other job performance criteria like tenure, training school scores or supervisory ratings has not been investigated.

Besides validity and adverse impact considerations in the use of the BQ in selection, there are several other concerns that warrant attention prior to its development and use. In the public sector it is important that items be job-related and not invade the privacy of applicants. Often, however, items that are not especially job-related are strong predictors of job performance criteria. Another concern which is closely related to the requirement that BQ questions have face validity, is the issue of fakability. When BQ items make obvious inquiries about job characteristics or personal motivations, there is some possibility that applicants will respond in a socially desirable manner to increase their chances of employment. The issue of fakability is not as great of a concern as some of the previous issues, because the scoring key for BQs is empirically developed and hence accuracy of biodata information is not critical.

Table 1

Comparison of Firefighter Applicants by Race on Selected Background Variables

| Background Variable | Total Sample | Black | | White | |
|---|---|---|---|---|---|
| Race | | | | | |
| Number | 447 | 356 | | 91 | |
| Percent | 100 | 79.6 | | 20.4 | |
| Hispanic (number) | 5 | 4 | | 1 | |
| Resident (percent) | 71 | 88 | | 4 | |
| Sex (percent) | | | | | |
| Male | 95 | 94 | | 99 | |
| Female | 5 | 6 | | 1 | |
| Education (percent) | | | | | |
| Less than high school | 5 | 5 | | 7 | |
| High school graduate/GED | 13 | 13 | | 11 | |
| Some college or technical training after high school; no degree | 4 | 3 | | 9 | |
| Two year degree, diploma | 36 | 37 | | 33 | |
| More than two years of college but no degree | 41 | 41 | | 40 | |
| Bachelor's degree or more | 1 | 1 | | 1 | |
| FIREFIGHTER SELECTION TEST * | | Mean | S.D. | Mean | S.D. |
| Reading Comprehension | | 47.82 | 9.52 | 58.53 | 6.74 |
| Using Formulas | | 48.05 | 9.55 | 57.62 | 7.87 |
| Judgment | | 47.88 | 9.52 | 58.29 | 7.13 |
| Reasoning | | 48.11 | 9.59 | 57.41 | 7.92 |
| Problem Identification | | 48.02 | 9.80 | 57.74 | 6.38 |
| Follow Directions | | 48.22 | 9.28 | 56.96 | 9.71 |
| FST Total (T score) | | 47.57 | 9.24 | 59.50 | 6.63 |

* T-scores standardized on total biodata sample with a mean= 50 and S.D. = 10
Note: All Black and White subgroup differences on the FST are significant at the .01 level.

Table 2

Description of the Developmental Group Sample (Profile)

| Item Number | Variable | Percent Total Sample | Percent Blacks | Percent Whites |
|---|---|---|---|---|
| 6 | WORK | | | |
| | Did not work | 18 | 20 | 8 |
| | Worked when not in school | 19 | 21 | 11 |
| | Less than 20 hours week | 18 | 16 | 23 |
| | 20 to 29 hours week | 31* | 29 | 41 |
| | 30 to 39 hours week | 9 | 9 | 8 |
| | 40 hours or more week | 6 | 6 | 8 |
| 8 | THOUGHT FF | | | |
| | Prior to 12th. grade | 40 | 34 | 66 |
| | 12th. grade or later | 60** | 66 | 34 |
| 9 | EXPERIENCE FF | | | |
| | None | 67** | 79 | 25 |
| | Less than one year | 11 | 12 | 7 |
| | One to five years | 12 | 6 | 31 |
| | More than five years | 10 | 3 | 38 |
| 10 | FRIENDS/RELATIVES FF | | | |
| | None | 30 | 37 | 2 |
| | One | 19 | 23 | 3 |
| | Two | 13 | 15 | 7 |
| | Three or more | 38** | 25 | 89 |
| 68 | PARTICIPATION IN CLUBS/ ORGANIZATIONS | | | |
| | Do not belong to any | 24** | 29 | 8 |
| | Not an active member of most | 6 | 6 | 3 |
| | Reliable member, but don't hold important position | 20 | 19 | 23 |
| | Like to hold office | 19 | 16 | 32 |
| | Have held one elected office | 20 | 20 | 22 |
| | Have held important offices in most | 11 | 10 | 12 |

Note: Percents do not always add to 100 due to rounding. Significance levels indicated are for the magnitude of the differences between Black-White subgroups. *= $p < .05$ **= $p < .01$ FF= firefighting/firefighter

Table 3

Percentage of the Developmental Group Sample, Whites, and Blacks Responding Positively to Biographical Questions

| Item Number | Variable | Percent Total Dev. Group | Percent Blacks | Percent Whites |
|---|---|---|---|---|
| PERFORMANCE OF ACTIVITIES | | | | |
| 14. | Plumbing | 41* | 38 | 52 |
| 17. | Carpentry | 57** | 53 | 74 |
| 21. | Landscape | 47** | 42 | 66 |
| PARTICIPATION IN RECREATION | | | | |
| 25. | Play Musical Instrument | 47* | 51 | 33 |
| 29. | Team Sports | 89* | 91 | 82 |
| 30. | Coaching/Officiating | 41* | 46 | 21 |
| 31. | Individual Sports | 83** | 87 | 66 |
| 39. | Cook | 93** | 96 | 82 |
| 40. | Attend Plays/Concerts/Movies | 88* | 90 | 80 |
| VOLUNTEER ACTIVITIES | | | | |
| 42. | Coach Youth Sports | 34** | 40 | 13 |
| 44. | Theater/Music Group | 25** | 30 | 5 |
| 46. | Big Brother/Big Sister | 17* | 19 | 7 |
| 48. | Counseling/Hotline | 15* | 17 | 5 |
| 49. | School Volunteer | 24** | 28 | 8 |
| ATTRACTIVENESS OF LISTED ASPECTS OF WORK | | | | |
| 53. | Work with People | 90* | 88 | 97 |
| 61. | Physical Work | 81* | 79 | 85 |
| 63. | Dangerous Situations | 65** | 62 | 83 |
| 64. | Emergencies | 68** | 64 | 84 |
| 65. | Heights | 44* | 44 | 56 |
| 66. | Dark | 34** | 31 | 47 |
| 67. | Closets | 30** | 26 | 42 |

Note: Positive responses to items were considered to be those ratings that were "Somewhat Attractive" and "Extremely Attractive". Discrepancies in the data are due to rounding and differences in the number of applicants responding to each question. * $p < .05$ and ** $p < .01$ for Black-White subgroup differences.

Table 4

Intercorrelations of Selected "Work Attractiveness" Biodata Variables

| Variable | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Security | .64 | .47 | .54 | .49 | .44 | .31 | .30 | .15 | .40 | .51 | .40 | .49 | .39 | .44 | .43 | .35 | .35 |
| 2. Pay | — | .65 | .67 | .58 | .55 | .45 | .47 | .20 | .41 | .69 | .52 | .57 | .46 | .47 | .46 | .31 | .37 |
| 3. Serve | | — | .77 | .58 | .52 | .32 | .49 | .00 | .29 | .69 | .54 | .62 | .45 | .49 | .38 | .25 | .28 |
| 4. People | | | — | .66 | .58 | .34 | .47 | .15 | .38 | .75 | .58 | .66 | .56 | .57 | .47 | .34 | .38 |
| 5. Equipment | | | | — | .63 | .35 | .52 | .20 | .39 | .64 | .60 | .51 | .46 | .58 | .54 | .44 | .40 |
| 6. Supervision | | | | | — | .38 | .41 | .29 | .43 | .56 | .53 | .54 | .45 | .54 | .53 | .53 | .51 |
| 7. Free Time | | | | | | — | .39 | .20 | .24 | .37 | .23 | .33 | .27 | .25 | .29 | .25 | .22 |
| 8. Prestige | | | | | | | — | .14 | .31 | .49 | .37 | .50 | .38 | .36 | .38 | .28 | .24 |
| 9. Risk | | | | | | | | — | .45 | .13 | .20 | .19 | .29 | .32 | .37 | .41 | .41 |
| 10. Shifts | | | | | | | | | — | .41 | .44 | .48 | .52 | .59 | .62 | .59 | .58 |
| 11. Opportunity | | | | | | | | | | — | .63 | .67 | .54 | .55 | .46 | .35 | .42 |
| 12. Physical Work | | | | | | | | | | | — | .69 | .55 | .61 | .52 | .46 | .48 |
| 13. Responsibility | | | | | | | | | | | | — | .62 | .61 | .51 | .40 | .46 |
| 14. Dangers | | | | | | | | | | | | | — | .74 | .64 | .50 | .52 |
| 15. Emergencies | | | | | | | | | | | | | | — | .69 | .62 | .60 |
| 16. Heights | | | | | | | | | | | | | | | — | .71 | .67 |
| 17. Dark | | | | | | | | | | | | | | | | — | .82 |
| 18. Closets | | | | | | | | | | | | | | | | | — |

Note: All correlations are significant at the .01 level, except Risk (9) with Security (1), People (4), Prestige (8) and Opportunity (11), which are significant at the .05 level. In addition, Risk (9) with Serve (3) is non-significant.

Table 5

Intercorrelations of Selected Biodata Variables and Part and Total FST Scores for the Developmental Group

| Biodata Variables | FST TEST SCORES | | | | | | |
|---|---|---|---|---|---|---|---|
| | RC | UF | J | LS | PI | OD | Total |
| Race | 41 | 39 | 42 | 37 | 39 | 35 | 47 |
| Residence | -35 | -33 | -37 | -32 | -39 | -31 | -42 |
| Education | 19 | 25 | 23 | 15* | 22 | 19 | 25 |
| Subjects 1) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Rank | -14* | -15* | -17 | -18 | -15* | -25 | -21 |
| High School Size | 15* | 0 | 17 | 0 | 19 | 0 | 16 |
| Thought FF 2) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Experience FF | 26 | 23 | 25 | 27 | 24 | 24 | 30 |
| Friends FF | 27 | 26 | 32 | 29 | 27 | 31 | 35 |
| Physical Exercise | 27 | 29 | 23 | 29 | 23 | 25 | 30 |
| Home Repair | 14* | 0 | 16 | 0 | 0 | 0 | 13* |
| Paint | 21 | 22 | 22 | 23 | 24 | 18 | 27 |
| Carpenter | 16 | 0 | 24 | 15* | 20 | 12* | 21 |
| Landscape | 12* | 0 | 13* | 14* | 20 | 0 | 17 |
| Car Repair | 18 | 15* | 24 | 0 | 24 | 0 | 21 |
| Coach | -18 | 0 | -15 | 0 | -18 | 0 | -16 |
| Relax Socially | 0 | 0 | 0 | 11* | 17 | 0 | 14* |
| Read 1) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Radio | 0 | 12* | 0 | 0 | 14* | 0 | 12* |
| Volunteer Theater | 0 | -14* | 0 | 0 | -13* | 0 | -12* |
| Volunteer Big Brother | -17 | 0 | 0 | 0 | -15* | 0 | -15* |
| Volunteer Hotline | -14* | 0 | -13* | 0 | -16 | 0 | -16 |
| Volunteer School | -20 | -18 | -19 | -16 | -22 | 0 | -22 |
| Security | 14* | 16 | 20 | 12* | 15* | 17 | 19 |
| Pay | 0 | 0 | 0 | 13* | 0 | 0 | 12* |
| Serve | 24 | 20 | 20 | 20 | 22 | 14* | 24 |
| People | 23 | 22 | 21 | 20 | 21 | 16 | 25 |
| Equipment | 19 | 19 | 20 | 18 | 17 | 14* | 21 |
| Prestige | 31 | 22 | 28 | 19 | 24 | 15* | 29 |
| Risk 2) | 0 | 0 | 0 | 0 | -17 | 0 | 0 |
| Shifts 2) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Opportunity | 21 | 22 | 18 | 23 | 20 | 20 | 25 |
| Physical Work | 16 | 16 | 0 | 16 | 0 | 14* | 15 |
| Responsibility | 20 | 14* | 15* | 15* | 15* | 0 | 18 |
| Dangers | 17 | 13* | 18 | 12* | 15* | 0 | 18 |
| Emergencies | 18 | 13* | 16 | 14* | 0 | 15* | 17 |
| Heights 2) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Dark 2) | 12* | 0 | 0 | 0 | 0 | 0 | 0 |
| Closets 2) | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Participation | 26 | 23 | 27 | 27 | 25 | 25 | 31 |

* Correlations significant at .05 level; all other correlations except for zeros are highly significant at .01 level.

1)= Biodata variables that have a cognitive component.

2)= Biodata variables that are job-related but did not correlate with the criterion.

## References


Asher, J.J. The biographical item: Can it be improved? Personnel Psychology, 1972, 25 (2), 251-269.

England, G.W. Development and use of weighted application blanks (Rev. Ed.). Minneapolis: University of Minnesota Industrial Relations Center, Bulletin No. 55, 1971.

Hsu, L., Darany, T.S., & Nettles, S., Cooperative Clerical Test Validation Study in Delaware. Newark, DE: Delware Public Administration Institute, University of Delaware, 1978.

Owens, W.A. Background data. In Dunnette, M.D. (Ed.) Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Owens, W.A., Glennon, J.R., & Albright, L.E. A catalog of life history items. Greensboro, North Carolina: Richardson Foundation, 1966.

Payne, S.S., & van Rijn, P. Development of a Written Test of Cognitive Abilities for Entry in the D.C. Fire Department: The Task-Ability-Test Linkage Procedure (Technical Memorandum 78-5). Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, 1978.

van Rijn, P. Biographical questionnaires and scored application blanks in personnel selection. Washington, D.C.: Personnel Research and Development Center, U.S. Office of Personnel Management (PRR-80-31), 1980.

AD P001361

O'Mara, Francis E., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Wed. A.M.)

## Battalion Effectiveness and Command Climate

It has long been known that effective Army units have high morale and favorable command climates. What has not been known, however, and what has remained a matter of debate, is how this comes about. Does command climate contribute to unit effectiveness or is unit effectiveness a necessary antecedent to a favorable command climate? Or, are command climate and unit effectiveness concurrent products of some third organizational attribute, such as the characteristics of the unit commander. The Command Climate Project was begun in February, 1978 to address these questions. Specifically, this research was aimed at defining the causal dynamics between command climate and morale on the one hand and battalion effectiveness on the other.

However, before command climate, or anything could be related to unit effectiveness, the Army had to understand how to define and measure this effectivenss. Thus a necessary prerequisite for examining the climate-effectiveness relationship was the measurement of unit effectiveness.

Approximately 55 battalions located at six CONUS installations were studied over two years (some measures beyond three years). At each of four waves (May and November of 1978 and of 1979), three types of data were collected on each test battalion: (1) Unit Performance/Readiness; (2) Command Climate Survey Measures; (3) Interviews with Senior Commanders.

Analysis of the results and implications for the assessment and diagnosis of unit effectiveness is presented and diagnosed.

# THE MEASUREMENT OF ARMY BATTALION PERFORMANCE

Francis E. O'Mara
Advanced Technology Incorporated

One of the most central but most perplexing issues in the study of organizations in general, and the study of military organizations in particular, is that of defining valid criteria of organizational effectiveness. This is a central issue inasmuch as any theory of organizational functioning must be validated by its successful prediction of organizational performance. Obviously, the strength of such a validation rests on the degree to which organizational performance is precisely yet comprehensively defined and measured. In this way, the quality of the theoretical and operational definition of organizational effectiveness defines the upward limit of the growth of organizational and military psychology.

In light of the importance of this issue, it is not surprising that a voluminous literature has developed around it. As far back as 1959, over 370 references could be cited as relevant to organizational effectiveness (Wasserman, 1959). In the intervening years, this literature has swelled further to impressive proportions (Goodman and Pennings, 1979). In an attempt to integrate this literature, Campbell, Bownas, Peterson, and Dunnette (1974), have identified two major approaches to this issue. The first of these, the Goal Model, assumes that there are clearly specified short and/or long-term goals for an organization. The assessment of organizational effectiveness then requires a determination of the organization's specific goal or mission, followed by the development of measures which reflect the degree to which these goals are met. Familiar examples of this approach would include cost/benefit analysis and a management-by-objectives approach to organizational effectiveness. The System Model approach, on the other hand, proceeds on the assumption that the goals of any sizable organization are so numerous, complex and dynamic that it is not possible to clearly define a small number of unambiguous, stable, and measurable goals. Rather, the best index of an organization's effectiveness is its general "health" (i.e., the ability of the organization to sustain its capabilities and operations without exhausting resources). "Good health" permits it to continually accomplish its goals, regardless of the nature of these goals or how they may change. Through this model, the assessment of organizational effectiveness does not proceed through a determination of goal attainment but rather through the direct measurement of the organization's "health." Thus, those subscribing to the Systems Model have emphasized organizational processes rather than organizational outcomes.

While this dichotomization of approaches has done much to provide coherence to a burgeoning literature, when it is applied to a military environment it can be seen that both models are simultaneously employed. The overall peacetime goal of Army units is to maintain their readiness to engage successfully in combat operations. In this way, the criterion of organizational effectiveness is singular and defined. Organizational

effectiveness is simply the extent to which the military unit maintains conditions which support its maximum performance in combat (e.g., equipment levels, equipment readiness, or the competence of unit personnel in their combat duties). In line with the Goal Model the Army devotes considerable resoures in monitoring such conditions and reporting their status through the Unit Status Report [USR].

The singular unit objective of combat readiness, however, when operationalized on a day-to-day basis, is transformed into a broad spectrum of taskings and requirements. In this way the sense of an orderly and integrated composite of clearly specified, prioritized and unchanging objectives breaks down at the unit level to a condition in which a multitude of often conflicting demands are placed upon the unit. Further, the unit must meet these demands in the fact of shifting priorities and with changing resource levels. In this environment, the unit's flexibility and capability to quickly respond to a wide variety of short-term objectives becomes a truer indication of its effectiveness. A Systems Model approach would therefore be more appropriate to assess the unit's generalized capability of meeting indeterminate objectives. Many measures of unit effectiveness employed by the Army today are congruent with the Systems Model approach. Most of these measures fall into the category of measures usually referred to as "Command Indicators" or "Traditional Indicators," e.g., rates of Absence Without Leave [AWOL], reenlistments, or courts-martial. These measures have been traditionally emphasized in the military because it is felt that they reflect the state of morale and discipline in the unit which in turn are thought to support the unit's general capabilities.

Thus, both models of organizational effectiveness are employed currently by the Army. The Goal Model is used in evaluating the accomplishing of the long-term goal of combat readiness while the unit's capability to meet varying day-to-day objectives is indirectly assessed through a Systems Model approach. In terms of contemporary models of oraganizational effectivensess then, the approach currently used to assess Army unit effectiveness is one which is quite complete.

The literature is far less supportive of the validity of the Army approach to unit effectiveness assessment. Many questions have been raised regarding the empirical adequacy of current Army unit effectiveness assessment procedures, particularly with regard to those involved with USR (Robinson, 1980; Ross, et. al., 1979; Sorley, 1979; Sorley, 1980; U.S. Army Concepts Analysis Agency, 1975; U.S. Army War College, 1976). The magnitude of the deficiencies identified with the USR can be best summarized in one of the many findings of the Army War College (1976) study on unit effectiveness reporting. Of the approximately 2,100 Army personnel surveyed as a part of this study a full 70 percent

reported that the Unit Status Report does not reflect the true readiness condition of a unit. The particular deficiencies in this system which had been advanced to explain its lack of manifest validity have included: (1) the subjectivity of measures--it is felt that there is a substantial degree of latitude for subjective interpretation of unit conditions that it is permitted in filling out the USR. For example, the estimate of training readiness, one of the major components of unit readiness measured on USR, is based upon the unit commander's estimate of the number of weeks of training the unit would need to be fully ready for combat. It is felt that in light of the pressures on commanders to maintain their units at maximum readiness, this unstubstantiated estimate is likely to be overly optimistic, (2) statistics management--evidence was found in the War College study to suggest considerable pressure being placed upon unit personnel to have the units portray a maximally positive readiness condition on the USR even to the extent of overlooking genuine unit deficiencies, (3) standards--it has been stated that the standards employed for determining readiness on the USR are either too lax (Sorley, 1979), too strict (U.S. Army War College, 1976), or inconsistent across branches of the Armed Forces (US Army War College, 1976), (4) complexity of procedures--the USR has been criticized as entailing complex procedures that are not always explained fully or well in the supporting regulations. Accordingly, there is a higher probability of errors being made in reporting unit readiness.

In addition to the measures reported on the USR, other factors employed in assessing unit effectiveness, particularly the command indicators, have been subjected to scrutiny and criticism. Unlike the USR measures, these indices are not systematically reported to the higher echelons of the Army command structure. However, unit measures on these variables are used quite frequently at the local level as indicants of unit conditions and problems. Sorley (1979) has been critical of the use of such measures inasmuch as he sees them leading to a "management by statistics" in which those factors, which are more readily quantifiable, are given greater command emphasis than those which more substantively support and reflect unit effectiveness but which are less readily measured. Too often, he feels, command attention is expended on "getting the numbers right" in such areas of questionable military value as motor vehicle accidents or letters of indebtedness among unit personnel at the expense of diverting command attention from such areas as unit training and equipment maintenance. The position underlying his assertions is that statistical indices of unit operations, particularly those relevant to the personnel area, are of questionable utility in assessing areas pertinent to unit effectiveness. Clearly, some of these statistical measures are more germane to unit effectiveness than are others. What is needed is an empirical

determination of the relative value of each of these measures for assessing unit effectiveness. In the absence of this, it is left to individual opinion as to which of the wide variety of possible measures are true indicants of unit capability. Such a condition can only lead to a proliferation of measures on which units are assessed and judged but which are of little or no value as true measures of unit effectiveness.

The purpose of this research is to examine the validity of the most commonly employed measures of Army unit effectiveness. To accomplish this, the interrelationship among variables that purport to measure the same organizational construct (e.g., combat readiness, morale) will be examined to determine the concurrent validity of these measures.

## METHODS

### SAMPLE

Measures of unit effectiveness were taken from a sample of 71 battalions located in USAREUR and CONUS. These battalions constituted a representative sample of Combat Arms, Combat Support, and Combat Service Support battalions.

### Measures

The measures of unit effectiveness which were employed in this research fell into three major categories: Direct Readiness measures, Command Indicators, and Personal Judgements. The specific measures used in the first two categories are displayed in Figure 1 and Figure 2, respectively. The Personal Judgements measures consisted of estimates of battalion effectiveness by the Division Commander, the Assistant Division Commander, and the Brigade Commander above the battalion in the chain of command and of the service members, Non-Commissioned Officers [NCOs], and officers within each of the battalions. The judgements of the Division Commander, Assistant Division Commander, and Brigade Commander were collected in the course of an interview with each of these individuals. In the course of these interviews these commanders were asked to assess the effectiveness of each of the sample battalions using a 13-point rating and also to rank each battalion relative to the other battalions in the rater's command. The rating and the ranking were each converted to standard scores and then combined into a single battalion effectiveness score for that rater. Estimates of battalion effectiveness were collected from service members, NCOs, and officers within each of the battalions in the course of a survey administered to unit personnel. Responses to each of three items regarding overall battalion effectiveness were averaged for each individual. These scores were then aggregated to produce an average SM, an average NCO, and an average officer estimate of battalion effectiveness for each of the battalions.

| | |
|---|---|
| OVERALL READINESS | A battalion's overall readiness status as reported in the monthly Unit Status Report. |
| PERSONNEL READINESS | A battalion's personnel readiness status as reported in the monthly Unit Status Report. |
| EQUIPMENT ON HAND | An index of the degree to which a battalion possesses all authorized equipment, a reflection of the battalion's supply system. |
| EQUIPMENT SERVICEABILITY | The maintenance status of a battalion's equipment, a reflection of the battalion's maintenance system. |
| EQUIPMENT ON HAND RATED RDY | The proportion of equipment a battalion actually has on hand that is operational. |
| ARTEP | The percentage of the missions/ tasks rated "satisfactory during a battalion's most recent field training exercise. |
| AGI | The percentage of the areas rated "satisfactory" during a battalion's most recent annual general inspection. |

FIGURE 1

READINESS MEASURES

| | |
|---|---|
| ARTICLES 15 | The percentage of enlisted personnel administered nonjudicial punishment (e.g., fines, reductions in grade) during a given month. |
| COURTS MARTIAL | The percentage of enlisted personnel receiving a court martial during a given month. |
| AWOL | The percentage of enlisted personnel who were involved in unexcused absences during a given month. |
| DESERTIONS | The percentage of enlisted personnel who deserted during a given month. |
| FIRST TERM RE-UP | The percentage of a battalion's first-term reenlistment objective that was achieved in a given month. |
| CAREER RE-UP | The percentage of a battalion's reenlistment objective for career personnel that was achieved in a given month. |
| CRIMES OF VIOLENCE | The percentage of a battalion's enlisted strength involved in crimes of violence in a given month. |
| PROPERTY CRIMES | The percentage of a battalion's enlisted strength involved in crimes against property in a given month. |
| DRUG ARRESTS | The percentage of a battalion's enlisted strength arrested for drug and marijuana violations in a given month. |

FIGURE 2

COMMAND INDICATORS

## Procedure

Data were collected in the course of four data waves in May and November of 1978 and 1979. In Wave 1, Direct Readiness and Command Indicator data were collected for the preceding five quarters, while in the subsequent waves these data were collected only for the intervening time periods. The Personal Judgements data were collected in the course of interviews and surveys conducted during each data wave.

## RESULTS

Pearson correlation coefficients were computed among the measures constituting each of the three groups. It was anticipated that the measures in each of the three groups would be significantly intercorrelated inasmuch as the same construct. That is, Director Readiness measures purportedly reflect direct assessments of the unit's capability to perform its mission while Command Indicators assess the degree of morale and discipline in the unit and the Personal Judgements measures reflect the reputation of the unit for effective performance.

The correlations among the Direct Readiness measures are displayed in Table 1. As shown, agreement of the USR measures with ARTEP and AGI results was restricted to USR measures pertaining to the maintenance status of equipment. In contrast to this, there was nearly universal agreement among the various USR measures. This agreement suggests that a substantial component of the variance of the USR measures in attributable to a general factor which supports high REDCON ratings in all areas. A possibility exists that this general factor is, at least in part, attributable to the priority given to the unit in allocating resources, that is, the unit's position on the Department of the Army Master Priority List [DAMPL]. This possibility is suggested by the differences in the significant correlations which the Equipment Serviceability REDCON rating shares with the other REDCON ratings in contrast to the lack of significant correlations between the Equipment Readiness measure and these REDCON ratings. The difference between the Equipment Serviceability REDCON and the Equipment Readiness measure is that the former measure reflects the condition of the unit's total allocated equipment while the latter reflects the condition of the equipment that the unit actually has on hand. Thus, the portion of variance which the Equipment Serviceability REDCON shares with the other REDCON ratings reflects the extent to which this measure is influenced by the units being supplied with its required resources while the variance it shares with the Equipment Readiness measure reflects the degree to which the unit successfully applies its level of resources to maintaining its equipment.

TABLE 1

MEAN CORRELATIONS AMONG DIRECT READINESS MEASURES
OF BATTALION EFFECTIVENESS

| | ARTEP | AGI | CO | CPER | CEOH | CESC | CTRN | WKS | EQRD |
|---|---|---|---|---|---|---|---|---|---|
| ARTEP | 1.00 | | | | | | | | |
| AGI | .19 | 1.00 | | | | | | | |
| Overall REDCON[2] (CO) | .27 | .16 | 1.00 | | | | | | |
| Personnel REDCON (CPER) | .13 | .06 | .69** | 1.00 | | | | | |
| Equipment on Hand REDCON (CEOH) | .02 | .04 | .48** | .11** | 1.00 | | | | |
| Equipment Serviceability REDCON (CESC) | .59** | .39** | .47** | .14** | .44** | 1.00 | | | |
| Training REDCON (CTRN) | .30 | .02 | .60** | .56** | .22** | .26** | 1.00 | | |
| Weeks to Readiness (WKS) | -.22 | .03 | .54** | .47** | .20** | .27** | .86** | 1.00 | |
| Equipment Readiness (EQURD) | .54** | .40** | .32** | .04 | .03 | .74** | .12** | -.10* | 1.00 |

* $p < .05$
** $p < .01$

[2]All REDCON ratings have been scored such that higher scores indicate greater readiness.

Therefore, the findings that both Equipment Serviceability REDCON and the Equipment Readiness measure are correlated with ARTEP and AGI performance while the other REDCON ratings are not indicate that it is the unit's ability to make the most of what it has which is the common element producing these correlations.

The mean correlations among the Command Indicators measures are exhibited in Table 2. There was a fair degree of agreement among these measures as 29 of the 55 mean correlations were significant. However, with a few exceptions, these significant means correlations were too small to support the interpretation of there being a single dimension underlying these measures. If one were to use a minimum cutoff value of only .20 as a criterion of practical utility, then only 9 of the 55 correlations would meet or exceed this value. Of these nine, one (between AWOL rate and Desertion rate) is preordained by the operational definitions of these measures, four (AWOL rate, Desertion rate, and Crimes Against Property rate with rate of Articles 15, and Drug Arrest rate with Courts-Martial rate) appear to describe linkages between specific violations of Army regulations and the punishment typically administered in response to team, while the remaining four mean correlations (between Articles 15 rate and Adverse Discharge rate, between First-Term Reenlistment rate and Career Reenlistment rate, and between Crimes Against Property rate and both Crimes of Violence rate and Drug Arrest rate) constitute agreement among measures tapping into similar functional ares of unit operation.

Considering the directionality of these significant mean correlations, it can be seen that the nature of the interrelationships among these variables is even more complex. First-Term Reenlistment rate, for example, correlates positively with rate of Adverse Discharges, Courts-Martial, Drug Arrests, and Career Reenlistment, yet negatively with rates of AWOLs, Desertions, Crimes Against Property, and Crimes of Violence. There is, thus, no simple pattern in which such a "positive" measure at First-Term Reenlistment rate is positively associated with other "positive" measures and inversely related to "negative" measures (e.g., Drug Arrest rate).

In summary, the correlations shown in Table 2 indicate that these measures are not alternative measure of a single underlying construct such as morale or unit effectiveness. Indeed, there appears to exist some trade-offs among the facets of unit operation reflected in these measures, so that high scores on some negative measures are associated with desirable unit outcomes (e.g., reenlistments).

The correlations among the Personal Judgments measures of battalion effectiveness can be seen in Table 3. In contrast to the other two groups of measures, there is a substantial degree

TABLE 2

MEAN CORRELATIONS AMONG COMMAND INDICATOR MEASURES
(11 Quarters, N = approx. 400 battalion-level observations)

| | EDP | ADVRS | ART | C-M | AWOL | DFR | FREUP | CREUP | COV | PROP | DRUG |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Expeditions Discharge (EDP) | 1.00 | | | | | | | | | | |
| Adverse Discharge (ADVRS) | .09 | 1.00 | | | | | | | | | |
| Articles 15 (ART) | .07 | .29** | 1.00 | | | | | | | | |
| Courts-Martial (C-M) | .08 | .03 | .11* | 1.00 | | | | | | | |
| AWOL | .15* | .09 | .26** | .07 | 1.00 | | | | | | |
| Desertions (DFR) | -.02 | .07 | .23** | .10 | .50** | 1.00 | | | | | |
| First-Term Reenlistment (FREUP) | .03 | .19** | .01 | .17** | -.10 | .09 | 1.00 | | | | |
| Career Reenlistment (CREUP) | .16** | .02 | .04 | .15** | .00 | .05 | .41** | 1.00 | | | |
| Crimes of Violence (COV) | .06 | .03 | .12** | .05 | -.01 | .02 | -.09* | .03 | 1.00 | | |
| Crimes Against Property (PROP) | .02 | .00 | .20** | -.01 | .14** | .09 | -.10 | .00 | .25** | 1.00 | |
| Drug Arrests (DRUG) | .18** | .02 | .05 | .23** | .03 | .02 | .15** | .14** | .16** | .42** | 1.00 |

* $p < .05$
**$p < .01$

TABLE 3

MEAN CORRELATION AMONG PERSONAL JUDGEMENTS
OF BATTALION EFFECTIVENESS (4 WAVES)

| | CG | ADC | BDE | OFF | NCO | SM |
|---|---|---|---|---|---|---|
| External Perceptions | | | | | | |
| Division Commander (CG) | 1.00 | | | | | |
| Ass't Division Commander (ADC) | .64** | 1.00 | | | | |
| Brigade Commander (BDE) | .44** | .50** | 1.00 | | | |
| Internal Perceptions | | | | | | |
| Officers (OFF) | .34** | .28** | .35** | 1.00 | | |
| NCOs | .10 | .10 | .33** | .55** | 1.00 | |
| Service Members (SM) | .16* | .17* | .29** | .26** | .26** | 1.00 |

* $p < .05$
**$p < .01$

of correlation among these independent estimates of battalion effectiveness. The extent of this agreement, however, is a function of the proximity in the chain of command of the individuals providing the judgment of effectiveness. Thus, there appears to be a perspective on battalion effectiveness which varies gradually across command echelons, such that individuals in intermediate echelons (i.e., Brigade Commanders, Battalion Officers) partially share the perspective of both those above and those below them in the chain-of-command while individuals in more extreme echelons seem to hold widely disparate views regarding what constitutes unit effectiveness.

## DISCUSSION

Of all the findings of the present research, that which is most serious in its consequence is the meager concurrent validity among the direct measures of unit mission capability, i.e., the Direct Readiness measures. The lack of concurrence between ARTEP and AGI results is somewhat understandable in light of the different emphases of these two evaluations. That is, ARTEP exercises strongly emphasizes the tactical proficiency of the unit while the AGI has a much stronger emphasis on garrison activities and procedures. Further, there is likely to be a considerable period of time separating a unit's ARTEP from its AGI. In the present results this time period was permitted to extend to as much as six months. It is not at all inconceivable that in the course of the time period separating these two evaluations a battalion could substantially increase or decrease in its overall effectiveness.

More troublesome than the lack of agreement between the ARTEP and AGI results in the rather spotty agreement between these two measures and those reported on the USR. Here, it was seen that the Overall REDCON measure--the "bottom line" on the USR--bore no relationship to battalions' effectiveness as measured by ARTEPs or AGIs. Further, the Training REDCON, which should have been the REDCON most closely aligned with performance on an ARTEP, a training evaluation, was not at all associated with it. These findings serve to support earlier criticisms of the USR (Robinson, 1980, Ross, et. al., 1979; U.S. Army Concepts Analysis Agency, 1975; U.S. Army War College, 1976; Sorely, 1979). However, the present findings suggest that to the degree that the Equipment Serviceability REDCON reflects not a battalion's level of allocated resources, but rather the battalion's ability to effectively apply its available resources in maintaining its equipment, it can constitute an acceptable measure of battalion effectiveness. This finding would therefore support a recommendation that the Equipment Serviceability REDCON could be improved if it were to be based on the percentage of the equipment that the unit actually has on hand which is ready rather than on a percentage of allocated equipment which is ready.

Even though Command Indicators have long been employed in the military as indices of overall unit morale, there is no evidence in the present results to suggest the presence of any global factor underlying these measures. Rather, these results show that the relationship among these variables are quite complex so that, for example, some "positive" indices (reenlistment) are directly related to some "negative" measures (e.g., Courts-Martial). Such findings clearly show the danger of evaluating the effectiveness of a battalion, or a battalion's command, by reference to a single score on a single measure without consideration being given to the dynamics and conditions which fostered that score. That is, these findings are an indictment of "management by statistics" approach to monitoring and maintaining unit effectiveness. A unit's high or low score on any of these Command Indicators cannot be taken as a sign of unit morale or discipline without additional information about why that score is at that level.

The relatively high degree of consensus along the professional judgments of individuals familiar with the unit suggests that such "soft" measures of unit effectiveness would be useful adjuncts to "hard" unit effectiveness measures in providing a total picture of unit capability. The broad consentual base which extends up to and includes the division commander shows that such an addition is not required at the local level. It is at the higher echelons where critical long-range decisions are made, with a heavy reliance on standard statistical indices of unit effectiveness, indices which the present results indicate have definite shortcomings. Future efforts should be expended to develop and evaluate methods for systematically and validly providing the potentially rich source of information which can be gleaned from judgments of military professionals familiar with the unit. The present results, in conjunction with related research (Kerner-Hoeg and O'Mara, 1981), indicates that this data should include the input of unit officers and NCOs at a minimum.

In conclusion, the results of this research indicate that the estimation of military unit effectiveness is an area which has some promise but a promise that is not presently being realized to a substantial degree. The findings support the position of neither the most optimistic supporter nor the most pessimistic critic of current methods of effectiveness assessment. Rather, they indicate that there are several shortcomings in current systems which can be fruitfully addressed at present while in other areas, ameliorative efforts must await a fuller articulation of the dynamics of military unit operation.

## REFERENCES


Barzily, A., Catalogne, P.R., & Marlow, W.H. Assessing Marine Corps Readiness (T-430). Washington, D.C.: The George Washington University, September 1980.

Bowser, S.E. Determination of Criteria of Operational Unit Effectiveness in the U.S. Navy (NPRDC TR 76TQ-41). San Diego, California: Navy Personnel Research and Development Center, August 1976.

Campbell, J.P., Bownas, D.A., Peterson, N.G. & Dunnette, M.D. Measure of Organizational Effectiveness: A Review of Relevant Research and Opinion. Final report to Office of Naval Research, under Contract N00022-73-C-0023, Minneapolis, Minnesota, 1974.

Kerner-Hoeg, S.E. & O'Mara F.E. Commanders' Assessment of Unit Effectiveness Measures. Paper presented at the 23rd Annual Conference of the Military Testing Association, Arlington, Va., 1981.

Robinson, R.M. Objective Measurement of Training Readiness. Carlisle Barracks, Pennsylvania: U.S. Army War College, May 1980.

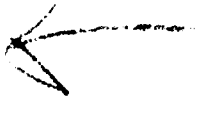
Ross, G., Murphy, J., March, M., Robinson & Tullington, B. Military Organizational Effectiveness/Readiness and Substainability. Final report to Headquarters, Department of the Army under Contract DAA1C21-79-0015, McLean, Virginia: Science Applications, Inc , September 1979.

Sorley, L. Professional Evaluation and Combat Readiness. Military Review, October 1979, 41-53.

Sorley, L. Prevailing Criteria: A critique. In Sarkesian, S.C. Combat Effectiveness--Cohesion, Strees, and the Volunteer Military. Beverly Hills: Sage Publications, 1980.

U.S. Army Concepts Analysis Agency, Readiness System Study, Phase I Analysis. AD-A029 387, August 1976.

U.S. Army War College, U.S. Army Unit Readiness Reporting. ACN 75025, June 1976.

AD P001362

Palmer, R. L., US Army Research Institute for the Behavioral and Social Sciences, Fort Hood, Texas. (Wed. A.M.)

## The Commander's Unit Analysis Profile (CUAP)

The CUAP questionnaire is a diagnostic tool for providing commanders of company-size units knowledge of enlisted attitudes related to such factors as Cohesiveness, Training, Leadership, Discipline, Job Satisfaction, Morale, Reenlistment, etc.

The questionnaire, which is completed in about 15 minutes, can be read by soldiers with minimal reading skills. Administration procedures require no special training and provide confidentiality for both respondents and commander. Only areas over which commanders exercise control are covered.

Timely, uncomplicated feedback is provided by two graphical unit-profiles. Profile 1 depicts for each factor the Unit Factor Score and the Average Score Other Units, which is the mean for all units recently utilizing the CUAP. Profile 2 depicts the Unit Percentile Rank for each factor.

The CUAP does not replace the commander's responsibility for judging the mission readiness of the unit; rather it identifies attitudinal proclivities that may detract from or contribute to overall operational effectiveness.

# The Commander's Unit Analysis Profile

R. L. Palmer

US Army Research Institute
Fort Hood, TX

## Purpose

The purpose of this report is to describe the Commander's Unit Analysis Profile project being conducted by the Army Research Institute's Field Unit at Fort Hood, Texas. The Commander's Unit Analysis Profile is a new leadership tool available to commanders of company-size units and their supervisors.

## Background

The Commander's Unit Analysis Profile project was begun in late 1978 under the sponsorship of Headquarters III Corps, located at Fort Hood, and the Army's Forces Command. At a basic level, the problem addressed by the project was that today's junior commanders--i.e., commanders of the Army's "rank and file,"--must contend with a variety of factors that tend to produce apathy, dissatisfaction, and disunity among soldiers, with the potential consequence of a degradation in the operational effectiveness of military units. The list of these factors is complicated and long, and it will not be discussed in detail here. Let it suffice to note that the list includes such subjects as the erosion of military benefits, the military leadership drain, drug usage, racial problems, problems associated with the increased numbers of women in the Army, inefficient and ineffective training methods and policies, attitudinal consequences of societal changes, and similar factors. Such factors are not, of course, all unique to today's Army; however, they have perhaps become more severe in recent years. The Commander's Unit Analysis Profile project does not address these factors directly, but deals instead with their consequences, insofar as they are reflected in the attitudes and performance of the "common" soldier.

The commanders of these soldiers--i.e., commanders of company-size units--are charged with maintaining the "mission readiness" of their units at all times regardless of the degree of apathy, dissatisfaction, or disunity that may exist in their units, and in spite of any personal problems among their soldiers. Thus, there is a need today for these leaders to be aware of the salient features of the attitudinal and social environments that prevail in their units. Until recently, however, there has been no satisfactory, standardized method for commanders to systematically identify problem areas in this realm with sufficient accuracy and specificity.

The Commander's Unit Analysis Profile project--known more succinctly (and picturesquely) as CUAP--addresses the problem described by making available to the Army a unit-diagnostic system that provides commanders of company-size units a working knowledge of troop attitudes about a variety of factors related to mission readiness and operational effectiveness, such as unit cohesiveness, morale, reenlistment potential, quality of training and leadership, etc. The system involves a simple, retribution-free procedure that yields substantial but concise information about the attitudinal environment of lower-ranking (E1 - E5) enlisted personnel.

## Development

Criteria. The nature of the CUAP system is portrayed by the list of developmental criteria that were established at the outset of the project and based in part upon suggestions and comments offered by Generals Robert Shoemaker and Marvin Fuller, who were successive commanders of III Corps and Fort Hood at the time. The criteria served the purpose of overcoming many problems commonly associated with surveys directed toward Army personnel. Not the least of these problems was that many questionnaires and surveys (administered by a variety of agencies) took troops away from their training missions and primary jobs during the administration period, but provided little compensation in the way of useful feedback to commanders, particularly at the company level. Furthermore, when feedback was provided, it was often so late in arriving that it was of little or no value to the commander. These and related concerns stimulated the following major developmental criteria associated with the CUAP project:

1. Administration of the questionnaire must involve minimal interference with normal troop training and work schedules; i.e., the instrument should be as short as feasible while maintaining its capacity to collect necessary and sufficient information.

2. The questionnaire must be easy to administer and the results easy to interpret, both without the assistance of specially-trained personnel.

3. Each questionnaire item must possess face-validity; i.e., its intent should be obvious, and there should be no so-called "double-meaning" items.

4. The questionnaire must be capable of being read by soldiers with minimal reading skills.

5. The questionnaire must be maximally sensitive to differences among company-size units; i.e., it should not contain items pertaining directly to battalion level, or higher, or items that all units tend to answer similarly.

6. The questionnaire should cover only subject areas over which the

small-unit commander can exercise significant influence.

7. The data format must facilitate rapid processing and timely feedback to the participating unit commanders (ideally, within 15 days).

8. The feedback must provide norms that permit commanders to compare their units with the combined results of all other units in the Army that have recently participated in the survey.

9. Anonymity must be provided for all questionnaire respondents, and confidentiality must be afforded all unit commanders who voluntarily request administration of the questionnaire in their units.

10. The feedback of questionnaire results should serve as a diagnostic tool for unit commanders to use in isolating and identifying factors that may be contributing to or detracting from unit operational effectiveness; but it should not be construed as providing an overall assessment of mission readiness, which should be based upon wider concerns and remain the responsibility of the unit commander.

Method. Development of the CUAP instrument was started with an original pool of 99 questionnaire items. These items were selected as a result of analyses of a variety of previously administered questionnaires and surveys: many of the items came from research questionnaires that had been used by the Army in other research programs; some were written especially for the CUAP. An attempt was made to tap all major topic areas considered related to the effective operation of company-size units and which lower-ranking enlisted personnel would find salient in their day-to-day lives. Of course, the topics also had to satisfy the criterion, specified earlier, that the unit commander must possess potential influence over the situation. Thus, for example, such topics as military pay were not addressed.

The 99 questionnaire items were formulated as interrogatives with 5-alternative, evaluative response scales, as in the example shown in Figure 1. The items were arranged intuitively into topic groups, or "factors," rather than randomly ordered--a procedure that could be viewed as possibly creating a response-set bias. However, for the

Does your company commander treat you with respect?

| | | |
|---|---|---|
| [+2] | _____ | Very often, or always |
| [+1] | _____ | Often |
| [ 0] | _____ | Sometimes |
| [-1] | _____ | Seldom |
| [-2] | _____ | Very seldom, or never |

Figure 1. Fictitious item, portraying CUAP questionnaire item format.

purposes at hand, it seemed desirable to focus the soldiers' evaluative processes on one general topic at a time and to allow the topics and the several questionnaire items pertaining to each topic to flow in a related sequence throughout the questionnaire.

The pilot version of the questionnaire was administered to 21 tank companies at Fort Hood during June and July of 1979--about 675 soldiers altogether. A statistical factor analysis of the collected data yielded 23 factors, which approximated the original intuitive grouping of the items into factor areas, as one might expect on the basis of the face validity inherent in the items. This factor analysis was used to eliminate items that loaded on two or more factors as opposed to one, and items that did not load reasonably heavily on any single factor. The purpose here was, of course, to eliminate overlap among topic areas and to rid the questionnaire of items that did not seem to contribute much to any particular area. Of course, the item composition of the factor areas was revised wherever the analysis conflicted with the original, intuitively laid-out structure, taking care to maintain the face validity of each item within each factor area. This step necessitated the throwing away or revising of a few items because they did not seem to "fit" where the analysis placed them.

Another analysis was carried out to eliminate items that would not distinguish among the 21 military units in the sample. The purpose of this analysis was to create an instrument whose primary purpose would be to measure differences among military units rather than their absolute standings. Thus, a one-way analysis of variance was conducted on the data from each of the 99 questionnaire items, and those items were eliminated for which there was not a statistically significant difference at the .01 alpha level among the 21 companies. A few items that failed to reach the .01 level in this analysis, but nevertheless showed promise, were retained, with revisions.

A third analysis was conducted to eliminate undue redundancy from the questionnaire. Here, the intercorrelation matrix for the 99 items was examined for the presence of high inter-item correlations, and where two items were highly correlated--say .70 or greater--the less desirable (either statistically or otherwise) was usually eliminated.

These analyses, plus a common-sense examination in which some items were eliminated because of unforeseen format inconsistencies and the like, left 63 items, which were factor analyzed again. In this analysis, several of the original factors collapsed together as a consequence of the item eliminations, leaving 13 factors. However, several of these factors were arbitrarily divided into two separate "factors" when their content provided a logical basis for the division and when there were reasons external to the analysis for maintaining such divisions. For example, "Sports Activities" and "Social Activities" appeared as a single factor in the analysis, yet the distinction was maintained because it was felt that such a division would be important to unit commanders. This procedure increased the number of factor areas to 23 again.

At this point, seven new questions were added to "round out" some of the factors, and rewording was accomplished wherever it seemed improvements in readability or clarity could be effected. Then a new version of the questionnaire was produced. During the following year the new version, CUAP-8004, was administered to approximately 3,850 soldiers in eight FORSCOM divisions, namely the 1st Cavalry and 2nd Armored Divisions at Fort Hood, the 5th Infantry Division at Fort Polk, the 82nd Airborne Division at Fort Bragg, the 101st Airborne Division at Fort Campbell, the 9th Infantry Division at Fort Lewis, the 7th Infantry Division at Fort Ord, and the 4th Infantry Division at Fort Carson.

During May and June 1981, the data collected with CUAP-8004 were analyzed in a manner that was essentially identical to the analytical procedure used for the original pilot version. These analyses, along with a year's experience with the 8004 version, led to several major and many minor revisions, the result being a new 96-item, pilot instrument for which the 8004 norms were not applicable. This new pilot instrument was administered to 30 companies at Fort Hood during late June and early July 1981. The number of soldiers in the sample was about 1,100.

Again the data were subjected to the same type of analysis,[1] which produced the current version, CUAP-8108, an 88-item questionnaire covering 21 basic topic areas. In addition to minor revisions the new CUAP instrument contains some entirely new items, and it will therefore have to be analyzed after sufficient data have been collected. As of November 1981, CUAP-8108 had been administered to approximately 960 soldiers in 18 companies in the 1st Cavalry Division at Fort Hood and about 1,000 soldiers from units within the 7th Corps' 72nd Field Artillery Brigade in Europe. Work is currently underway for developing Army-wide norms.

Feedback profiles. Figures 2 and 3 are examples of the first of two feedback profiles provided to each unit commander who utilizes the survey. These examples depict actual data from the highest- (Figure 2) and lowest- (Figure 3) scoring units surveyed to date. Down the left-hand column of the profiles are listed the 21 topic, or "factor," areas. To the right of each factor title is a scale that ranges from -100 to +100. The military unit surveyed receives a score on each factor, which is portrayed by the large open triangle on the factor scale. Essentially, the scale can be interpreted as ranging from "very bad" to "very good,"--although emphasis is not placed on the absolute judgment implied by these terms; rather, it is placed on the the unit's obtained score in relation to the average of all units recently utilizing the survey. That average is shown on the scale by the small solid arrowhead. Thus, the column of solid arrowheads represents the norms associated with

---

[1] The item composition of the factors, item contents, and factor analytic results of the analysis are available from the author, upon request.

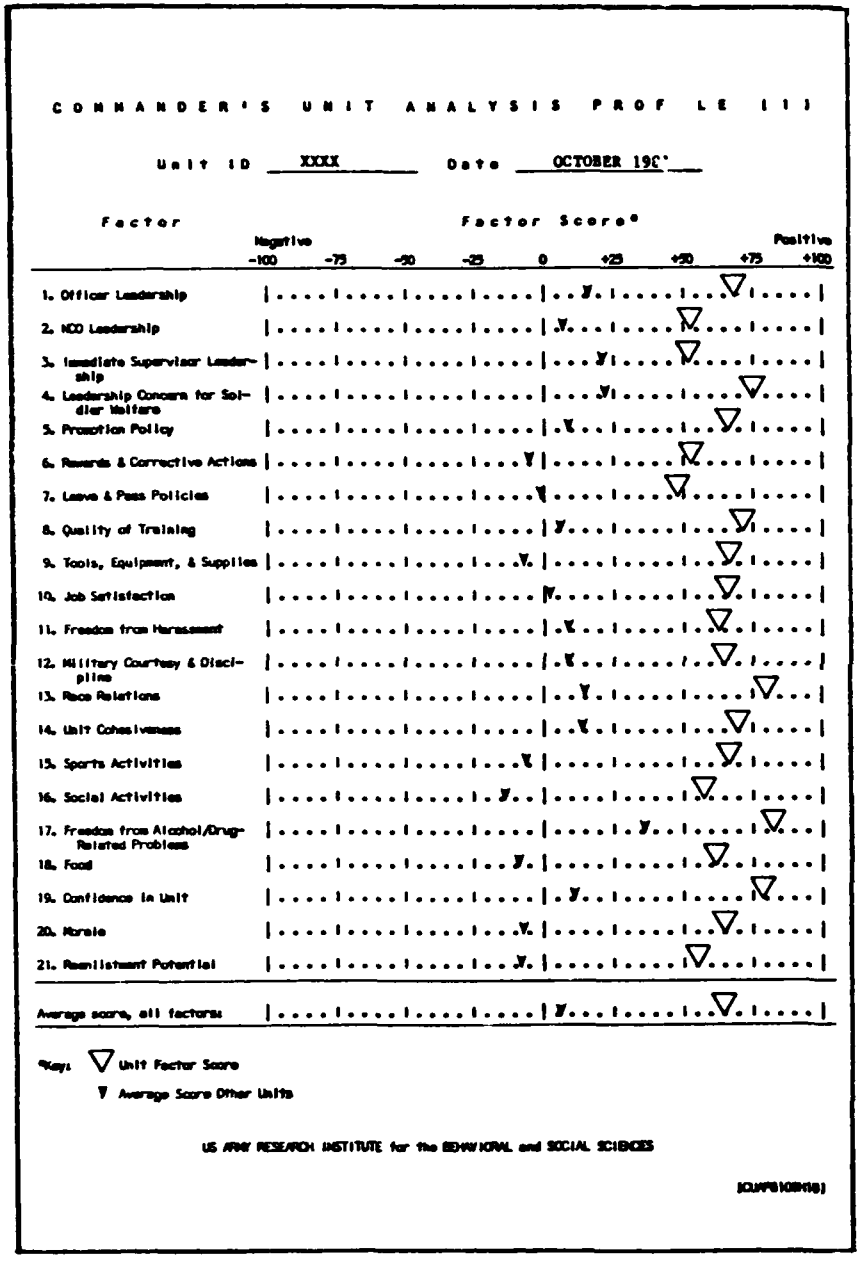



Figure 2. CUAP factor-score feedback profile, depicting data from highest-scoring unit to date.

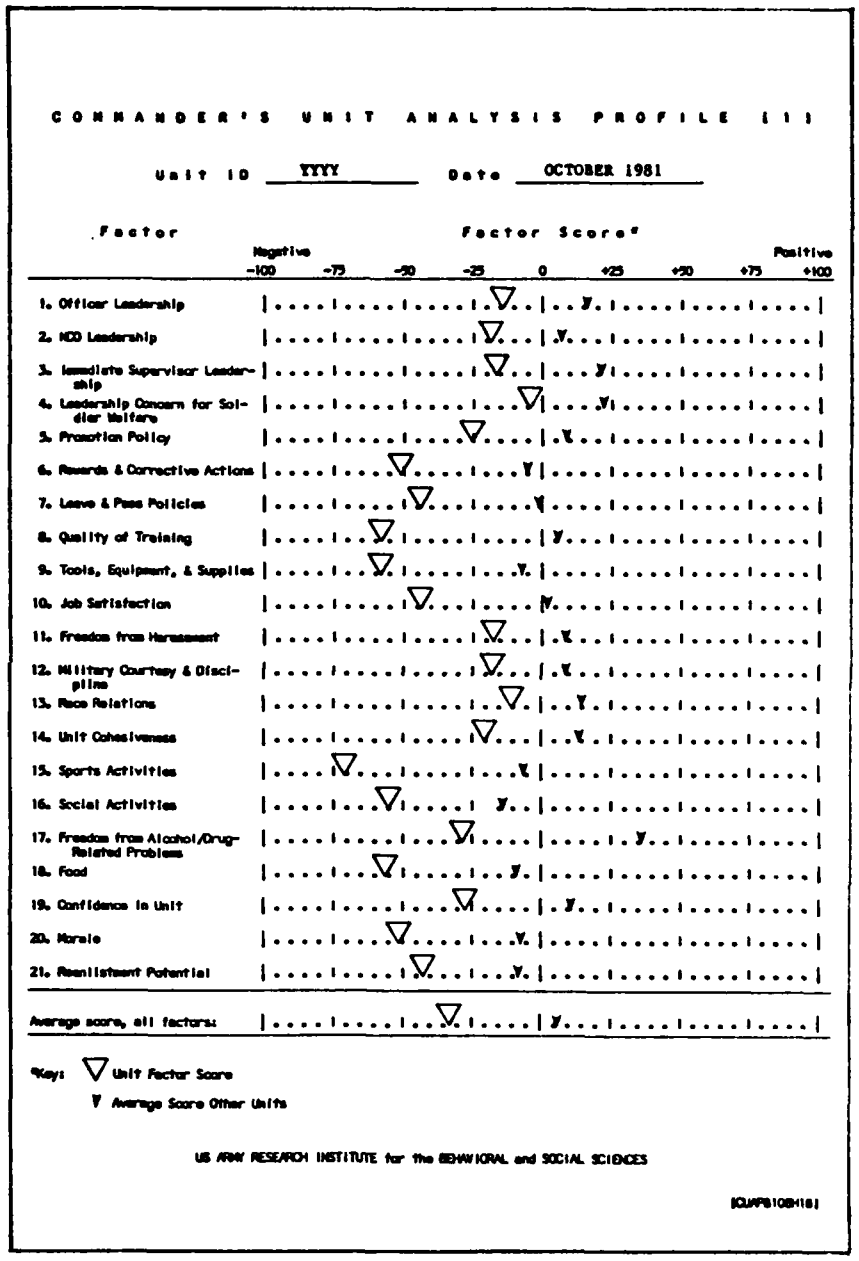



Figure 3. CUAP factor-score feedback profile, depicting data from lowest-scoring unit to date.

the questionnaire. It should be added that because the current version of the questionnaire has been in use for a short time only, the norms shown in Figures 2 and 3 must be considered tentative. As mentioned earlier, Army-wide norms are in the process of being developed.

The unit commander is also provided a second feedback profile that shows the percentile rank of the unit for each factor area. For each factor the percentage of all military units receiving equal or lower scores is indicated on a scale ranging from 0 to 100. This profile simply provides another way for commanders to see their units in comparison with other units that have utilized the survey. The percentile profiles for the highest- and lowest-scoring units to date are shown in Figures 4 and 5, respectively.

## Utilization

The research described has resulted in the development of an easy-to-understand, multiple-choice questionnaire that can be administered by a single person to one, two, or three company-size units at a sitting. The amount of time required for a soldier to complete the instrument is typically 15 to 20 minutes; the total time required for instructions, administration, and collection of completed questionnaires is usually about 30 minutes. In its several developmental versions, the CUAP questionnaire has been administered to approximately 8,000 soldiers in more than 100 different company-size units in nine divisions within the continental US and US Army Europe.

Adequate future research and development for the CUAP project is greatly dependent upon continued enthusiastic reception of the project by Army leaders. Although the project requires close supervision by the Army Research Institute for the next few years, it is intended that the CUAP be implemented as soon as possible into the Army as a standard tool for leadership development. A complete system package, including automatic data processing, is scheduled for release by the end of the current fiscal year (FY 82).

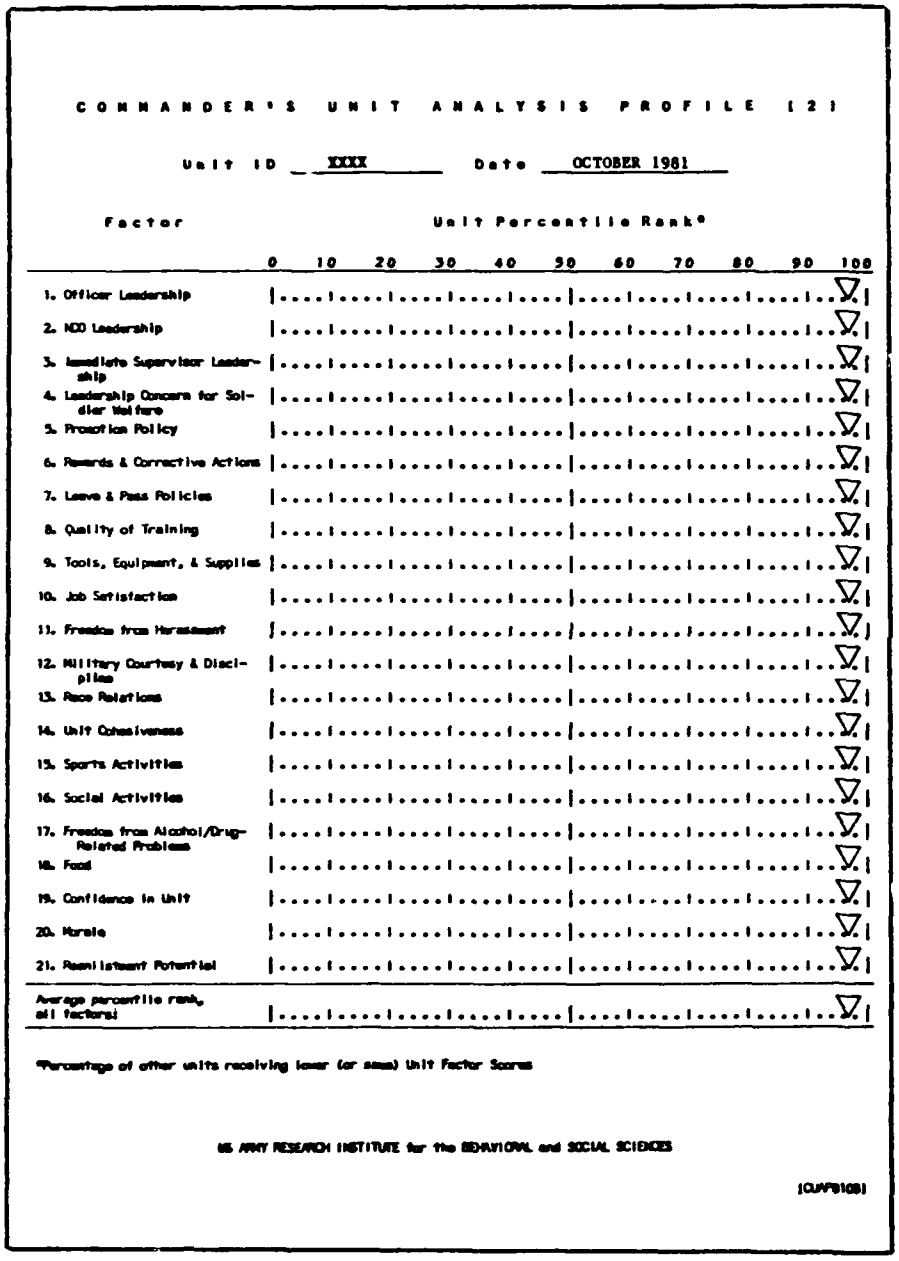



Figure 4. CUAP percentile-rank feedback profile, depicting data from highest-scoring unit to date.

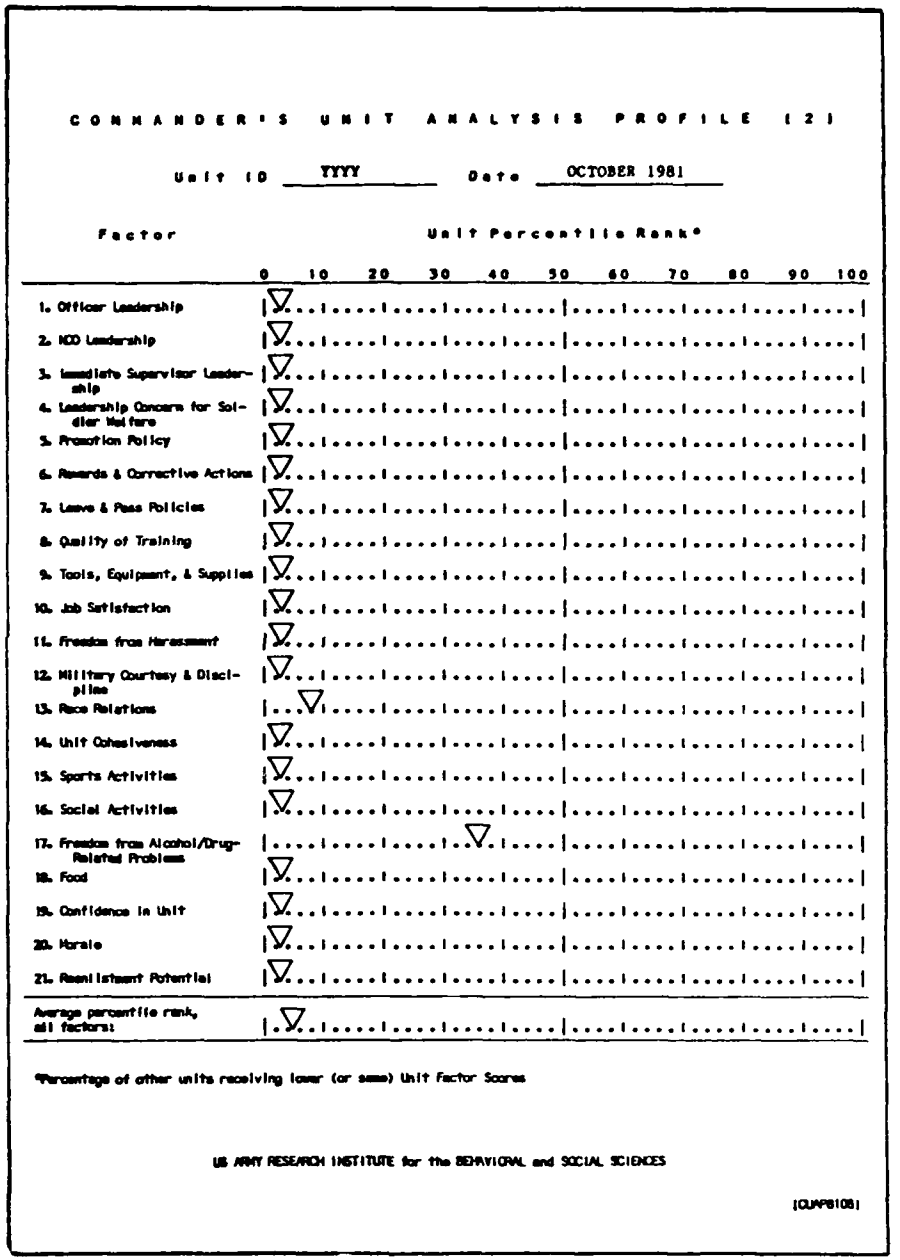



Figure 5. CUAP percentile-rank feedback profile, depicting data from lowest-scoring unit to date.

AD P001363

Phalen, William J., Air Force Human Resources Laboratory, Brooks Air Force Base, Texas. (Wed. A.M.)

CODAP: New Techniques to Improve Job-Type Identification and Definition

This paper will discuss previously used indices of "within" and "between" group overlap and their limited effectiveness in selecting and defining job types from a CODAP job clustering. Newly developed indices of core-task homogeneity and task uniqueness will be defined, demonstrated, and contrasted with the old indices as to quantity and quality of information provided. Printed output of the CODAP CORSET program, which computes and displays data on computer or analyst selected job groups in terms of the new indices, will be presented and interpreted in terms of their application in several occupational analysis studies.



PREVIOUS PAGE IS BLANK

# CODAP: SOME NEW TECHNIQUES TO IMPROVE JOB TYPE IDENTIFICATION AND DEFINITION

by

William J. Phalen
and
Johnny J. Weissmuller

Air Force Human Resources Laboratory
Brooks AFB, Texas

The most complex and controversial task of Air Force occupational analysts is the selection of homogeneous and meaningful job clusters (job types) from a CODAP hierarchical clustering. As many of you already know, CODAP is an acronym for Comprehensive Occupational Data Analysis Programs and refers to a software package that manipulates and reports task-level and biographical survey data gathered from job incumbents and expert raters for the purpose of identifying and analyzing current job structures and task characteristics within a target occupational area. Over the years, a number of CODAP programs have been devised to assist occupational analysts in this task, but even the most useful programs, singly or in combination, have fallen short of automating the selection of job types. Occupational analysts, in all their variability of knowledge, experience, and judgment, are still the most important links in the selection process--and they always will be, whenever subtle delineation of significant job types for specific purposes is required. Nevertheless, there is still a sizable region in the job type selection process for which subjective judgment could be partially or totally eliminated. Manual calculations and intuitive decisions are still being made which the computer could make or simulate more accurately and efficiently than the occupational analyst. This paper will discuss somewhat briefly a new set of analysis techniques which are in the process of being incorporated into the CODAP system. These techniques are designed to enhance the job type selection process. They will significantly improve the analyst's ability to accurately delineate and differentiate job types, and do so with greater speed and less manual effort.

Some of the ideas expressed in this paper were communicated at the Occupational Analysts' Conference at Randolph AFB, San Antonio TX, last May. This paper will review those ideas and show how they have been expanded upon and incorporated into a new CODAP program called "CORSET" and a program yet to be developed called "GRPSET."

## Development of a Core-Task Homogeneity Index

Throughout the 20 years that have witnessed the evolution of the CODAP system, the emphasis in job typing has been on various measures of between-group similarity. The "BETWEEN" or "BEST" value has, from the beginning, been the objective function that drives the clustering process. From the beginning, also, the initial selection of job types has been tied to a range of "BETWEEN" values, as reported in the cluster merger diagram (DIAGRM). The matrix print overlap program (MTXPRT) was also developed to

provide an index of similarity between pairs of composite (average) job descriptions. Simultaneously, the group difference (GRPDIF) program was developed to provide a means of pinpointing the components of difference between groups, namely, task differences ordered from highest positive difference to highest negative difference in terms of percent members performing or average percent time spent by all members. Still later, the Automated Job Type Selection Program (AUTOJT) was designed to assess six kinds of differences between pairs of groups. Meanwhile, the "WITHIN" function, which is an overall measure of within-group homogeneity computed during the hierarchical clustering process, was virtually ignored in job typing, and there were good reasons why this occurred. While the "BETWEEN" function calculates a precise estimate of similarity between two groups independently of group size, the "WITHIN" value is heavily influenced by group size, being relatively insensitive to the contribution of small groups when they are merged with large groups. The "WITHIN" also gives a distorted picture of homogeneity for small groups, because of the disproportionate weight contributed by the diagonal cells of the "WITHIN" matrix, i.e., the 100% overlap of every case with itself. In a word, the "WITHIN" index presented a major interpretability problem. Another measure of within-group homogeneity which was used in the early years and later dropped was the "percent time perfectly described," which was the average overlap of each case in a group with the average job description for the group. Users of CODAP data had difficulty comprehending the meaning of this measure, and researchers relied on it even less than they relied on the "WITHIN" to select meaningful job types. Within the last year or so, this measure has been revived as a means of determining the overlap of jobs of individual workers with task profiles of various civilian job classifications. It is only within the last several months that experimentation has hit upon a promising new measure of within-group homogeneity that should prove to be a useful tool in selecting meaningful job types. It is called the "core-task homogeneity index," because it concentrates on those tasks that are most representative of a specific group of workers and the amount of time the group devotes to those tasks. This index ignores the many tasks in a group job description which are specific to individual incumbents and subgroups and which are, at best, peripheral to the identification of a job type. The measure is also designed to be easily interpretable by occupational analysts and other users. To simplify the explanation of core-task homogeneity, I will first lay out the steps for calculating the core-task homogeneity index and follow this with a discussion of its properties.

## Calculation of Core-Task Homogeneity Index

Step 1. Select all tasks in a group job description performed by at least 66 2/3% of the group. The value "66 2/3%" was chosen as the minimum criterion of performance, because it seemed only fitting that any task to be used as a core task to identify a job type should be performed by at least 2/3 of the group; i.e., the ratio of performers to non-performers should be at least 2-to-1. The maximum possible percentage of group members performing a task is, of course, 100%.

Step 2. Compute the cross-product of "percent members performing" and "average percent time spent by all members" for each selected task. The cross-product serves to weight the average percent time spent on

each selected core task in direct proportion to the percentage of group members performing the task.

Step 3. Sum the cross-products and divide the result by 100 to arrive at the core-task homogeneity index, which is nothing more than the weighted sum of average percent time spent by all group members on the selected core tasks based on a standard of 100% members performing each task.

Table 1 is an example of how the core-task homogeneity index is computed.

---

Table 1. Simulated Task Data Exemplifying Calculation of Core-Task Homogeneity Index

| Task | % Members Performing | Average Percent Time Spent by All Members | Cross-Product of Selected Core Tasks |
|---|---|---|---|
| * A1 | *100.00 | * 5.00 | 500.00 |
| * A2 | * 90.00 | * 4.70 | 423.00 |
| * A3 | * 75.00 | * 4.40 | 330.00 |
| * A4 | * 69.88 | * 4.25 | 296.99 |
| A5 | 55.00 | 4.50 | |
| B1 | 60.00 | 4.00 | |
| * B2 | * 66.67 | * 3.00 | 200.01 |
| * B3 | *100.00 | * 2.50 | 250.00 |
| C1 | 65.00 | 2.35 | |
| C2 | 33.33 | 2.15 | |
| CORE-TASK TOTAL | 501.55 | 23.85 | 2,000.00 |

*Selected as core task ( $\geq$ 66 2/3% performing). Number of tasks selected = 6.

$$\text{Core-Task Homogeneity Index} = \frac{2{,}000.00}{100} = 20.00$$

$$\text{Average percentage of members performing core tasks} = \frac{501.55}{6} = 83.59.$$

---

Now, how should the index value "20.00" be interpreted? Table 2 will be helpful at this point.

Table 2. The Interpretation Limits for the Core-Task Homogeneity Index Computed from the Simulated Task Data in Figure 1.

| % Performing | | Sum of Percent Time Spent | | | | Homogeneity Index |
|---|---|---|---|---|---|---|
| 100 (Maximum) | X | 20.00 (Minimum) | ÷ | 100 | = | 20.00 |
| 83.59 (Average) | X | 23.93 (Average) | ÷ | 100 | = | 20.00 |
| 66 2/3 (Minimum) | X | 30.00 (Maximum) | ÷ | 100 | = | 20.00 |

Although Table 2 shows the calculation of only three cross-products which yield a core-task homogeneity index of 20.00, there is really an infinite number of cross-products that could be computed between the specified "percent performing" limits of 100% (maximum) and 66 2/3% (minimum) that would yield a homogeneity index of 20.00.

However, "% performing" minimum and maximum limits define the "sum of percent time spent" limits of 20.00% (minimum) and 30.00% (maximum). Thus, if no further information were given than the index number "20.00," we would know with certainty that the core-task homogeneity index for this job description could be the result of a set of one or more tasks which are performed by 100% of the job incumbents and account for a total of 20.00% of work time, or which are performed by 66 2/3% of the job incumbents and account for a total of 30.00% of work time, or some other combination of "% performing" and "sum of percent time" falling between the minimum and maximum limits and which yields a cross-product: "sum of percent time spent" ÷ 100 = 20.00. Note that the maximum "sum of percent time spent" (30.00) is 1 1/2 times the minimum (20.00). This multiple (1 1/2) holds true for any value of the core-task homogeneity index. For example, the interpretation limits or "window" for an index value of "40.00" would be: 100% of group members performing tasks accounting for 40.00% of the group's work time, or 66 2/3% of group members performing tasks accounting for 1 1/2 X 40.00 = 60.00% of the group's work time, with the actual average falling somewhere between these limits. Table 2 shows the "average" cross-product of "average % performing" and "average sum of percent time spent" for the simulated data in Table 1 to be 83.59 X 23.93. The computation of the "average % performing" value (83.59) is shown in Table 1; the "average sum of percent time spent" value (23.93) was arrived at by solving the equation: $83.59(x) \div 100 = 20.00$.

Another important property of the core-task homogeneity index is that it can be rapidly and efficiently calculated for all groups reported in a cluster-merger diagram, because no ordering of tasks is involved in the calculation. The index was computed for numerous clustered groups in several Air Force Specialties which had recently been analyzed and job typed. About 90% of job types analyzed had a core-task homogeneity index between 20%-30%. There were a few job types that fell in the teens. A close inspection of the task data for the deviant groups confirmed our suspicion that groups having a core-task homogeneity index below 20.00 are not sufficiently homogeneous to warrant being selected as job types. However, groups with index values in the

teens are fitting candidates for selection as job clusters (umbrella groups containing two or more job types, or job types subsuming a number of add-on cases). Probably the most appealing aspect of the core-task homogeneity index is its concreteness. It tells you rather precisely what degree of within-group cohesiveness you will find among the top tasks in a group job description. This is not true of an overall index of similarity, such as the average "WITHIN," whose value is derived from all sorts of pairwise combinations of tasks. Put another way, the core-task homogeneity index eliminates a large part of the "noise" generated by the "WITHIN" with minimal loss of relevant information.

The next step in the use of the core-task homogeneity index is to substitute it for the "WITHIN" in the cluster merger DIAGRAM printout and use it in the selection of starter groups; i.e., the top row of groups on the DIAGRAM printout. This is one of the things the proposed GRPSET program will do, and it will ensure that every starter group is at the appropriate level of homogeneity to be considered for selection as a job type.

Going a step further, if the core-task homogeneity index is computed for every group formed in a CODAP hierarchical clustering, each cluster group can be evaluated in terms of whether it is sufficiently homogeneous; i.e., has a group of tasks performed by a sizable proportion of its members, e.g. 66 2/3%, which account for a sizable percentage of group work time, e.g., 20.00%, to be considered as a potential job type. If a group passes the homogeneity test, it must then be determined whether a significant portion of the group's core time is sufficiently unique as to differentiate that group from other homogeneous groups. The next part of this paper will address this issue in detail by proposing and describing two new task indices: a task uniqueness index and a task discrimination index.

## Development of a Task Uniqueness Index and a Task Discrimination Index

"Task Uniqueness," as defined in this paper, refers to the extent to which a task is or is not performed exclusively by one group of workers as compared to this group's complementary group (complementary uniqueness/discrimination), or as compared to all other groups in a selected set of mutually exclusive groups (contextual uniqueness/discrimination). A set of contextual groups may be defined in terms of KPATH ranges derived from clustering, or in terms of one or more background variables, such as grade level, major command, time in service, type of aircraft maintained, etc. If a complementary group is used, it will usually be the total sample minus the target group, although it may also be some restricted sample, such as a single career ladder, minus a target group wholly contained in the restricted sample.

"Task Discrimination" and "Task Uniqueness" are related, but not equivalent, concepts. A discriminating task is a unique task that is also a core task for one or more groups. Thus, all discriminating tasks are also unique tasks, but not all unique tasks are discriminating tasks. For example, if all performers of a rarely performed task are contained in a single subgroup of the total sample, that task is uniquely associated with that subgroup, even though it is performed by too few members, i.e., < 66 2/3%, to be considered a core task.

While the core-task homogeneity index tells whether there is a sufficiently large block of group work time concentrated in tasks performed by a sufficiently large percentage of incumbents to give it potential job type status, the core-task discrimination index determines what portion of the group's core tasks and core time are sufficiently peculiar to that group as to identify it as a job type unto itself, rather than as a homogeneous part of a larger job type group. Carried a step further, the task uniqueness and core-task discrimination indices can also be used to detect secondary functions of subgroups within an otherwise homogeneous job type group.

## Procedures for Computing Task Discrimination and Task Uniqueness Indices

### I. Complementary Approach

Let:

a. Task Discrimination $(TD_i) = P_{T_i} - P_{C_i}$

where:

$P_{T_i}$ = % members in target group T performing task i

$P_{C_i}$ = % members in complementary group (total sample or restricted sample, with target group excluded) performing task i

b. Task Uniqueness $(TU_i) = \dfrac{TD_i}{MAX\ (P_{T_i}, P_{C_i})}$

See Table 3 for some example calculations of task discrimination (TD) and task uniqueness (TU) indices.

Table 3. Example Calculations of Task Discrimination (TD) and Task Uniqueness (TU) Indices Using the Complementary Approach

| TASK | $P_{T_i}$ | $P_{c_i}$ | $P_{T_i} - P_{c_i}$ | $TD_i$ | $\frac{TD_i}{MAX\ (P_{T_i}, P_{c_i})}$ | $TU_i$ |
|---|---|---|---|---|---|---|
| A | 80 | 30 | 80 - 30 | 50 | $\frac{50}{80}$ | .63 |
| B | 20 | 90 | 20 - 90 | -70 | $\frac{-70}{90}$ | -.78 |
| C | 100 | 0 | 100 - 0 | 100 | $\frac{100}{100}$ | 1.00 |
| D | 0 | 70 | 0 - 70 | -70 | $\frac{-70}{70}$ | -1.00 |
| E | 40 | 40 | 40 - 40* | 0* | $\frac{0}{40}$ | 0.00 |
| F | 0 | 100 | 0 - 100 | -100 | $\frac{-100}{100}$ | -1.00 |
| G | 10 | 0 | 10 - 0* | 10* | $\frac{10}{10}$ | 1.00 |

Several observations can be made upon inspection of Table 3:

1. $TD_i$ can range from "+100" (uniquely performed, highly discriminating) to " − 100" (uniquely not performed, highly discriminating), with a midpoint of "0" (totally lacking uniqueness, no discrimination).

2. $TU_i$ can range from "+1.00" to "-1.00" with midpoint of "0," and with the same interpretation that extreme and midpoint values have for TD.

3. $|TU_i \times 100| \geq |TD_i|$ in all cases, because, unlike TD, TU is proportionate to the percentage of workers who perform a task; i.e., TU can have a large value even if both $P_T$ and $P_C$ are low; TD can be large only if $P_T$ or $P_C$ is high and the other is low.

In lieu of practical experience, tentative cutoffs for classifying a task as discriminating or unique have been set at: $33.33 < TD < -33.33$ & $.50 < TU < -.50$.

II. Contextual Approach

The contextual approach, which calculates task discrimination and task uniqueness with reference to a set of groups derived from a common context,

*TD is normally not computed if $P_{T_i}$ or $P_{c_i}$ does not qualify task as a core task; i.e., $P_{T_i}$ and $P_{c_i} < 66\ 2/3$.

i.e., representing categories of a discrete or continuous variable or combination of variables, is a linear extension of the two-group solution, as described in the complementary approach, to a k-group problem. The k-group solution is as follows, with example calculations shown in Table 4:

Let:

$$TD_{T_i} = \left| Q_{T_i} - R_{T_i} \right| \cdot \frac{Q_{T_i} - R_{T_i}}{(k-1)^2}$$

$$TU_{T_i} = \left| \hat{Q}_{T_i} - \hat{R}_{T_i} \right| \cdot \frac{\hat{Q}_{T_i} - \hat{R}_{T_i}}{(k-1)^2}$$

where:

$TD_{T_i}$ = task discrimination index for target group T on task i

$TU_{T_i}$ = task uniqueness index for target group T on task i

$$Q_{T_i} = \sum_{j=1}^{k-1} \sqrt{P_{T_i} - P_{c_{ji}}} \quad \text{for all } P_{T_i} - P_{c_{ji}} \geq 0$$

$$R_{T_i} = \sum_{j=1}^{k-1} \sqrt{\left| P_{T_i} - P_{c_{ji}} \right|} \quad \text{for all } P_{T_i} - P_{c_{ji}} < 0$$

k = number of contextual groups before target group is selected

$P_{T_i}$ = percent members performing (P) on task i by target group T

$P_{c_{ji}}$ = percent members performing (P) on task i by contextual group j

$$\hat{Q}_{T_i} = \frac{Q_{T_i}}{\sqrt{P_{T_i}}} \quad \text{for all } P_{T_i} - P_{c_{ji}} \geq 0$$

$$\hat{R}_{T_i} = \sum_{j=1}^{k-1} \sqrt{\left| \frac{P_{T_i} - P_{c_{ji}}}{P_{c_{ji}}} \right|} \quad \text{for all } P_{T_i} - P_{c_{ji}} < 0$$

Table 4. Example Calculations of Task Discrimination (TD) and Task Uniqueness (TU) Indices Using the Contextual Approach

Given:

Matrix of Percent Members Performing Values for the Core Tasks of 11 Contextual Groups

| CORE TASK ID | GROUP ID 0631 | 0220 | 0586 | 1122 | 1001 | 0097 | 0154 | 0488 | 0086 | 0372 | 0947 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0001 | 70 | 70 | 90 | 100 | 20 | 15 | 5 | 0 | 80 | 80 | 100 |
| 0003 | 85 | 100 | 60 | 20 | 0 | 5 | 5 | 95 | 90 | 80 | 100 |
| 0025 | 5 | 100 | 30 | 20 | 0 | 10 | 5 | 100 | 90 | 80 | 100 |
| 0057 | 10 | 50 | 10 | 80 | 10 | 15 | 5 | 50 | 100 | 80 | 100 |
| 0112 | 30 | 25 | 100 | 70 | 100 | 20 | 60 | 50 | 100 | 40 | 100 |
| 0351 | 100 | 20 | 90 | 70 | 100 | 25 | 70 | 30 | 60 | 40 | 0 |
| 0380 | 95 | 0 | 40 | 60 | 100 | 30 | 80 | 40 | 70 | 40 | 0 |
| 0722 | 50 | 30 | 0 | 90 | 90 | 35 | 80 | 0 | 75 | 40 | 0 |
| 0814 | 60 | 50 | 10 | 30 | 90 | 40 | 70 | 0 | 20 | 0 | 0 |
| 0948 | 40 | 90 | 10 | 40 | 90 | 45 | 60 | 60 | 0 | 0 | 0 |
| 1020 | 75 | 50 | 30 | 20 | 10 | 55 | 60 | 70 | 75 | 75 | 75 |

Let:

$T_i$ = group 0631 and task 0003

then:

$$Q_{T_i} = \sqrt{85 - 60} + \sqrt{85 - 20} + \sqrt{85 - 0} + \sqrt{85 - 5} + \sqrt{85 - 5} + \sqrt{85 - 80}$$

$$= 42.41$$

$$R_{T_i} = \sqrt{|85 - 100|} + \sqrt{|85 - 95|} + \sqrt{|85 - 90|} + \sqrt{|85 - 100|}$$

$$= 13.14$$

$$TD_{T_i} = \left|42.41 - 13.14\right| \cdot \frac{42.41 - 13.14}{(11 - 1)^2} = 8.57$$

$$\hat{Q}_{T_i} = \frac{42.41}{\sqrt{85}} = 4.60$$

$$\hat{R}_{T_i} = \sqrt{\left|\frac{85 - 100}{100}\right|} + \sqrt{\left|\frac{85 - 95}{95}\right|} + \sqrt{\left|\frac{85 - 90}{90}\right|} + \sqrt{\left|\frac{85 - 100}{100}\right|} = 1.33$$

$$TU_{T_1} = \left|4.60 - 1.33\right| \cdot \frac{4.60 - 1.33}{(11 - 1)^2} = .11$$

Complementary vs. Contextual Discrimination/Uniqueness

I. Complementary Approach

a. Advantages

1. The idea of using a complementary group to determine the discrimination and uniqueness of tasks in a target group is appealing and easy to understand.

2. Computation of task discrimination and uniqueness indices is straightforward and simple.

3. Task discrimination and uniqueness indices can be computed for any selected group and the indices are comparable across multiple groups as long as the groups are drawn from the same total or restricted sample.

b. Disadvantages

1. The complementary group is not a meaningful group in its own right, but a potpourri of unrelated as well as related individuals and groups.

2. The complementary group is insensitive to job group structures outside the target group and, therefore, provides a somewhat inaccurate standard against which to measure task discrimination and uniqueness in the target group.

II. Contextual Approach

a. Advantages

1. The contextual approach is sensitive to job structures existing within a well-defined context of groups.

2. The contextual approach results in the pinpointing of a larger number of discriminating and unique tasks having higher index values.

b. Disadvantages

1. Task discrimination and uniqueness indices cannot be computed without first selecting a set of contextual groups. When the goal is the identification and selection of job types from the hierarchical clustering, the contextual set should consist of the full array of potential job type groups.

2. Task discrimination and uniqueness indices computed for a set of contextual groups are not valid outside the context of the set of groups for which they were computed.

3. Task discrimination and uniqueness indices are unduly influenced by the presence of small groups in the contextual set, since group size is ignored in the computation of the indices. On the other hand, weighting the computations by group size would give undue weight to unusually large groups.

## The Development of a Measure of Average Discrimination Per Unit of Core-Task Homogeneity (ADPUCTH)

Although the core-task homogeneity index is a good measure of within-group cohesiveness, which is one essential consideration in determining whether a clustering-derived group should be selected as a job type, it provides no information as to whether the group is or is not distinguishable from other clustering-derived groups that have been selected as potential job types. A well-chosen job type group must be internally cohesive and externally distinguishable from all other selected job type groups. The availability of the task discrimination index can solve the distinguishability problem through the computation of an average discrimination per unit of core-task homogeneity (ADPUCTH) index. It is derived by computing and summing the cross-product of three vectors of task values: the vector of core-task discrimination indices, the "percent members performing" vector, and the "average percent time spent by all members" vector (only the last two vectors are used in computing the core-task homogeneity index) and dividing the sum of cross-products by "10,000" (the sum of cross-products used in computing the core-task homogeneity index is divided by "100"). Symbolically, the computation is as follows:

Let:

$$\text{ADPUCTH} = \frac{1}{10{,}000} \sum_{i=1}^{k} (P_i \cdot T_i \cdot D_i)$$

where:

k = number of core tasks in selected group

$P_i$ = "percent members performing" value for core task i

$T_i$ = "average percent time spent by all members" value for core task i

$D_i$ = discrimination index for core task i

## The CODAP CORSET and GRPSET Programs

### CORSET

The development of the core-task discrimination index has led to the development of an experimental CODAP program which is currently being tested at the Air Force Occupational Measurement Center as a tool for identifying and defining clustering-derived job type groups, as well as for distinguishing between skill-level groups, time-in-service groups, and type-of-aircraft-maintained groups within the aircraft maintenance area. The new program is called "CORSET," because it provides homogeneity and discrimination data on sets of

core tasks for a set of selected contextual groups of job incumbents. There are so many sections and subsections within the CORSET report that I will not be able to include samples of output of the report in this paper, which already exceeds the guideline for length. Rather I will give a brief descriptive title for each section and subsection. After more experience and further planned improvements to CORSET, a future paper will be presented which will explain and evaluate all aspects of the CORSET program in detail.

List of CORSET Report Contents*

1. Program control card listing, including specified uniqueness (discrimination) cutoff values and identification information for each criterion (contextual) group.

2. Composite job description for the criterion (contextual) groups ordered from most to least frequently occurring core tasks.

3. Task uniqueness (discrimination) report for each criterion (target) group.
   a. Core tasks in descending order of uniqueness
   b. Non-core unique tasks
   c. Tasks uniquely not performed
   d. Highly unique tasks in descending order on time spent
   e. Fairly unique tasks in descending order on time spent
   f. Fairly common tasks in descending order on time spent

4. Summary of group statistics for this group (target group).

5. Similarity of this group (target group) to each criterion group (all remaining contextual groups) based on various measures of common core time.

6. Index of criterion (contextual) groups, together with summary statistics for each group for the task categories listed in item 3 above.
   a. Index in input order
   b. Index in core-task homogeneity order
   c. Index in core-task uniqueness (discrimination) per unit time spent order
   d. Index in group ID sequence

The current version of the CORSET program promises to save 10% to 20% of the time it now takes an occupational analyst at the Air Force Occupational Measurement Center to accomplish a complete, large-sample, job type analysis.

*Words in the list which are in parentheses represent vocabulary changes not yet reflected in the current version of the CORSET report.

## GRPSET

A planned, but not yet developed, program titled "GRPSET" (Group Set) will select those groups from a job clustering which have the greatest potential for being meaningful job types. It will use the core-task homogeneity index and the average discrimination per unit of core-task homogeneity index to select the groups. Abbreviated job descriptions and a CORSET report will be produced for the selected groups, as well as a cluster merger diagram in which these groups will form the topmost row of groups (starter groups).

The GRPSET program will probably save at least 10% to 20% of the time it takes to accomplish a thorough job type analysis over and above the CORSET savings, since it will greatly cut down the length of time it now takes an occupational analyst to make the initial selection of potential job types. An important spinoff of the automated selection procedure will be a greater standardization of the job type selection process, an especially important consideration when there is a relatively frequent turnover of occupational analysts. A future paper will discuss in detail the group selection algorithms which will be used in GRPSET. A promising algorithm has already been developed, but it would be premature to describe it before it has been tested.

## Some Potential Applications of the CORSET/GRPSET Technology

The principal arena in which the CORSET/GRPSET technology will be used is in the identification and definition of job type groups from a hierarchical clustering.* Of almost equal importance is the ability of the technology to pinpoint task differences among groups of job incumbents defined in terms of one or more background variables. These are the more obvious uses. Some other, less obvious, applications of the technology might be:

1. Built into the CORSET program is a task-modularizing, group-subsetting capability that may lead to a technique for deriving taxonomic networks of clustering-derived job type groups. Such a technique would identify the multidimensional relationships that elude current automated techniques.

2. The core-task homogeneity index might be the best available means for evaluating clustering approaches which transform the original scale values of task data at input time in such a way as to give greater weight to the most time-consuming tasks. Since the most time-consuming tasks tend to drive the clustering, it stands to reason that the heavier weighting of the most time-consuming tasks should result in a greater number of groups which display more core tasks performed by larger percentages of incumbents and which account for more core time than would be found in a standard clustering using the original scale values.

3. The task discrimination/uniqueness indices might be used to select the smallest subset of tasks capable of accurately assigning job incumbents to previously defined job type groups. It would be the fastest and most economical way of classifying the jobs of incumbents who have not yet been surveyed.

*The CORSET/GRPSET programs do not replace the Automated Job Type Selection Program (AUTOJT), but enhance its usefulness by screening out those groups that are sufficiently distinct as not to require the detailed pairwise comparisons performed by AUTOJT.

## Conclusion

The possibility and need to continue making substantive improvements to occupational analysis methodology, in general, and to CODAP technology, in particular, is still very much alive. The hardware and the software are just now beginning to catch up with some of the old ideas that were not feasible to automate when they were first expressed, such as those presented in this paper. New ideas are constantly being generated, as new applications and changing needs of technology users provide new and deeper insights into the analysis process. The challenge to technology developers is to find practical ways to automate those aspects of the analysis process which assist the human judgmental process without usurping its proper domain.

## REFERENCES


Anderberg, Michael R. Cluster Analysis for Applications. New York: Academic Press, 1973.

Archer, Wayne B. Computation of Group Job Descriptions from Occupational Survey Data. PRL-TR-66-12, AD-653 543. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, December 1966.

Christal, Raymond E. The United States Air Force Occupational Research Project. AFHRL-TR-73-75, AD-774 574. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

Phalen, William J. & Christal, R.E. Comprehensive Occupational Data Analysis Programs: Group Membership (GRMBRS/GRPMBR) and Automated Diagramming (DIAGRM) Programs. AFHRL-TR-73-5, AD-767 199. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, April 1973.

Thew, Michael C. & Weissmuller, Johnny J. CODAP: A Current Overview. Paper presented to the 21st Annual Conference of the Military Testing Association, U.S. Navy, San Diego, CA, October 1979.

Ward, Joe H., Jr. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 1963, 58, 236-244.

## Conclusion

The possibility and need to continue making substantive improvements to occupational analysis methodology, in general, and to CODAP technology, in particular, is still very much alive. The hardware and the software are just now beginning to catch up with some of the old ideas that were not feasible to automate when they were first expressed, such as those presented in this paper. New ideas are constantly being generated, as new applications and changing needs of technology users provide new and deeper insights into the analysis process. The challenge to technology developers is to find practical ways to automate those aspects of the analysis process which assist the human judgmental process without usurping its proper domain.

## REFERENCES


Anderberg, Michael R. Cluster Analysis for Applications. New York: Academic Press, 1973.

Archer, Wayne B. Computation of Group Job Descriptions from Occupational Survey Data. PRL-TR-66-12, AD-653 543. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, December 1966.

Christal, Raymond E. The United States Air Force Occupational Research Project. AFHRL-TR-73-75, AD-774 574. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

Phalen, William J. & Christal, R.E. Comprehensive Occupational Data Analysis Programs: Group Membership (GRMBRS/GRPMBR) and Automated Diagramming (DIAGRM) Programs. AFHRL-TR-73-5, AD-767 199. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, April 1973.

Thew, Michael C. & Weissmuller, Johnny J. CODAP: A Current Overview. Paper presented to the 21st Annual Conference of the Military Testing Association, U.S. Navy, San Diego, CA, October 1979.

Ward, Joe H., Jr. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 1963, 58, 236-244.

AD P001364

Pigeon, E. Richard, National Defence Headquarters, Ottawa, Ontario

## Effects of Psychological Differentiation and Cognitive Consistency on Students' Course Evaluations

(Abstract)

Concern has arisen on student consistency in their course evaluation reports. Since there might possibly be different evaluations given by different groups of raters, it has been suggested that cognitive styles and affective reactions should be taken into account when evaluating instruction. This research investigates in the literature and in a natural observational setting the influence of one set of two variables, namely "Interest" and "Expected Score", on the evaluation of another set of five course components. Two groups of raters are studied: field dependent and field independent. This research also suggests through the theory of cognitive consistency an explanation for the higher correlation found within one of the two groups of raters between the two sets of variables.



PREVIOUS PAGE IS BLANK

# EFFECTS OF PSYCHOLOGICAL DIFFERENTIATION AND COGNITIVE CONSISTENCY ON STUDENTS' COURSE EVALUATIONS

## INTRODUCTION

Accountability has become quite important in the field of education for the last fifteen years or so. The origin of this movement can be traced back to the final version of the Congress Elementary and Secondary Act of 1965 Title I and III stipulating that each project conducted under their support had to be evaluated and evaluations reported to the federal government. Since one of the main sources of data collection in an educational setting is the student himself, it is worthwhile to consider the degree of variability between students in their course evaluations. The present paper will address this matter mainly by getting to an understanding of why students evaluate a course the way they do.

A good definition of evaluation is provided by Crawford (1979):

"Program Evaluation is a process involving the making of judgments about a program, or a part thereof, to determine its worth. The process involves systematic efforts to gather, analyze, and interpret information about the program. This information is then used to determine if the program is accomplishing what it is set out to accomplish. This does not, however, preclude the possibility of attaining unforeseen objectives."

The same author notices that the content and usefulness of ratings is a function of their use and the intended audience.

The study of Penfield (1972) is of a particular interest in that it gathers students' comments on the usefulness of evaluation forms. The results of his enquiry from 125 students show:

a. their uncertainty as to information results;

b. their appreciation of the instrument as an effective means of evaluating;

c. that retaliation from the institution was not a major preoccupation.

As to the validity of evaluations, Costin ET AL (1971) report in their review of literature positive correlations between higher evaluations and the estimation of course objectives attainment. They report on studies having obtained positive correlations between present and past students' evaluations of the same teachers and between

students and colleagues of teachers. They also report on higher evaluations given to more experienced teachers. Our review of more recent articles confirms or adds up to the validity of students' evaluations. For instance Marsh ET AL (1979) continue to observe an impressive agreement between the institution and students' evaluations.

But problems start when we consider the so-called reliability of evaluations. Reliability coefficients reported in the literature are generally in the order of 0.80. This is the degree to which various evaluations from various evaluators are proportional when expressed in terms of deviations from the mean. If instead we consider agreement between evaluators, as Feldman (1977) did, coefficients drop from 0.7-0.9 to 0.1-0.3. For this author, then: "To the degree that students within class are not consistent, a concomitant concern arises - namely, whether there are differences among students that may be producing systematic variability in ratings."

The results of a study by Marsh and Overall (1979) on long-term stability of evaluation have indicated that "there are systematic differences among raters that cannot simply be attributed to carelessness." Consequently their suggestion is to report mean group evaluations according to various groupings of variables. Crichton and Doyle had suggested that: "Perhaps it will be possible to group raters according to some theory of how they will rate in a particular situation and their ratings within subgroups will be more uniform than ratings within the total group."

Messick (1970) has recommended to take into account cognitive styles and affective reactions when evaluating instruction. This research investigates in the literature and in a natural observational setting the influence of those variables.

## REVIEW OF LITERATURE

The purpose of our review of literature was threefold. First, to investigate among a constellation of variables those related to the evaluator and the course most likely to influence students' evaluations. Secondly, to investigate the relation of cognitive style on students' evaluations. Thirdly, to explain by the theory of cognitive dissonance the effect of certain variables on evaluations for given cognitive styles.

### 1. Influence of variables on evaluations

This first step of our review has brought forward the following considerations:

a. there are a great many variables (personal, course related, mixed) that do influence course evaluations by students;

b. links between variables and evaluations are often weak and the results of some studies even conflict with others;

c. given their popularity and conspicuous success in explaining differences in evaluations, course interest, course importance and expected score in the course would be variables to be kept in a study on variabilitity of evaluations;

d. as recommended in the literature on the subject, more specifically in the case of expected scores (Feldman, 1976) it would be important in future research to take into account conditions in which variables are more or less associated to evaluation.

2. Cognitive styles and evaluation

According to Messick (1970) cognitive styles and affective reactions are two classes of variables which should be taken into consideration when evaluating: the interactions between these variables and objects evaluated are probable and should be systematically appraised. From three articles, Musella (1969), Greenfeld and Arbuthnot (1969), and Drummond and McIntire (1977), some statements can be made about these relations:

a. breakdown of evaluators according to their cognitive style produces inter-group differences on overall evaluation;

b. breakdown of evaluators according to their cognitive style produces inter-group differences on items of evaluation questionnaires;

c. certain evaluators will be uniform in the evaluation of different objects, others will show great variation.

This influence of cognitive styles on evaluation has made us to consider more specifically a well-known cognitive style termed by Witkin "field dependency". Field dependency is a measure of psychological differentiation. The last formulation of the psychological differentiation model (Witkin, Goodenough, Oltman: 1979) identifies field dependency as self-nonself segregation. This segregation plus the segregation of psychological functions and the segregation of neurophysiological functions are the three major indicators of differentiation. "Segregation of self-nonself or field independency means autonomy of external referents. Limited self-nonself segregation, responsible for less autonomous functioning or a field dependent cognitive style, signifies continued connectedness with others." (Witkin et al, 1979, p.1138).

It is our belief that field dependency as a cognitive style can explain the presence or absence of certain variables' effects on evaluation. Students identified as field dependent, given their little differentiation in their behaviour and global perception of situations rather than their distinct elements, would not have a strong tendency to evaluate a course in function, for instance, of their interest for this course. Field independent evaluators, given their analytical perception of a situation, will discern their course interest and their course evaluation as two entities and will naturally tend to maintain a balance between them or their elements.

### 3. Cognitive dissonance and Cognitive Styles

The fundamental background of the theory according to Festinger (1968) is the need for a human being to maintain the greatest coherence possible: "human organism tries to establish internal harmony, consistency, or congruity among his opinions, attitudes, knowledge, and values" (p.260).

Shaffer and Hendrick (1974) summarize the equilibration process in the individual in the following way. The individual first experiences a discomfort after a cognitive maladjustment. Ther he tolerates it to a certain degree. Finally if the discomfort is not tolerable, he will choose a strategy to reduce the lack of balance and its unpleasant consequences. According to Festinger the usual way is to reevaluate the eventualities. This principal mode of dissonance reduction consists in increasing the difference in subjective values assigned to eventualities between which the choice was made.

Even though according to Shaffer (1975) the details of the relation between cognitive dissonance and cognitive style are not clear, this relation does nonetheless seem real. It is to be expected that field independent evaluators being aware of elements in a situation, some of which possibly conflict with one another, will try to maintain a coherence between these elements more than field dependent evaluators who perceive the situation globally rather than analytically.

## PROBLEM AND RESEARCH HYPOTHSES

The foundation of this research is that there is effectively variation between students' evaluations. Also the influence of the three following variables can be recognized on course evaluation: interest, importance and expected score in relation with the course. Moreover because of lesser or greater need in the evaluator of cognitive consistency between what he feels towards the course and its evaluation one should find a more or less strong correlation between the above three variables and course evaluation according to the evaluator's cognitive style.

The research problem is thus the following:

Given the field dependence of the evaluator and his need for cognitive consistency, what is the effect of the course interest, course importance, and expected score in the course on course evaluation by the student?

The proposed research hypothesis is the following:

There is a relation between course interest, course importance, expected score in the course and evaluator field dependency on one side, and course evaluation on the other side.

## EXPERIMENTAL DESIGN

### Courses Evaluated

Data were gathered from participants of the Officer Professional Development Program (OPDP) of the Canadian Armed Forces. As stated in the administrative orders governing the program, the aim of the OPDP is "to broaden and deepen the Canadian Forces Officer's knowledge and understanding of the military profession beyond the specific technical expertise of classification training, and to contribute to the foundation of knowledge upon which further professional development will be built (CFAO 9-60).

Canadian Forces officers in the ranks of second lieutenant to major who were commissioned in 1971 or later are required to successfully complete six courses deemed to be fundamental to their professional development. The six courses are:

a. OPDP 2 - General Service Knowledge;
b. OPDP 3 - Personnel Administration;
c. OPDP 4 - Military Law;
d. OPDP 5 - Financial Administration and Supply;
e. OPDP 6 - National and International Studies; and
f. OPDP 7 - War and the Military Profession.

For officers commissioned in 1971 or later, completion of the six courses is a selection prerequisite for further professional development training, namely, attendance on the Command and Staff Course. The OPDP is a self-study program with successful completion of each course determined by measurement of performance on an objective (usually multiple choice) examination.

## Questionnaire Administration and Sample of Participants

At the beginning of this year OPDP management requested the cooperation of 29 major bases of the CF in order to gather participants' comments about the OPDP and to answer their questions. Teams of 3 or 4 members of the OPDP personnel visited these bases. Questionnaire administration was always performed by one of the two evaluating officers of the program, helped out by a colleague.

Some 800 candidates attended these meetings. However, 594 candidates fully responded to the evaluation questionnaire and the GEFT designed to measure the field dependency of the candidate. These candidates were kept for the research.

## Instruments

The OPDP evaluation questionnaire is essentially a modified version of the Course Comment Questionnaire and of the Personalized Course Analysis from Hogan (1976). These two instruments take up the question of student course evaluation by focusing attention on the student himself. In order to adapt these two instruments to OPDP needs they were discussed in-house and modified, and an instrument of 41 items was put together. This first form was administered in the fall of 1980 to 60 officers at a large CF base. From statistical analysis performed, comments gathered from these officers and their supervisors, an instrument of 21 items was kept. Factor analysis of this instrument came up with 6 factors with eigenvalues greater than 1.00 and explaining 74.5% of the variance.

The Group Embedded Figures tests (Oltman, Raskin, and Witkin, 1971) was administered to measure the perceptual field dependency of subjects, which is an indication of the psychological differentiation of subjects. The required task of the subject in the GEFT is to find a single figure hidden in a greater more complex one. There are 18 complex figures. A score of 13 or above makes a subject field independent while a score under 13 makes him a field dependent.

## Statistical Analysis

In operational terms our research hypothesis can be stated as the expectancy to find a greater relation between the two sets of variables for the field-independent group than the field dependent.

Two canonical correlation coefficients are calculated between the two sets of variables for each group of evaluators. Those coefficients can be defined as the maximal correlation produced by the linear combinations of the two sets of variables. In this present case

we will calculate part canonical correlation coefficients. Indeed, since in the experimental situation uneven pressures are put on candidates from their supervisors to be successful, we eliminate the effect of course importance on course interest and expected score. The second set of variables related to evaluation remains unchanged.

Once the part canonical correlation coefficients are calculated for the two groups, we compare these two coefficients. Even though we are not aware of any statistical comparison technique for this case, we can consider the p-value of each one of these coefficients and come up to some decision on the difference of these coefficients' magnitude.

The computer program used for the canonical analysis is the Canonical Analysis Program from Carlson and Timm (1976). This program does allow for part-canonical analysis.

## Data Analysis and Discussion

Before reporting the results of the canonical analysis, a consideration of descriptive statistics for the variables under study for both groups, is necessary. Table I reports for field independent (n=324) and field dependent (n=270) averages and standard deviations.

TABLE I
Descriptive Statistics of Variables

| | Field Dependent | | Field independent | |
|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. |
| Interest | 2.91 | 1.17 | 2.83 | 1.14 |
| Expected Score | 3.30 | 0.89 | 3.34 | 0.94 |
| Importance | 3.07 | 1.20 | 3.05 | 1.14 |
| General Cognitive Development | 2.77 | 0.99 | 2.64 | 0.96 |
| Specific Cognitive Development | 2.61 | 0.86 | 2.54 | 0.84 |
| Course Relevance | 2.76 | 0.88 | 2.65 | 0.86 |
| Personal Development | 2.64 | 0.87 | 2.58 | 0.87 |
| Course Content | 3.27 | 0.75 | 3.19 | 0.77 |

As can be observed in Table I there is not much variation between the FD and FI variables' means. However, as one can notice for the five evaluation variables, FD rated the course higher than FI but not significantly. It might be that FD are more lenient towards courses.

From the correlation matrix in Table II it can be observed that correlations between expected score and interest and the five evaluation variables are generally greater for FI. Also it seems that we are justified to partial out the effect of Course Importance if we consider its high correlation especially with course interest.

TABLE II
Correlation Matrix

| | Interest | Expected Score | Importance | Gen.Cog. Dev. | Spec. Cog. Dev. | Course Relevance | Pers Dev. | Course Content |
|---|---|---|---|---|---|---|---|---|
| Interest | 1.00 | | | | | | | |
| Expected Score | 0.28(FI)<br>0.34(FD) | 1.00 | | | | | | |
| Importance | 0.63<br>0.66 | 0.34<br>0.28 | 1.00 | | | | | |
| Gen.Dev. | 0.37<br>0.36 | 0.21<br>0.18 | 0.32<br>0.46 | 1.00 | | | | |
| Spec. Dev. | 0.47<br>0.41 | 0.42<br>0.29 | 0.44<br>0.51 | 0.61<br>0.71 | 1.00 | | | |
| Course Relevance | 0.47<br>0.49 | 0.34<br>0.24 | 0.52<br>0.58 | 0.57<br>0.69 | 0.69<br>0.75 | 1.00 | | |
| Pers. Development | 0.36<br>0.34 | 0.29<br>0.27 | 0.34<br>0.41 | 0.54<br>0.57 | 0.63<br>0.69 | 0.56<br>0.64 | 1.00 | |
| Course Content | 0.16<br>0.16 | 0.20<br>0.17 | 0.24<br>0.12 | 0.26<br>0.25 | 0.29<br>0.28 | 0.32<br>0.30 | 0.20<br>0.21 | 1.00 |

The canonical analysis itself has produced the part canonical correlation coefficients reproduced in Table III. Clearly the FI group have greater coefficients than the FD's. A comparison between the two significant coefficients shows a difference of some 12%.

TABLE III

Canonical Analysis

| | Part Canonical Correlation Coefficient | p-value |
|---|---|---|
| Field Independent | First: 0.364 | $p < 0.0001$ |
| | Second: 0.158 | 0.086 |
| Field Dependent | First: 0.237 | 0.033 |
| | Second: 0.128 | 0.324 |

Taking now only the first two coefficients which were significant, and calling U the set of variables Interest and Expected Score, and V the set of the five evaluation variables, we can look at correlations between variables of one set and the canonical variate of the other set.

In Table IV it can be seen that Expected Score is more related to V, the composite of evaluation variables, than Course Interest. This is true for either FD or FI. In Table V the picture is a bit more complex. Specific Cognitive Development is highly related to U, the canonical variate of Interest and Expected Score, and Course Content the least for FI. But for FD the least related variable is General Cognitive Development while all other variables have about the same coefficients.

TABLE IV
Correlations between U set variables
and V canonical variate

| | Field Independent | Field Dependent |
|---|---|---|
| Interest | 0.24 | 0.11 |
| Expected Score | 0.29 | 0.23 |

TABLE V
Correlations between V set variables
and U canonical variate

| | Field Independent | Field Dependent |
|---|---|---|
| General Cognitive Development | 0.21 | 0.07 |
| Specific Cognitive Development | 0.36 | 0.16 |
| Course Relevance | 0.24 | 0.11 |
| Personal Development | 0.25 | 0.17 |
| Course Content | 0.11 | 0.16 |

These findings were to be expected. First, the prediction of higher correlations for field independent between the course interest and expected score and their course evaluation was found to be true. As stated previously, because of their way of functioning, FI need to maintain a greater consistency between components of a situation. The situation is a given course, and they cannot tolerate cognitive dissonance between what they feel towards the course (Interest) and Expected outcome (Expected Score) and their rating of the course.

Even though there is such a need for FD people - there is a respectable relation between the two sets of variables - it is probably not as great for them. They do not have an analytical perception of the situation and its components but rather a global perception.

Secondly, it is quite interesting to consider evaluation of particular components. For instance a high correlation coefficient was to be expected for FI in their evaluation of course contribution to their Specific Cognitive Development. After all, they go much for details. But evaluation of Course Content is quite a vast area and a specific evaluation of each part of the course would have been more appropriate. We may consequently not be surprised at the low relation between FI interest and expected score and their evaluation of course content. As for FD one was not to expect much difference between the last four of the five evaluation variables - correlations vary between 0.11 and 0.17. The small relation for FD between General Cognitive Development and Interest - Expected Score can be understood in the very fact that since it is "General" and FD have global perception of a situation relation between this evaluation variable and students' interest and their expected score is even less seen than for the other evaluation variables.

## Summary of Findings

The major finding of this research is that there is a greater relationship between course-student related variables (Interest and Expected Score), partialling out the effect of Course Importance, and five Course Evaluation variables among field independent evaluators than among field dependent evaluators. It was suggested that since field independent people have an analytical perception of a situation, their need for cognitive consistency between course-student related variables and course evaluation variables drives them to maintain balance between these two sets of variables. Field dependent evaluators, being more global in their perception of a situation, do not have such a great need to maintain balance between the two sets of variables.

Other findings of the research are the following ones:

a. field dependents' course evaluations are higher (more positive but not significantly) than field independent ones;

b. among FI the greatest relationship with student-course variables composite was found with the evaluation of Specific Cognitive Development and the lowest with the evaluation of Course Content;

c. among FD the lowest relationship was found with the evaluation of General Cognitive Development, and all other relationships were found to be about equal; and

d. Expected Score on the course is more related to the evaluation composite variate than Course Interest.

## Conclusions

Even though studies in different educational settings are needed, the present research contributes toward adding to our comprehension of factors affecting course evaluations.

This research points out the relevance of cognitive styles and cognitive dissonance in the study of course evaluations. If, as suggested earlier, evaluation reports should be broken down by certain characteristics particular to groups, it might be worthwhile to use in such breakdowns cognitive styles and specific student-course variables. Also one should not report only a global evaluation score, but report scores for particular components being well identified factors in the overall course evaluation. Finally other studies are needed to assess the impact of these groupings in improving the agreement between course evaluators.

## References


Canadian Forces Administrative Orders No. 9-60. Department of National Defence, Ottawa, Ontario.

Carlson, J.E. and Timm, N.H. "Canonical Analysis Program". Psychometrika, 1976, 41, 159-176.

Costin, F., Greenough, W.T. and Menges, R.J. "Student ratings of college teaching: reliability, validity, and usefulness". Review of Educational Research, 1971, 41, 511-535.

Crawford, I. "Program Evaluation". Comment on Education, 1979, 9, 6-8.

Drummond, R.J. and McIntire, W.G. "The role of cognitive style in student evaluation of instruction". College Student Journal, 1977, 11, 220-223.

Feldman, K.A. "Consistency and variability among college students in rating their teachers and courses: a review and analysis". Research in Higher Education, 1977, 223-274.

Festinger, L. A theory of cognitive dissonance. Standord, C.A.: Stanford University Press, 1968.

Greenfeld, L. and Arbuthnot, J. "Field independence as a conceptual framework for prediction of variability in ratings of Others". Perceptual and Motor Skills, 1969, 28, 31-44.

Hogan, T.P. Student evaluation of courses in terms of personal development. Paper presented at the annual AERA meeting, San Francisco, California, April 1976.

Marsh, H.W. and Overall, J.U. "Long-term stability of students' evaluation: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses". Research in Higher Education, 1979, 10, 139-147.

Marsh, H.W., Overall, J.U. and Kesler, S.P. "Validity of student evaluations of instructional effectiveness: a comparison of faculty self-evaluations and evaluations by their students". Journal of Educational Psychology, 1979, 71, 149-160.

Messick, S. "The criterion problem in the evaluation of instruction: assessing possible, not just intended, outcomes". in The Evaluation of Instruction, Wittrock, M.C. and Wiley, D.E. (Ed.). Holt, Rinehart and Winston, Inc. 1970.

Musella, D. "Perceptual-cognitive style as related to self-evaluation and supervisor rating by student teachers". 1969. Eric Document #031 450.

Oltman, P.K., Raskin, E. and Witkin, H.A. Group Embedded Figures Test. Consulting Psychologists Press, Inc. Palo Alto, CA, 1971.

Penfield, D.A. "Student ratings of college teaching: rating the utility of rating forms". The Journal of Educational Research, 1971, 72, 19-22.

Shaffer, D.R. "The effects of cognitive style upon the inconsistency process". JSAS Catalog of Selected Documents in Psychology, 5, 283 (MS no. 1017), 1975.

Shaffer, D.R. and Hendrick, C. "Dogmatism and tolerance for ambiguity as determinants of differential reacting to cognitive inconsistency". Journal of Personality and Social Psychology, 1974, 29, 601-608.

Witkin, H.A. Goodenough, D.R. and Oltman, P.K. "Psychological differentiation: Current status". Journal of Personality and Social Psychology, 1979, 37, 1127-1145.

AD P001365

# PREDICTORS OF SUCCESS IN BASIC ENLISTED SUBMARINE SCHOOL

LCDR Earl H. Potter III, USCG
Assistant Professor of Psychology
United States Coast Guard Academy

## SUMMARY

Attrition, academic performance, and adaptive personality changes were examined in two studies of students in the U. S. Navy's Basic Enlisted Submarine School at Groton, Connecticut. Academic aptitude, the quality of the students' homelife, previous school experience, and test anxiety were the primary factors which predicted academic achievement. Significant positive changes over the course of the program were noted in test anxiety, hostility, and depression indices. Implications of these findings for interventions in the program designed to increase the rate of student success were considered.

Recent discussions of military preparedness in the popular press have highlighted the importance of our submarine force to national defense. While the launching of the first Trident-class submarine, the USS OHIO, was a striking technological achievement which signaled future increases in the strength of our submarine force, it is the crews of these potent weapons systems that will determine their effectiveness. The selection and training of these crews is, therefore, an issue vital to the national defense.

For each enlisted submariner, the first step toward becoming an effective crew member is graduation from the Basic Enlisted Submarine School (BESS). BESS, which is located in Groton, Connecticut, is a five-week training program which provides students with a basic understanding of submarine systems and life aboard a submarine. Graduates of this program must master a large body of technical information and demonstrate the ability to work as a team under pressure. This latter ability is tested in a simulation room which is fitted with leaking pipes. Students must repair the leaks before the room floods. Another test of performance under pressure is submarine escape training which requires the student to make a free ascent after entering the school's underwater escape training tower through a hatch at a depth of 50 feet. The Basic Enlisted Submarine School presents the student with a significant challenge. Graduation from the school represents a significant personal achievement--both academic and emotional.

The goals of BESS are three: 1) to graduate students with the technical foundation upon which to build future qualifications; 2) to screen out those persons who are not suitable for submarine service; 3) by providing an atmosphere of challenge to inspire and motivate graduates toward future performance. While much research has considered military attrition (Hand, Griffith, and Mobley, 1977) most of this research has attempted to account for the numbers of persons lost from training environments. The focus of this research is on determining the factors which contribute to success in the training environment. If one understands the goal by which success is measured in this particular training program, however, it is clear that the success of this program cannot be determined without a followup study of the performance of its graduates in the submarine fleet. This paper reports the first phase of an effort to examine the degree to which the goals of BESS are met. The second phase will examine the performance of successful graduates of BESS in the fleet. Together these studies should provide a better understanding of the role which initial training and screening plays in determining the effectiveness of the submariner. Hopefully, such an understanding will lead to improvements in the training program which will not be measured simply in terms of reduced attrition from BESS but in terms of the eventual effectiveness of submarine crews in the fleet.

## Theoretical Framework

Students come to BESS with a varying array of resources and liabilities. At BESS they confront an environment which poses certain challenges and offers certain supports. If the student perceives that these challenges exceed his resources, he experiences stress. This stress is greatest when the student's doubt about his ability to succeed is greatest (McGrath, 1976). The condition of greatest doubt is most likely to occur when the match between the student's resources and the demands of the environment is closest. In this condition the student's ability to cope with the resulting stress and draw on his resources to meet the demands of the environment will play a large part in determining eventual success or failure. Therefore, student characteristics which contribute to success in BESS are likely to include psychological factors which reflect the student's strategies for dealing with stress as well as more obvious resources such as cognitive ability. One such factor may be the student's desire for excitement and stimulation. Zuckerman, Kolin, Price, & Zoob (1964) have called this factor "Sensation Seeking" and related this tendency to a desire for challenge and experience. Another such factor may be a history of past success which would lead to positive expectations for success and increased motivation (Weiner, 1970).

Liabilities, too, include psychological factors related to the interpretation of demanding environments and coping strategies such as test anxiety, depression, and hostility. They also include aspects of the student's life such as an accumulation of stressful life events which has been related to poor academic (Carranza, 1972) and work (Harris, 1972) performance.

A BESS student is considered successful: 1)if he graduates; 2) to the degree that his final standing in his class (GPA) is higher than his peers; 3) that the resources which he takes to future assignments are greater than those with which he entered. The first two of these points are common criteria for all training research. The third criteria assumes that success in school will lead to greater expectation of future success and therefore greater motivation to attempt success. As a corollary, positive change in those psychological factors which are related to student responses to stress is also expected of a successful student.

## METHOD

Subjects

Two studies of students at BESS are summarized here. Study I was conducted in the winter of 1980 with 291 students from three class groups. Students ranged in age from 17 to 37 with 74% being younger than 21. Twenty-three percent of the students in Study I were not high school graduates. Study II was conducted in the winter of 1981 with 302 students from three class groups. Student ages were comparable to those in Study I (range 17-35, 73% less than 21) while only 16% were not high school graduates.

Testing

Students were given a rather lengthy package of instruments at the beginning of their second week at BESS and again just prior to graduation in their fifth week. A previous study (Antoni, 1980) demonstrated that the pre-test itself had no significant impact on post-test values. Testing took from one to two hours depending on student reading ability.

Instruments

Cognitive Ability. The measure of cognitive ability was the Armed Forces Qualification Test (AFQT). Mean AFQT scores for Studies I and II were 69.67 (range 41-98) and 68.75 (range 31-98).

Sensation Seeking. A desire for challenge was measured with a shortened version of Zuckerman's Sensation Seeking Scale (Zuckerman, et al, 1964). The possible values ranged from 0 to 22. Students in Studies I and II had values ranging from 0 to 21 ($\bar{X}$ = 12.6) and from 1-20 ($\bar{X}$ = 12.57).

Test Anxiety. Test anxiety was assessed using a short form of Sarason's Test Anxiety Scale (Sarason, 1978b). Possible values ranged from 0 to 25 while student scores in Studies I and II ranged from 0 to 25 ($\bar{X}$ = 9.70) and from 0 to 24 ($\bar{X}$ = 9.13).

Stressful Life Events. Students completed the Life Experiences Survey (Sarason, 1978a) which lists 47 significant life events. Weights (0 to 3) and positive/negative evaluations are assigned by the student. This method allows the computation of several scores including a total of good and bad events and a weighted total of good and bad events. As Sarason suggests, good life events seem to have little impact on later success. The total number of bad life events and the weighted total do seem to be related, to some degree, to student success. Students in Studies I and II report 0 to 45 ($\bar{X}$ = 5.85) and 0 to 27 ($\bar{X}$ = 3.67) bad life events. Weighted scores range from 0 to 82 ($\bar{X}$ = 10.25) for Study I and from 0 to 44 ($\bar{X}$ = 6.52) for Study II.

Anxiety, Hostility, Depression. Zuckerman and Lubin's (1965) Multiple Affect Adjective Checklist (MAACL) yields three sub-scores: anxiety, hostility, and depression. Anxiety scores for students in Studies I and II ranged from 0 to 18 ($\bar{X}$ = 5.82) and from 0 to 20 ($\bar{X}$ = 5.52) with a possible range of 0 to 21. Possible values for hostility ranged from 0 to 28. Student values ranged from 0 to 19 ($\bar{X}$ = 7.33) in Study I and 0 to 27 ($\bar{X}$ = 8.13) in Study II. For depression the possible range was 0 to 40. In Study I, students ranged from 0 to 31 ($\bar{X}$ = 10.21); in Study II, the range was 0 to 34 ($\bar{X}$ = 10.78).

Anger. A student's tendency to respond to frustration with anger was assessed with Novaco's Angel Scale (1975). Possible values ranged from 80 to 400. Students in Studies I and II ranged from 80-385 ($\bar{X}$ = 261.2) and 80-380 ($\bar{X}$ = 260.0).

Past Success. The student's perception of his past academic success was measured with a one-item, five-point likert scale which asked the student to rate his experience in school from good (1) to terrible (5). Students in Studies I and II ranged from 1 to 5 with means of 2.07 and 1.91.

Quality of Homelife. A one-item, five-point Likert scale was also used to assess homelife. Student values ranged from 1 (good) to 5 (terrible) with means of 1.70 and 1.69.

## RESULTS

### Graduation/Non-graduation

BESS students were divided into three groups: 1) those students who graduated with the class with which they started school; 2) those students who were setback into later classes to correct academic deficiences; and, 3) those students who did not graduate. The first set of analyses considered the differences between graduates and non-graduates with respect to psychological characteristics (AFQT, Test Anxiety, MAACL-Anxiety, MAACL-Hostility, MAACL-Depression, and Sensation Seeking) and aspects of the student's history (homelife quality, school experience, and bad life experiences--weighted and unweighted). Tables 1 and 2 summarize the results of the two-tailed t-tests between the independent groups, graduate and non-graduate, for Studies I and II.

Graduates in both studies have higher AFQT scores, lower test anxiety, more positive past experiences (school and home life) and evidence less anxiety, hostility, and depression. In Study I, sensation seeking is greater for graduates ($p = .009$); in Study II, the difference between the sensation seeking scores of graduates and non-graduates is in the same direction but not significant. In Study II, students with a greater number of recent bad life experiences (weighted and unweighted) are more likely to be graduates. While the differences between graduates and non-graduates with respect to life events is in the same direction for Study I, the difference is not significant.

These results suggest that the graduate has greater cognitive ability and is more likely to have had past experiences which lead him to expect success. The fear of failure which may be characteristic of the non-graduate is evidenced by greater test anxiety, hostility, and depression. These findings are significant because of the increase in the percentage of non-high school graduates in the submarine force which seems to coincide with the advent of the all volunteer armed forces. Students with a record of past failures may be vulnerable to defeat in a demanding academic program. On the other hand, if expectations and experience and not academic ability alone determine success, interventions may be possible which will increase the chances of these students for success.

STUDY I

Differences Between Graduates and Non-Graduates
of Basic Enlisted Submarine School

| | GRAD | NON-GRAD | P | |
|---|---|---|---|---|
| Test Anxiety | 9.51 | 11.7 | .014 | |
| AFQT | ____ | ____ | ____ | |
| Bad Life Experiences | 5.53 | 7.43 | .086 | |
| B LE Weighted | 9.65 | 12.79 | .173 | |
| Homelife | 1.63 | 2.04 | .019 | |
| School Experience | 2.00 | 2.36 | .047 | |
| MAACL-A | 5.48 | 7.18 | .016 | |
| MAACL-D | 9.80 | 12.18 | .044 | |
| MAACL-H | 6.87 | 8.75 | .017 | |
| Sensation Seeking | 12.90 | 10.86 | .009 | Table 1 |

STUDY II

Differences Between Graduates and Non-Graduates
of Basic Enlisted Submarine School

| | GRAD | NON-GRAD | P | |
|---|---|---|---|---|
| Test Anxiety | 8.92 | 13.78 | .008 | |
| AFQT | 69.40 | 58.29 | .005 | |
| Bad Life Experiences | 3.42 | 5.81 | .028 | |
| B LE Weighted | 5.78 | 13.82 | .005 | |
| Homelife | 1.59 | 2.67 | .009 | |
| School Experience | 1.81 | 2.89 | .007 | |
| MAACL-A | 5.07 | 8.17 | .037 | |
| MAACL-D | 9.99 | 15.39 | .031 | |
| MAACL-H | 7.56 | 12.06 | .022 | |
| Sensation Seeking | 12.67 | 11.27 | .131 | Table 2 |

Academic Performance

Table 3 summarizes regression equations which predict GPA using AFQT, test anxiety (TA), the AFQT x TA interaction, school experience and homelife quality. In Study I, the multiple R for this set of variables is .50 accounting for 25% of the variance in GPA. In Study II, the multiple R is .57 which accounts for 32% of the variance in GPA. In general, students with higher AFQT's and lower test anxiety tend to have higher GPA's. These effects are additive; the interaction term is not significant in either Study I or Study II. The relationship of school experience and homelife to academic performance is ambiguous both being significant in Study II but not in Study I.

Again, students who do well have greater cognitive ability and a greater expectation of success. The relationship of test anxiety to GPA is independent of AFQT and, therefore, offers a promise of possible intervention to improve student performance.

STUDY I

Hierarchical Multiple Regression on GPA

| VAR | MULT | $R^2$ | $\Delta R^2$ | Significance of $\Delta R^2$ | Simple R |
|---|---|---|---|---|---|
| AFQT | .44 | .19 | .194 | $<.001$ | .44 |
| Test Anxiety | .48 | .23 | .037 | $<.001$ | -.31 |
| AFQT·TA | .487 | .24 | .005 | NS | -.16 |
| School Experience | .498 | .25 | .011 | NS | -.19 |
| Homelife | .500 | .25 | .002 | NS | -.15 |

STUDY II

| | | | | | |
|---|---|---|---|---|---|
| AFQT | .45 | .21 | .21 | $<.001$ | .45 |
| Test Anxiety | .523 | .273 | .067 | $<.001$ | -.38 |
| AFQT·TA | .524 | .274 | .001 | NS | -.23 |
| School Experience | .54 | .29 | .015 | .025 | -.23 |
| Homelife | .57 | .32 | .036 | $<.001$ | -.30 |

Table 3

## Adaptive Personality Change

Those students who graduate from BESS experience a significant personal success--one of the few for some students. If success leads to greater expectations for future success and can be related to a more positive evaluation of self, than facing and meeting this challenge of BESS should lead to change in personality measures which could be characterized as adaptive. In other words, success in BESS may be the best influence on the student's mix of resources and liabilities and may lead to a greater chance of future success. Table 4 summarizes paired t-tests for students who graduate from BESS in Studies I and II. Graduates in both studies are less anxious, more desirous of challenge, less depressed, and less easy to anger than they were when they entered BESS. It is noteworthy that academic success results in a decrease in test anxiety which, our results suggest, should be reflected in better performance in more advanced schools.

Change in Personality Scale Values

| | STUDY I | | STUDY II | |
|---|---|---|---|---|
| Sensation Seeking | 12.98 | p=.001 | 12.51 | p<.001 |
| | 13.50 | | 13.08 | |
| Test Anxiety | 9.28 | p<.001 | 9.13 | p<.001 |
| | 6.67 | | 6.81 | |
| Anger Scale | 257.6 | p<.001 | 252.7 | p<.001 |
| | 243.9 | | 239.3 | |
| MAACL-A | 5.55 | p=.001 | 5.52 | p=.034 |
| | 4.89 | | 5.21 | |
| MAACL-D | 9.80 | p<.05 | 10.79 | p<.01 |
| | 9.13 | | 10.12 | |

Table 4

## DISCUSSION

This paper has briefly reviewed some of the preliminary findings of a two-phase effort to understand the role which the Basic Enlisted Submarine School plays in determining the success of enlisted submariners in the submarine fleet. Clearly the cognitive and psychological resources which the students bring to the school are major determinants of the student's success in BESS. Foreknowledge of these characteristics should allow BESS to target students for extra attention and assistance. Some potential students may also be identified as poor risks for completion before they begin BESS. The most important finding, however, may be that students leave BESS more prepared for success than when they started BESS. This finding supports efforts to encourage and support students who might otherwise fail.

Given adequate cognitive skills, the student's perception of the degree of threat posed by the environment and his strategy for coping with the resulting stress does seem to play an important role in determing success in BESS. Interventions which focus on improving students' coping strategies do seem to offer hope for increasing the rate of success in BESS and, as a consequence, in the submarine fleet as well.

## REFERENCES


Antoni, M.H. An analysis of test effects and the susceptibility of several types of tests to these effects. Unpublished, 1980.

Carranza, E. A study of the impact of life changes on high school teacher performance in the Lansing School District as measured by the Holmes and Rahe Schedule of Recent Experiences. (Doctoral dissertation, Michigan State University, 1972). Dissertation Abstracts International, 1973, 33, 4995A-4997A. (University Microfilms No. 73-5340)

Hand, H. H., Griffeth, R. W., and Mobley, W. H. Military enlistment, reenlistment, and withdrawal research: A critical review of the literature. Columbia, S.C.: Center for Management and Organizational Research, University of South Carolina, Office of Naval Research Technical Report No. 3 (ONR: TR3), ADA048955, April 1977.

Harris, P. W. The relationship of life change to academic performance among selected college freshmen at varying levels of college readiness. (Doctoral dissertation, East Texas State University, 1972). Dissertation Abstracts International, 1973, 33, 6665A-6666A. (University Microfilms No. 73-14285)

McGrath, J. E. Stress and behavior in organizations. In M.D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology, Chicago: Rand McNally, 1976.

Novaco, R. W. Anger control: The development and evaluation of an experimental treatment. Lexington, Massachusetts: D.C. Heath, Lexington Books, 1975.

Sarason, I. G. The test anxiety scale: Concept and research. In C. D. Spielberger and I. G. Sarason (Eds.) Stress and Anxiety (Vol. 5). New York: J. Wiley & Sons, 1978a.

Sarason, I. G., Johnson, J. H., and Siegel, J. M. Assessing the impact of life changes. Development of the Life Experiences Survey. Journal of Consulting and Clinical Psychology, 1978b, 46, 932-946.

Weiner, B. New conceptions in the study of achievement motivation. In B. A. Maher (Ed.), Progress in Experimental Personality Research (Vol. 5). New York: Academic Press, 1970.

Zuckerman, M., Kolin, G. A., Price, L., & Zoob, I. Development of a Sensation-Seeking Scale. Journal of Consulting Psychology, 1964, 28 477-482.

Zuckerman, M. & Lubin, B. Manual for the Multiple Affect Adjective Check List. San Diego, California: Educational and Industrial Testing Service, 1965.

AD P001366

Rankin, William C. & McDaniel, William C. US Navy Training Analysis and Evaluation Group, Orlando, Florida. (Wed. A.M.)

## Aviation Training Task Proficiency: A Probabilistic Approach

The assessment of flight task proficiency by instructor pilots in Navy training squadrons poses a problem in determining the point at which performance on most subsequent trials would be judged "proficient." Because of the inherent variability in trainees, instructor judgments, and training tasks, a probabilistic decision must be made. It is proposed that Wald's technique of sequential sampling applied to trail-by-trial performance data (instructor judgments of "proficient vs. not proficient") can be adapted to the flight training situation. Sequential sampling models can be used consistently by letting the instructor focus on a single binary decision each time the trainee performs a task. Then, the appropriate model determines with controllable risks, when to terminate training on the task. The model(s) does this because a sampling threshold has been reached and there is a high probability that most or all subsequent performance for a given task will be judged proficient.

# AVIATION TRAINING TASK PROFICIENCY: A PROBABILISTIC APPROACH

William C. Rankin and William C. McDaniel

Training Analysis and Evaluation Group (TAEG)
Orlando, Florida 32813

Determination of the proficient performance of aircraft flying tasks continues to be a subjective judgment made by instructor pilots. Current practice in training squadrons consists of "flights" during which a subset of tasks from the training syllabus are performed a varying number of times by the pilot trainee at the discretion of the instructor pilots. During or shortly after each flight, the instructor pilot "grades" the pilot trainee on the tasks performed using a standard scale but also employing his own personal criteria. While instructors differ in their personal rating bias (hard-easy), they attempt to grade in terms of "average performance at this stage of training." It is usual for the pilot trainee to be exposed to several different instructor pilots. After a specified minimum number of flights, and a recommendation by an instructor pilot, the pilot trainee is scheduled for a final "check flight." His performance on selected tasks is graded by an instructor pilot acting in the independent role of "check pilot." Should the pilot trainee not perform the flight consonant with the standards of performance expected of him by the "check pilot," he is rescheduled for additional "check flights" until he is deemed proficient.

Student exposure to training tasks can be variable due to instructor differences and varying performance standards. In addition, each individual pilot trainee exhibits variability in successive performances on complex procedural and psychomotor tasks. This variability of skilled task performance has been well documented (Fitts and Posner, 1968). Further compounding this problem of inconsistent performance, the pilot trainee is transitioning from a level of performance well below the required level to a required standard of performance. This transition reflects different learning rates by the individual pilot trainees. Learning rates are also highly variable within and between individuals (Sidman, 1960). It is quite obvious that determination of asymptotic performance commensurate with desired performance standards is difficult to ascertain using the current practice.

## PROFICIENCY GRADING SYSTEM

In a series of studies conducted by the Training Analysis and Evaluation Group (TAEG) to determine the effectiveness of Device 2F87F (P-3 Operational Flight Trainer) in the FRS, the inadequacies of current grading procedures were recognized (Browning, Ryan, Scott, and Smode, 1977; Browning, Ryan, and Scott, 1978). To overcome these inadequacies, the TAEG instituted a "proficiency grading system." The system provided a clearer picture of the trainee's flight task performance in both simulator and aircraft training. The proficiency grading system still required a subjective judgment by instructor and check pilots. However, the instructors graded task performance against a precise standard: "P was defined as performance estimated to be equivalent to that required to demonstrate competence in that task on the conventional FLY 6 check" (Browning, et al., 1977, p. 20). This standard focuses on the required terminal level of performance; i.e., the objective of training. Actual grading of performance was accomplished using a dichotomous scale. Task performance that met or exceeded the standard was recorded as "P"; task performance that did not meet the standard was recorded as "1." The proficiency grading introduced

by the TAEG had a further requirement. Performance was graded each time the task was performed and this series of graded trials was recorded and kept in the sequence of presentation. The procedure of grading each task trial as it was performed eliminated the requirement for the instructor to make a summary judgment of task proficiency based on pilot trainee performance of successive task trials during a flight.

The performance standard used in the proposed system is defined as task performance estimated to be equivalent to that required to earn an adjective rating of "Qualified" and/or a numerical score of 4 on the Naval Air Training and Operating Procedures Standardization (NATOPS) Program flight evaluation. The proposed system uses the same proficiency grading procedure as discussed previously. Although the grading procedure increases the precision, it does not reduce several sources of variability in trainee performance; e.g., task difficulty and learning rates.

The proficiency grading procedure results in a task performance or training protocol for each task. Two hypothetical trainee records (protocols from the same trainee) are shown in table 1.

TABLE 1. HYPOTHETICAL TASK PERFORMANCE OF ONE TRAINEE FOR TWO DIFFERENT TASKS

| Task | Training Protocol |
|---|---|
| Task A | 11P1P1PPP1PP |
| Task B | 1PPPPPPPPPP |

It could be inferred that "Task A" is more difficult than "Task B" or it could be inferred that the trainee is more proficient on "Task B" than "Task A."

Table 2 contains examples of trainee task performance protocols for two different kinds of tasks and hypothetical task protocols for a trained pilot. The pilot trainees exhibit different protocols initially (more "1's" than "P's") but the variability eventually will diminish. Learning rates differ among tasks as shown by comparing Task A with Task B. During later flights/sessions the protocols for the pilot trainee are not readily distinguishable from those of a trained pilot. A procedural problem remains in determining when task performance protocols for trainees matched the protocols of trained pilots.

TABLE 2. COMPARISON OF HYPOTHETICAL TASK PERFORMANCE PROTOCOLS FOR TWO DIFFERENT TASKS AND TWO LEVELS OF AVIATOR PROFICIENCY

| Task/Aviator | Training Protocol During Flights/Sessions | | | | | |
|---|---|---|---|---|---|---|
| | One | Two | Three | Four | Five | Six |
| TASK A | | | | | | |
| Pilot Trainee | 111 | P11 | 1P1 | 1PP | PPP | PP |
| Trained Pilot | PPP1 | PPP | 1PP | P | P1 | PPP |
| TASK B | | | | | | |
| Pilot Trainee | 11 | 1P | P | PPP | PP | PP |
| Trained Pilot | PP | PPP | P1 | PP | P | PP |

The essence of the problem lies in assessing, with a specified degree of confidence, the point at which proficiency has been obtained.

Several ways to deal with the problem were explored. Two approaches were found in previous research concerned with proficiency assessment. The first approach was to define arbitrarily the point at which proficiency was attained by the following rule:

> (1) over 50 percent of the trials (for a given task) on any flight had to be "P" and (2) at least 50 percent of the trials were $\overline{P}$ on all subsequent flights (Browning, et al., 1978, p. 23).

The second approach was used in the evaluation of the Initial Entry Rotary Wing Flight Training Program by the Army (USAAVNC Evaluation Team, 1979). The tasks were graded by daily performance rather than by individual trials; however, the approach used to determine proficiency could also be incorporated with graded trials.

> The point of principal concern was the training day on which the student achieved proficiency on each maneuver. Achievement of maneuver proficiency was defined as that training day on which the third successive (+) grade on the maneuver was given the student. That is, the student was required to perform a maneuver in accord with established USAAVNC standards on three successive occasions before he was judged to be proficient on that maneuver (USAAVNC Evaluation Team, 1979, p. 21).

While both of the above approaches are logical, objective, and expedient, they are faulty. Both require training protocols that include initial and final levels of proficiency to make accurate performance determinations. In other words, they are "after the fact" rather than predictive. Another flaw is that an arbitrary number of "P" trials is not realistic across all tasks due to differences in task difficulty. In addition, these approaches may not accommodate situations where only a small number of training trials are given or where there are wide differences in learning rates of trainees. Finally, the instructor's judgment may be biased if he has knowledge of an arbitrary decision rule.

## PROPOSED APPROACH

SEQUENTIAL METHOD. Both of the above approaches require a sample of trials of trainee performance before the rule can be applied. An alternate approach would be to examine trials taken one at a time and accumulate the information for input into the decision model (Hoel, 1971). Using this approach, one would expect to be in a better position to make decisions than if no attempt were made to look at the data until a sample of fixed size had been taken.

One sequential method that may be used as a means for making statistical decisions with a minimum sample was introduced by Wald (1947). Probability ratio tests and corresponding sequential procedures were developed for several

statistical distributions. One of the tests, the binomial probability ratio test, was formulated in the context of a sampling procedure to determine whether a collection of a manufactured product should be rejected because the proportion of defectives is too high or should be accepted because the proportion of defectives is below an acceptable level. The sequential testing procedure also provides for a postponement of decisions concerning acceptance or rejection. This deferred decision is based on prescribed values of alpha (a) and beta (B). Alpha (a) limits errors of declaring something "True" when it is "False"" (Type I error). Beta (B) limits errors of declaring something "False" when it is "True" (Type II error).

In an industrial quality control setting, the inspector needs a chart similar to figure 1 to perform a sequential test to determine if a manufacturing process has turned out a lot with too many defective items or whether the proportion of defects is acceptable. As each item is observed, the inspector plots a point on the chart one unit to the right if it is not defective, one unit to the right and one unit up if the item is defective. If the plotted line crosses the upper parallel line, the inspector will reject the production lot. If the plotted line crosses the lower parallel line, the lot will be accepted. If the plotted line remains between the two parallel lines of the sequential decision chart, another sample item will be drawn and observed/tested.

This sequential sampling procedure decision model has been previously used in educational and training settings. Ferguson (1969) used the sequential test to determine whether individual students should be advanced or given remedial assistance after they completed learning modules of instruction. Similarly, Kalisch (1980) employed the sequential test for an Air Force Weapons Mechanics Training Course (63ABR46320) conducted at Lowry Air Force Base, Colorado. Results from both applications of sequential testing indicate greater efficiency than for tests composed of fixed numbers of items. It appears sequential testing may substantially reduce testing time. It should be noted that in the industrial quality control setting, sampling occurs after the manufacturing process. In the educational and training applications cited above (Ferguson, 1969 and Kalisch, 1980), sequential sampling occurred <u>after</u> the learning period. In the proposed system, the sequential sampling occurs <u>during</u> the learning period and eventually terminates it.

MODEL PARAMETERS. The decision model can be described as consisting of decision boundaries. Referring to figure 1, the parallel lines represent those decision boundaries. Crossing the upper line, or boundary, results in a decision to "Reject Lot"; crossing the lower line, or boundary, results in a decision to "Accept Lot." In the proposed system, these decision boundaries translate to "Proficient" and "Not Proficient." Calculations of the decision boundaries require four parameters. These four parameters are:

$P_1$ Lowest acceptable proportion of proficient trials (P) required to pass the NATOPS flight evaluation with a grade of "Qualified." Passage of the NATOPS flight evaluation is required to be considered a trained aviator in an operational (fleet) squadron.

$P_2$ Acceptable proportion of proficient trials (P) that represent desirable performance on the NATOPS flight evaluation.

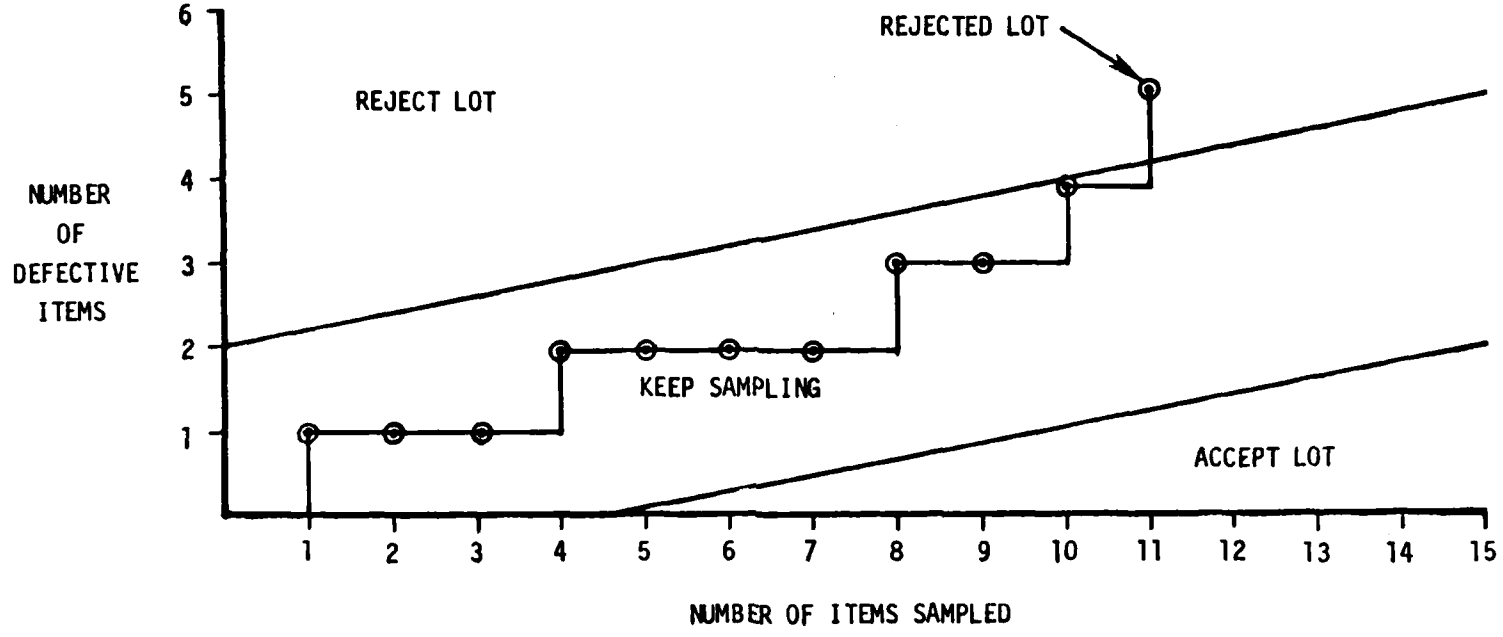



Figure 1. Hypothetical Sequential Sampling Chart

Alpha (a) The probability of making a TYPE I decision error (deciding a student is proficient when in fact he is not proficient).

Beta (B) The probability of making a TYPE II decision error (deciding a student is not proficient when in fact he is proficient).

Parameter setting is a crucial element in the development of the sequential sampling decision model. Kalisch (1980) outlines three methods for selecting proficient/not proficient performance ($p_2/p_1$ values) as:

> Method 1--External Criterion. Individuals are classified as masters, non-masters, or unknown on the basis of performance on criteria directly related to the instructional objectives.
>
> Method 2--Rationalization. Experts in the subject area who understand the relation of the training objectives to the end result; e.g., on-the-job performance, select the $p_2$ and $p_1$ values to reflect their estimation of the necessary levels of performance.
>
> Method 3--Representative Sample. The scores of prior trainees, who demonstrate the entire range from extremely poor to exemplary performance on objectives, are used to estimate $p_2$ and $p_1$.

Selection of values for $P_1$ and $P_2$ for the proposed decision model incorporated Method 1 for setting of $P_1$ and Method 3 for setting of $P_2$.

The selection of alpha (a) and beta (B) should be based on the criticality of accurate proficiency decisions. Small values of alpha (a) and beta (B) require additional task trials to make decisions with greater confidence. Factors that are important in selecting values for alpha (a) and beta (B) are outlined below:

1. Alpha (a) values

   a. Safety--potential harm to the trainee or to others due to the trainee's actual non-mastery of the task.

   b. Prerequisite in Instruction--potential problems in future instruction, especially if the task is prerequisite to other tasks.

   c. Time/Cost--potential loss or destruction of equipment either in training or upon fleet assignment.

   d. Trainee's View of the Training--potential negative view by trainee when classified as proficient although the trainee lacks confidence in that decision. Also, after fleet assignment if previous training has not prepared him sufficiently the trainee may also have a negative view of the training program.

2. Beta (B) values

   a. Instruction--requirement for additional training resources (personnel and materials) for unnecessary training in case of misclassification as not proficient.

   b. Trainee Attitudes--the attitude of trainees when tasks have been mastered yet training continues; trainee frustration; corresponding impact on performance in the remainder of the training program and fleet assignment.

   c. Cost/Time--the additional cost and time required for additional training that is not really needed.

Alpha (a) and beta (B) values used in the proposed decision model were arbitrarily selected as .10. A confidence level of 90 percent in decisions made by the model appears reasonable when the previously discussed factors are considered. As rigorous field testing of the model is conducted, these parameters may be modified as indicated by empirical evidence and command policy. At present, values of .10 appear quite reasonable. After the model parameters have been selected, calculation of the decision boundaries may be accomplished using the Wald Binomial Probability Ratio Test.

To illustrate the differences in task difficulty, two tasks were selected from the HS-1 training syllabus, and the decision models for these tasks were calculated. To further show how the decision models serve to aid in making proficiency decisions, task protocols of a pilot trainee are imposed on the model.[1] Figure 2 shows the model for the task "Running Takeoff," and figure 3 shows the model for the task "Free Stream Recovery."

Empirical data reflect a relative difference in task difficulty. The sample of NATOPS evaluation scores indicates the proportion of "Qualified" scores on the Running Takeoff task was .92, while the proportion of "Qualified scores on the Free Stream Recovery task was .77. This relative difference in task difficulty is represented in the model as differences between the slopes and the widths between the parallel lines of the two models. In the case of the Free Stream Recovery task (figure 3), the slopes are less steep (indicating more trials to reach proficiency) and the parallel lines are farther apart (indicating there will typically be more uncertainty about individual trials before a decision can be reached).

In these examples, the probability of making decision errors (both type I and type II) as indicated earlier was set at .10 for both tasks. If this level of confidence was increased (lower values of alpha (a) and beta (B)), the region of uncertainty would also increase. The overall result is that more trials are required to make a decision with increased confidence.

Both models, then, reflect rather well the true state of affairs between different tasks and their impact on a rational decision process. The differences in task difficulty relate directly to differences in the model parameters.

[1]Actual trial data for a pilot trainee undergoing training at HS-1, NAS Jacksonville, FL.

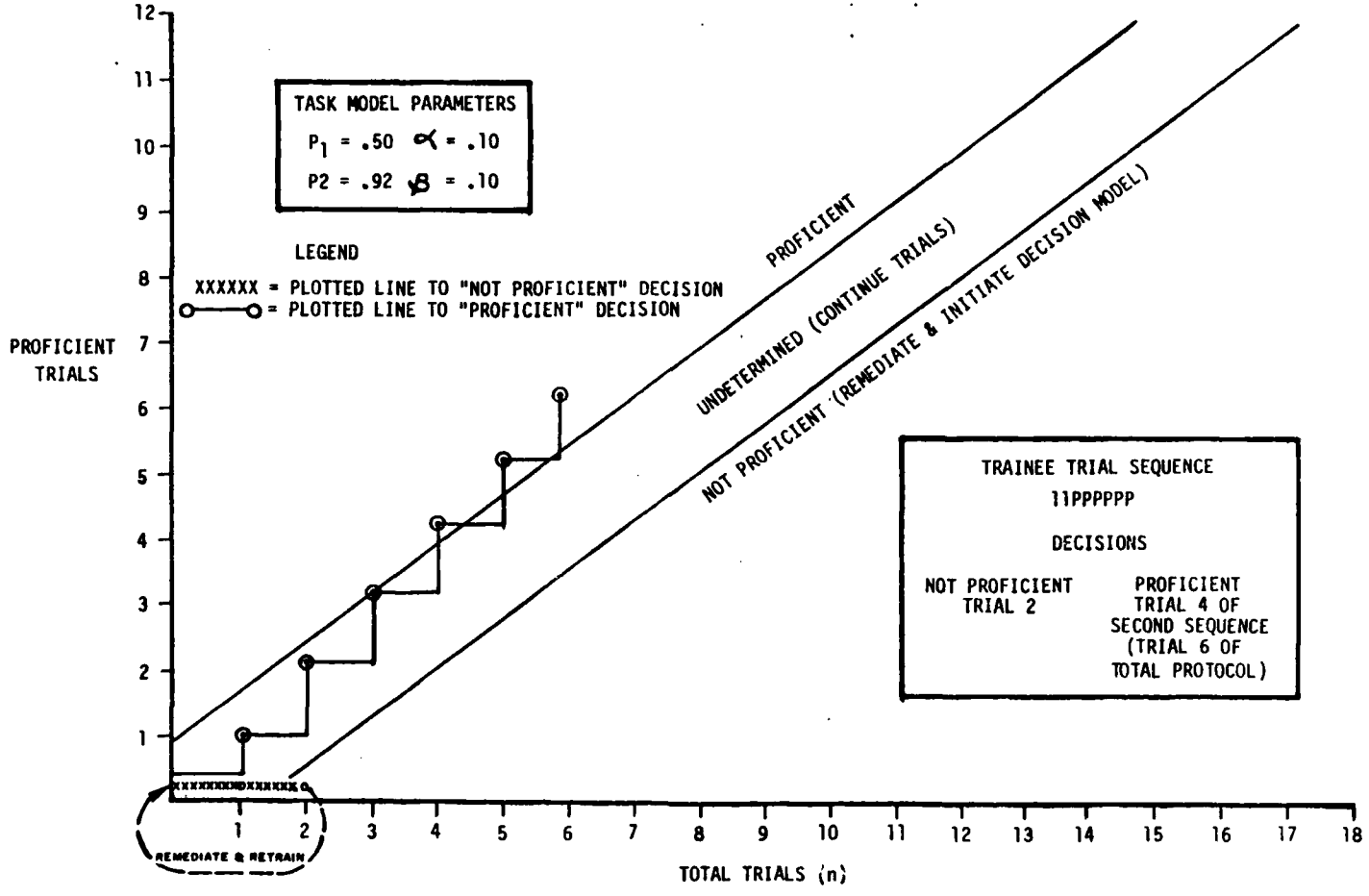



Figure 2. Sequential Sampling Decision Model for Running Takeoff Task

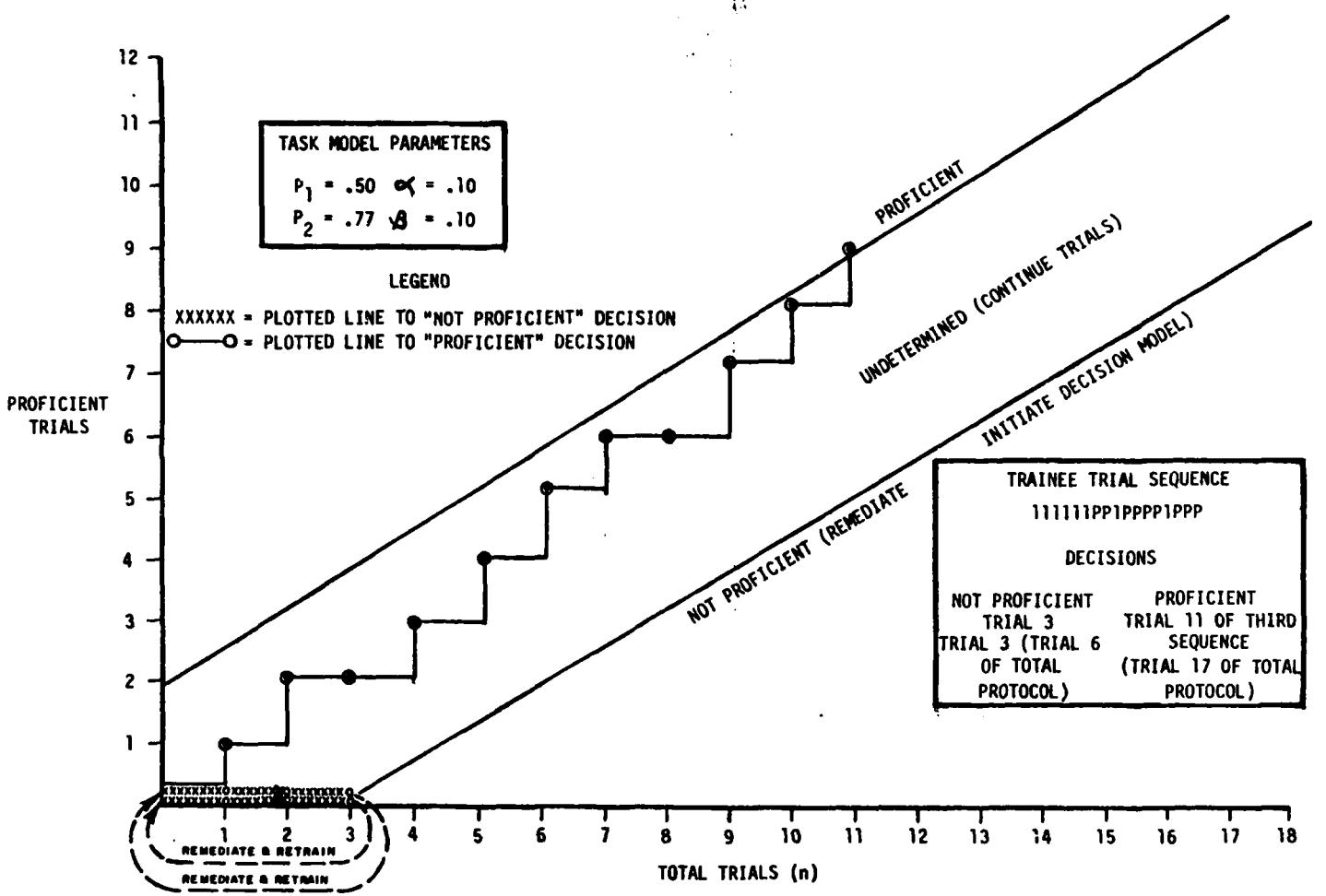



Figure 3. Sequential Sampling Decision Model for Freestream Recovery Task

Figures 2 and 3 also show the decisions reached by the model on student performance. The student received a total of eight trials on the Running Takeoff task during the training program. The sequence of graded trials and the graphical plots of the sequence are shown in figure 2. The first two trials were judged to be below the standard of performance. On the second trials the decision model indicated the student was "Not Proficient" and logically should be given remedial or additional training. The sequence is initiated again on trial three, and on the fourth trial of that sequence (sixth trial given) the model decision was "Proficient."

Figure 3 shows the protocol for the Free Stream Recovery task. Perhaps because of slower acquisition of a more difficult task, two decisions were made declaring the student "Not Proficient" in the earlier sessions of task exposure. The model does show that more task trials were required before a decision could be made about proficiency. This can be attributed to increased task difficulty and variability of performance.

## PLANNING FOR IMPLEMENTATION

A study is currently underway to test the concept at the East Coast SH-3 FRS, HS-1, NAS, Jacksonville, Florida. The study is broadly planned as follows:

1. identify a syllabus of specific training tasks

2. establish proficiency decision model parameters from prior data collected at HS-1

3. train instructors to render performance judgments on task trials; i.e., was performance a "1" or a "P"?

4. collect data on each trainee's task performance by trial

    a. The current decision model (unique to each instructor) will determine when to terminate training the task.

    b. Instructors and training managers will have no knowledge of models decisions regarding task proficiency.

5. analytically compare the models using final performance criterion (NATOPS flight evaluation performance)

6. make recommendations as to feasibility.

Assuming the results of the study are promising, it will be desirable to look toward incorporating or designing a CMI system for which these models are readily amenable. Implementing the proficiency determination concept advanced in this paper can only be done efficiently with on-line computer support. The work of Ferguson (1969) and Kalisch (1980) would have been virtually impossible without on-line computer support. Also planned are future efforts to determine the range of applicability to other FRS settings.

POSTNOTE:

For a fuller description of this work see Rankin, W. and McDaniel, W. Computer Aided Training Evaluation & Scheduling (CATES) System: Assessing Flight Task Proficiency. TAEG Report No. 94. December 1980. TAEG, Orlando, FL 32813 (AD A095007)

## REFERENCES


Browning, R. F., Ryan, L. E., Scott, P. G., and Smode, A. F. Training Effectiveness Evaluation of Device 2F87F, P-3C Operational Flight Trainer. TAEG Report No. 42. January 1977. Training Analysis and Evaluation Group, Orlando, FL 32813 (AD A035771)

Browning, R. F., Ryan, L. E., and Scott, P. G. Utilization of Device 2F87F OFT to Achieve Flight Hour Reductions in P-3 Fleet Replacement Pilot Training. TAEG Report No. 54. April 1978. Training Analysis and Evaluation Group, Orlando, FL 32813. (AD A053650)

Ferguson, R. The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction. Unpublished dissertation, University of Pittsburgh. 1969.

Ferguson, R. " A Model for Computer-Assisted Criterion-Referenced Measurement." Education. 1970. 91. pp. 25-31.

Fitts, P. M. and Posner, M. J. Human Performance. Belmont, CA: Brooks and Cole, 1968.

Hoel, P. G. Introduction to Mathematical Statistics. New York: John Wiley & Sons, Inc. 1971.

Kalisch, S. J. Computerized Instructional Adaptive Testing Model: Formulation and Validation. AFHRL-TR-79-33. February 1980. Air Force Human Resources Laboratory, Brooks AFB, TX.

Sidman, M. Tactics of Scientific Research. New York: Basic Books. 1960.

United States Army Aviation Center Evaluation Team. Evaluation of the 175/40 Initial Entry Rotary Wing Flight Training Program. TR 79-02. May 1979. U.S. Army Aviation Center, Fort Rucker, AL.

Wald, A. Sequential Analysis. New York: John Wiley & Sons, Inc. 1947. (Reprinted by Dover Publications. 1973.)

AD P001367

Roll, Charles Robert, Jr. and Berger, B. Michael, Selective Service System, Washington, D.C.

## "Traffic Cop" - Automated Prediction and Control of Registrants to AFEES

Possible military mobilization poses significant problems associated with peacetime planning for and mobilization processing of inductees through Armed Forces Examining and Entrance Stations. 'Traffic Cop' is an automated process which predicts, in peacetime, the likely flow of registrants at various levels of induction call, and permits control of the actual registrant flow following mobilization.

Since shutdown of the entire system is possible due to regional bottlenecks, Selective Service and the Military Enlistment Processing Command (MEPCOM) must be able to allocate resources and "swing" inductees to larger AFEES if workloads exceed station capacities.

The presentation discusses the likelihood of overloading AFEES based upon induction call levels and AFEES rated capacities. The methodology of the flow prediction in the 'Traffic Cop' model is explained as is the method for identifying 'swing' AFEES and shifting the workload via the computer.

The mechanics of issuing the induction call via Mailgram, and the lead-time for decision-making on registrant flow are discussed.

Computer graphic outputs demonstrate the utility of the system for planning and operational use. The presentation concludes with a discussion of the installation of the model on main and mini-computers in the Selective Service/MEPCOM network.

# Mobilization Planning & Control
# A Model for the Selective Service System

by

Charles Robert Roll, Jr.
Policy and Management Planning Group
Science Applications, Inc.
&
B. Michael Berger
Analysis and Evaluation Division
Selective Service System

## INTRODUCTION

During a mobilization, the military services must be quickly augmented to provide the military capabilities required for the defense of the nation. Volunteers cannot be depended upon to meet all of these manpower requirements. Thus, the Director of Selective Service is charged with providing the Department of Defense with enough inductees on appropriate days to fill the balance of training base requirements.

In doing this, Selective Service is charged with maintaining a uniform national call so that the burden of conscription is distributed equitably among those eligible for induction. While a lottery system has been chosen to make the sending of induction notices "uniform," uniformity also requires that proper priority be given to individuals in processing, and that no significant processing backlogs develop. If such backlogs develop and induction does not proceed at a pace such that inductees with the same lottery numbers are processed at similar times, then the system may face a challenge to the process of induction call. This dimension of responsibility for the Selective Service System requires its involvement in all policy decisions related to mobilization planning and processing. Selective Service must be prepared to adjust its induction calls to account for the variations in AFEES processing capacity, must add in the number of "no shows," conscientious objectors or rejected inductees, and further consider the ability of the military training facilities to absorb the number of inductees which can be provided.

This paper describes two distinct models: one for planning, a second for control.* The planning model is used prior to a mobilization to assess policy alternatives and to test the sensitivity of the mobilization plan to respond to various uncertainties inherent in a mobilization. The control model is used for real-time adjustment of planned actions once a mobilization and draft have begun. While the basic processing procedures in each of these models will be the same, the control model differs from the planning model in that it focuses on the actual daily experiences of the operating induction and processing systems and thus requires a data monitoring capability.

## THE PLANNING MODEL

The planning model is the first stage in simulating mobilization flows. It keeps track of the number of individuals in all parts of the mobilization system from the issuance of draft notices or contacts with a recruiter through shipment of qualified individuals from AFEES. The planning model serves as the starting point for the control model described in the next section. The planning section contains many control-related aspects in order to maintain manpower shipments to training bases at the desired levels.

This section describes how the model is used to capture the anticipated flow and the options in the model for modifying the flow. A flow chart of the model is presented, then each part of the model is described.

## FRAMEWORK OF THE PLANNING MODEL

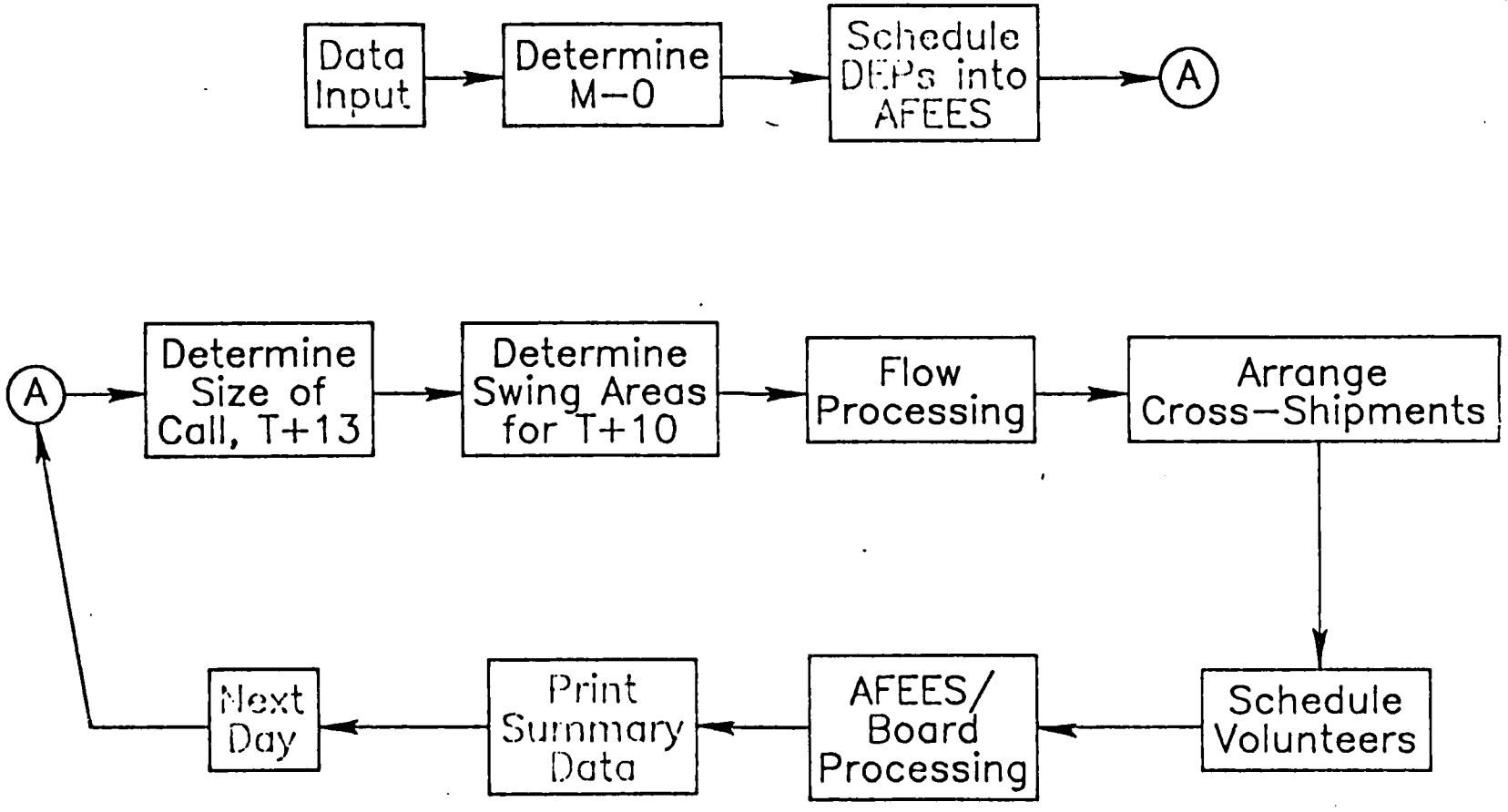
Figure 1 describes the processing flow of the planning model. Data is input both from a file on disk and through interactive query by the user. Some of the large data arrays are initialized at this point. Part of this information establishes the first day of mobilization. Data is transformed into week-day and day-of-the-year format by a perpetual calendar subroutine. Individuals in DEP (Delayed Entry Program) pool are scheduled into the AFEES during the first thirteen days of mobilization.

---

* The details of these models and the basis for this paper may be found in Bruce Wm. Bennett and Charles Robert Roll, Jr., Selective Service Mobilization Planning and Control, Science Applications Incorporated, July 1981.

Figure 1

# Flow Chart for the Planning Model



The remainder of the model works on a day-by-day basis. Each day, the size of the induction call required in the next 13 days is determined, then various administrative areas are set-up for the induction notices to be sent that day. Flow processing is then run for each AFEES. First, the responses to induction notices sent ten days previously are determined, indicating the number of induction candidates that have arrived at each AFEES or have taken some other action. The priority pools, those postponed to a specific data, are updated to include those who have been denied deferments or postponements, and those whose postponements have ended (such as students who have completed a school semester). Based upon these priority pools and the general pool of registrants, induction notices are "issued." The final flow process step takes any unplanned AFEES capacity for 10 days hence and schedules in CO (Conscientious Objectors) applicants for early processing before adjudication of their CO request.

After the basic flows have been determined for each AFEES, the model looks across all of the AFEES to determine those which are overloaded and cross-ships some of the men they are scheduled to process to an AFEES that has excess capacity. It is important to note that we define a group of individuals in each AFEES that are candidates to be "swung" to AFEES other than their home AFEES. Individuals in these "swing" areas are those who satisfy certain maximum distance criteria in terms of distance to their home AFEES and contiguous AFEES regions. Volunteers are then scheduled for AFEES exams 16 days hence or more. The AFEES/local draft board processing is performed on an AFEES-by-AFEES basis. The summary data for the model are printed, the day incremented, and the day-by-day analysis begun again. This process continues until 180 days of processing have been completed.

## THE CONTROL MODEL

Early in the planning of this research effort, it was anticipated that the planning and control models would be similar. Moreover, as work progressed, it became clear that the planning model was going to require significant "control" components in order to match appropriate training shipment requirements and, that the control model would, in any case, require almost exactly the same framework as the planning model. At the same time, it was recognized that in some sense two versions of the control model existed, though the differences between them were slight. In the first, the statistics being monitored (such as induction call response rates) are supplied as rates reflecting the use of the model in a planning or exercise evaluation mode. The second version of the model, to be used in an actual mobilization, uses actual counts of people as inputs to calculate actual rates.

Obviously, the only real difference between these two versions is the type of data used, and the extra step of converting numbers into rates.

Because the planning and control models are implemented in the same computer program, the description of the planning model above describes fairly completely the control model as well, and we simply note the major differences below.

In the planning model, it is assumed that the observed ("real") rates and other factors will be the same as the planning factors used in the model, and thus there is no need to monitor the rates as time goes on. In both versions of the control model, it is recognized from the beginning that observed rates and factors will vary, potentially by large amounts, from the factors initially used for planning purposes. Therefore, the observed values are monitored and used to modify the planning factors so that the model is as responsive as possible to the situation encountered. Further, it is recognized that the actual factors and rates can change over time, and thus the monitoring process cannot simply be done early on or on a cumulative basis; rather, continuous monitoring and updating of planning factors is required.

To this end, the control model establisnes a monitoring function for the most important rates and factors. These include the induction notice response rates and the rates associated with AFEES processing of COs, induction candidates, volunteers and those coming from DEP. These factors are the rates used in the planning and control process, and thus the essential values to know in order to schedule properly over time. It is also important to note the program is set up to handle the input observed rates as average values, and to sample around those values in a way that reflects the variations that should be expected in any set of observed data.

## APPLICATIONS

Space limitations preclude an examination of particular model simulations, and therefore we now turn to some more general applications of the concept of using essentially similar models for both planning and control purposes. The mobilization planning and control models were developed to manage the control of volunteers and inductees to the AFEES, however, there are other applications for the model within the training and personnel management communities.

For example, while the model deals with AFEES input, it can also be applied to AFEES "output" or the planning for and control of trainees into the training bases. Just as the model describes "swing" AFEES, it would be possible to apply these techniques to service schools or programs of instruction within service schools. As the predictor model identified the approach of training or course capacity limits, the model could

recommend alternatives (swing schools or courses) which would take maximum advantage of AFEES outputs by location. An example which considers only the flow of inductees might predict an overload at a particular Army training facility and recommend such actions as shifting personnel to another base offering the same training, shifting personnel into an alternative training program, or even reallocating inductees to another service which has excess training capacity. The model could take into account such factors as date of departure from the AFEES, the most economical shipment of trainees, and might even be expanded to consider results of mental examinations or other variable offering a prediction of success in training. Since the model works "in advance" of actual AFEES processing, some means of predicting the likely performance by personnel being processed at each AFEES might be necessary. Assuming some reasonable level of homogeniety among the population processing through a given AFEES, this form of prediction should not be beyond the capabilities of the system.

A second application involves management of the occupational structure itself. Using the "swing" concepts, one could envision prediction and control of training outputs using the concept of cross-leveling adapted for the mangement and control of trainees through AFEES and the training base. Again, applying the predictive capabilities of the system, one could input military skill requirements (either uni-service or multi-service) in terms of predicted need for replacements, new units, etc., and manage the entry of volunteers and inductees into training programs which would provide the desired outputs. As training seats in a given specialty reached "overload" levels, the model would "swing" trainees into best alternatives, either at the same or another training location. If the demand was for personnel with only "basic training" skills, the model might even provide opportunity for cross-leveling between military services. While this is certainly a radical departure from traditional uni-service military personnel management concepts, it is possible that a future mobilization will require optimum flexibility in personnel resource management. The mobilization planning and control models discussed in this paper offer some potential for development of innovative approaches to the military manpower problem.

# HUMAN APTITUDE ABILITY ASSESSMENT TECHNIQUES FOR SYSTEM DESIGNERS

AD P001368

By

Paul G. Rossmeissl, Stanley J. Kostyla and
James D. Baker
US Army Research Institute
Alexandria, Virginia

## ABSTRACT

Modern weapon systems are increasing in sophistication and man-machine interface complexity, while the manpower pool to operate and maintain these systems is decreasing in terms of both numbers of individuals and the aptitudes, abilities and skills those individuals bring into the Army. This situation leads to the necessity of considering human resources as a parameter of weapon system design, but such an effort is severely handicapped by a lack of efficient and reliable techniques that can be used by designers to estimate the human resource implications of their designs. The Army Research Institute (ARI) is currently pursuing a research program to develop a human aptitude/ability assessment technique for use during weapon system design. The basic approach is a taxonomy similar to that developed by Fleishman but computerized for greater efficiency and with heavier emphasis on cognitive factors. Research concerns within this project include: whether a branching or an exhaustive assessment technique is more effective, what is the appropriate level of analysis jobs or tasks, should the method of qualitative analysis be discrete or continuous, and what is the effect of having different types of users of the assessment procedures.



PREVIOUS PAGE IS BLANK

# HUMAN APTITUDE ABILITY ASSESSMENT TECHNIQUES FOR SYSTEM DESIGNERS

by

Paul G. Rossmeissl, Stanley J. Kostyla and
James D. Baker
US Army Research Institute
Alexandria, Virginia

## I. INTRODUCTION

The Army presently has a weapon system modernization program underway which is so large in scale that even the former head of the Army Force Modernization Coordination Office, General Lawrence (1979), has voiced concern about the Army's ability to absorb the impact of this "bow-wave of modernization." Unfortunately, this modernization program is taking place in a period of a severely constrained resource, namely, a people constraint. We know from census data (Bureau of the Census, 1977) there will be fewer 18-24 year olds between 1980 and the late 1990s, so the Army can expect persistent shortages of qualified recruits to operate and maintain these technologically advanced systems. As a consequence, the Chief of Staff of the Army, General Meyer, is quoted as saying that we have become a hollow Army, principally because of manpower shortages, (National Review 1980). While the sharp decline in the size of the future pool can hardly be questioned, there has been a good deal of controversy regarding a similarly sharp decline in the quality of both the recent military accessions, and conceivably, the future military manpower pool (Rimland and Lawson, 1980). Numerous examinations are now underway to assess this particular "supply-side deficit." For example, the opening panel of this 23rd Annual Conference of the Military Testing Association is entitled: "Profiling the Aptitudes of the Current Mobilization Population." In all, there is a growing concern among the manpower, personnel and training (MP&T) community, that when we finally know the true dimensions of the quality issue, and couple it with the known quantity shortfall, the bow-wave of modernization may in reality be a tsunami.

The problems of human resource supply and demand considerations during the force modernization program are further aggravated by the increasing sophistication and man-machine interface complexity. The long-term impact of this increase in complexity is not fully understood but preliminary studies (i.e., Kerwin and Blanchard, 1980; GAO Report 1981) suggest that increasing the sophistication of a weapon system often leads to an increase in the skills and abilities of the people required to operate and maintain that system. Since it appears that these highly skilled individuals may be in particularly short supply it is possible that sufficient quantities of individuals with the required aptitudes, abilities, and skill levels will not be available to effectively operate, maintain, and support the new and developing weapon systems (Kerwin and Blanchard, 1980).

Anticipating what the world will be on the supply-side, it becomes imperative that we work toward producing a detailed picture of the demand-side characteristics. In the past, skilled manpower for military systems were provided after hardware was delivered to the military unit. Personnel selection and training was accomplished as a form of reaction to the demands of the equipment. Today, because of the problems cited above, there is a dire need for human resource planning which will permit us to predict manpower

requirements during system development. Success in doing this will permit us to compare supply to demand and allow us to use human resource data as criteria in system design. Such data will provide us the means for making judgments about (a) the impact of design alternatives on our human resource pool, as well as (b) judgments about the constraints our human resource pool imposes on design alternatives (Askren, 1976). Let us take a moment here to briefly sketch-in how this might be done and to highlight a critical problem which must be overcome if we are to achieve this goal.

System design begins with a statement of purposes for the system; one or more "missions" the system is expected to perform. The purposes set the stage for the derivation of what the system's characteristics will be, i.e., mission profiles. Following the determination of system requirements and mission profiles, a functional analysis is undertaken which attempts to allocate functions between men and machines (see Figure 1). The traditional method for allocating functions between men and machines is to consider the relative superiority of the machine (e.g., microsecond response times; precise performance on boring, repetitive operations, etc.) or the human component (e.g., handling unanticipated occurrence; ability to reason inductively, etc.) in performing a particular function and assign that function accordingly.

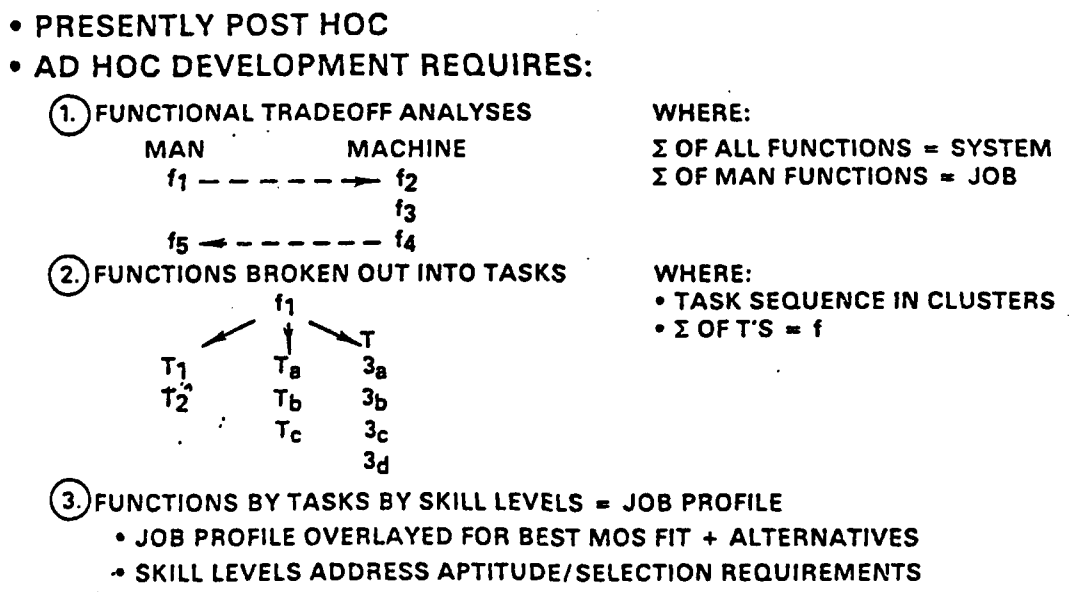



**Figure 1: Graphic illustration of how functional allocation/tradeoff analysis data leads to task analysis which, in turn, provides a basis, early in system design for MOS best fit and/or skill level determination and/or aptitude/selection requirements.**

Given the specification of the functions to be performed by the man in the system, it is now possible to break down the functions into component tasks. The classic definition of task is that of Miller (1953): "A group of discriminations, decisions and effector activities related to each other by temporal proximity, immediate purpose and a common man-machine output." The elements of a task are, therefore, the stimulus to the operator, which triggers performance

of the task, the required response to that stimulus (i.e., the performance criterion), a procedure for performing the response (which includes the equipment to be utilized for performing the task), and a goal or purpose (mission element) that organize the whole.

Now it merits comment that even though the designer routinely goes through a functional allocation/task analysis procedure in the front-end design process, these data typically are not used (although they could be) as detailed input for early MP&T considerations. Data, usually at the functional level, may be considered late in the life cycle of the system (Milestone II and beyond), to provide a reasonable estimate of the quantity of people the system will require, but these data contribute little to the quality determinations. Usually a post hoc determination is made in terms of military occupational speciality (MOS) requirements (quasi-quality determination) with only a "guestimate" of skill requirements (e.g., in terms of grade E-4, E-5, etc.). Numbers of people, by MOS and grade level, are thus selected in a somewhat "artistic" post hoc fashion. What is needed is a refined technique for making ad hoc determinations. This overall relationship is briefly summarized in Figure 1.

One way to achieve this goal is to use the early functional/task analysis data to provide a "job profile." This job-profile could then be overlayed on multiple MOS profiles to provide best, and alternate fits. But to the extent that the new job does not overlay precisely on an extant MOS, we are faced with a crucial problem....in the words of Meister (1976): "How does one derive from task characteristics, guidelines for operator selection and training and for prediction of operational performance?" He goes on to say (p. 101): "Despite written guides, the derivation of selection and training requirements is still largely an intuitive process."

What follows is a description of our attempt to resolve the problem of translating functional/task analysis data into behavioral components, such as aptitudes and skills, and to remove this process from the intuitive realm of art and to attempt to transform it into a behavioral science.

## II. BASIC APPROACH

As part of its investigation into the broad range of manpower, personnel, and training (MP&T) issues in weapon system acquisition, the Army Research Institute (ARI) is currently pursuing a research program to develop a technique that can be used by system designers to estimate the human resource implications of their designs. This effort entails the development of a computer-based assessment procedure to aid in identifying and quantifying estimates of the human aptitudes/abilities of projected tasks implied by system design concepts.

One of the first problems encountered by this project was the realization that there is no uniformly applied term referring to the human attributes under investigation. This problem is illustrated in Figure 2. Not only do different researchers, in this case Dunette (1976) and Fleishman (1975), use different terms to describe similar or equivalent concepts, but similar or identical terms are often used to describe distinct concepts. Thus, while Dunette uses the term ability to refer to a fairly specific cognitive trait, Fleishman uses the same term to refer to a general trait that can be either cognitive or physical.

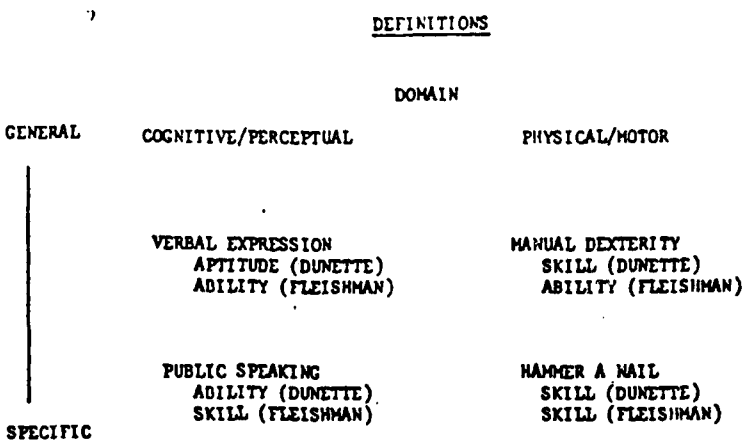



**Figure 2: Definitional discrepancies within the realm of human resource requirements.**

To try to minimize this confusion, it was decided to use the term aptitude to refer to the human resource traits of interest and then systematically define what is meant by the term. Aptitude will be used to refer to a general characteristic of an individual that affects his or her performance on a task or set of tasks. Aptitudes are assumed to be the result of a multitude of factors and therefore, are enduring traits that are difficult or impossible to alter through cost-effective training. Consequently, aptitude requirements should be of crucial consideration during weapon system development since any discrepancy in aptitude between the manpower required by a system and the personnel available to operate and maintain that system will be very difficult to overcome.

The basis or starting point of the current approach to the analysis of aptitude requirements is the extensive research of Fleishman (1972, 1975) in the identification of basic human aptitudes or abilities and their relationship to performance on a wide range of tasks. A particular advantage of using the Fleishman approach is that his basic procedure has been extended (Mallamad, Levine and Fleishman, 1980) to utilize binary decision-flow diagrams to assist in identifying the aptitude requirements of jobs and tasks. These decision flow-diagrams are structured in a binary (yes, no) format to reduce the information processing and decision-making demands on the analyst. The diagrams (see Figure 3) contain information relating to critical characteristics of an aptitude that suggest either its presence or absence and help differentiate that aptitude from similar aptitudes. A yes/no decision is required at each node in the decision flow structure. To aid in these decisions a list of task examples which are relevant to the aptitude in question is provided at each decision point. The diagrams function to determine the presence or absence of

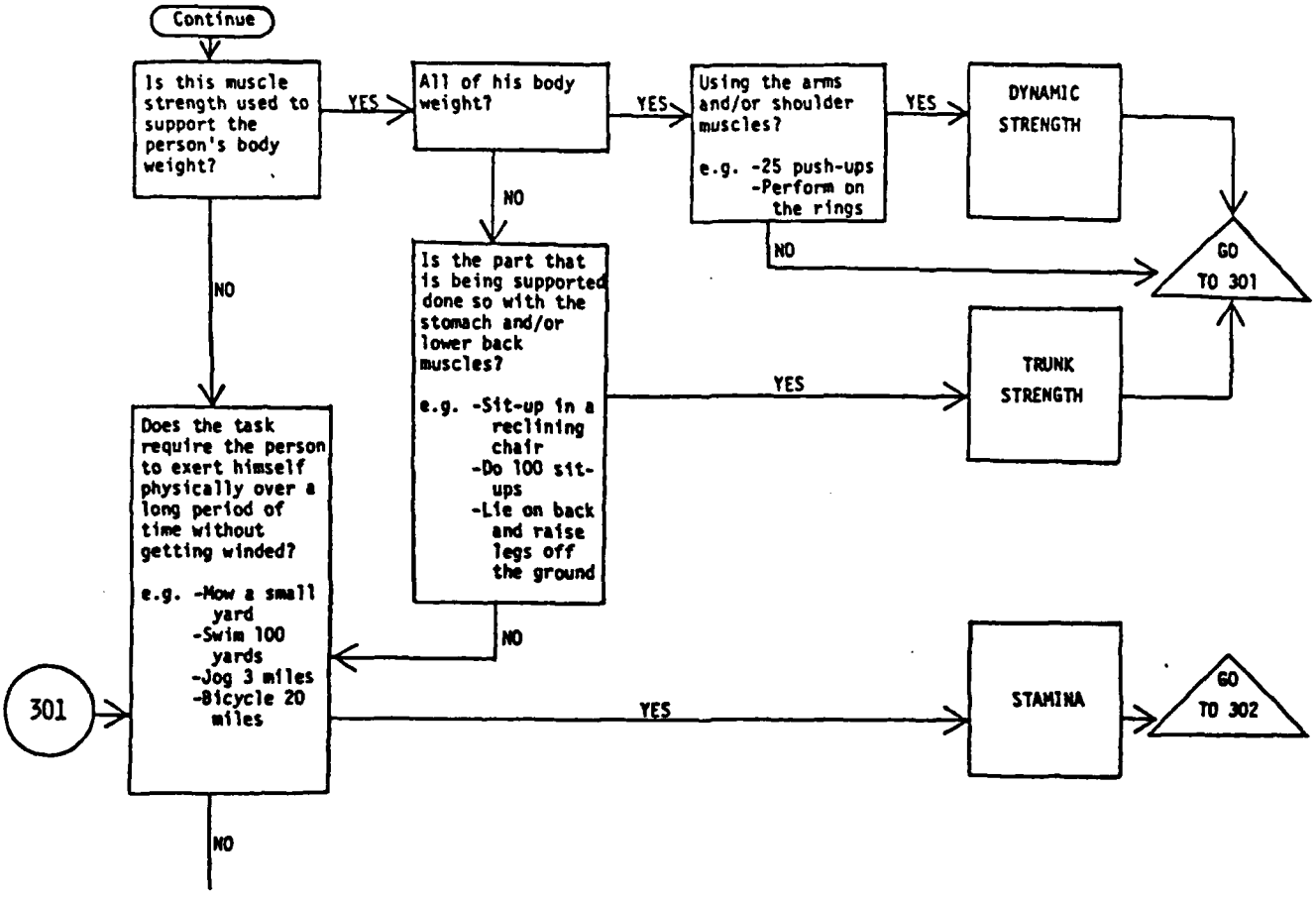



**Figure 3: An example of a portion of the binary decision flow diagrams (from Mallamad, Levine, and Fleishman, 1980).**

an aptitude judged to be necessary for task or job performance. Each diagram can then be further supplemented by a rating scale (see Figure 4) which may be used to quantify the relative level of a particular aptitude required to perform a given job or task.

The present developmental effort is designed to build and expand on the results of Fleishman and others in developing an aptitude oriented taxonomy that can relate tasks to their aptitude requirements (Fleishman, 1975; Mallamad et al, 1980; Siegel, Federman and Welsand, 1980) in a procedure that can easily be used by weapon system designers.

## III. COMPUTERIZED APTITUDE ASSESSMENT

The first phase of this project, currently being developed in conjunction with McFann Gray and Associates, will result in a research tool based on Fleishman's structured procedure mentioned above but computerized for greater accuracy and efficiency. This system will utilize a standard portable CRT display and off-the-shelf microcomputer components. The disc-based software will be in modular format in both its initial and upgrade versions. The software will consist of three basic elements: a binary decision flow skeletal

DELAY TOLERANCE

This scale is a measure of how much delay work performance can be tolerated between the time the soldier becomes aware that the work must be performed and the time he must begin doing it. Must the soldier begin immediately, or does he have time to consult a manual, seek guidance, or even be taught to do it? The work is to be rated on a scale from 1 (Very Long Delay Tolerance) to 7 (Very Short Delay Tolerance) with intermediate levels defined as follows:

How much delay before performing the job is acceptable?

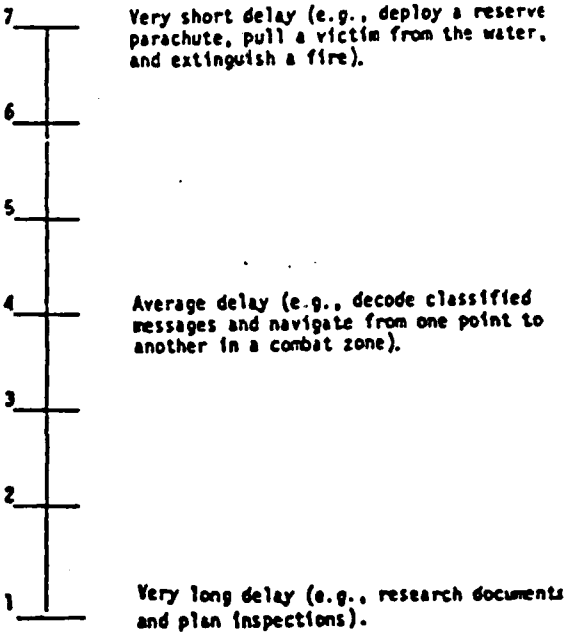



**Figure 4: An example of an aptitude rating scale (from Fleishman, 1975).**

structure, the capability for a variety of types of rating scales and task examples, and data aggregation, processing, reduction and analysis routines.

The immediate goal of the project is to refine this aptitude assessment methodology and to adapt its use to an Army environment. Initially this will entail the redefinition of the task exemplars and rating scale anchors to render them more appropriate to Army tasks. Other developmental considerations that will be addressed are outlined in Figure 5 adapted from Siegel et al (1980).

The research issues during the initial phase of development will be concerned primarily with structural and procedural variables such as rating scale format, rater variability, and task and job descriptions. The development, refinement and adaptation of the aptitude assessment technique will gradually shape the instrument to its dual function as a computer-based decision aid for system designers and a research tool for further analysis of the human aptitude/ability requirements of Army jobs and tasks.

Reliability--the scheme should be amenable to psychometrically reliable data acquisition methods.

Validity--the sheme should be based on acceptable constructs relevant to Army job content, and seem reasonable to the Army users.

Practicality--the scheme should be relatively simple to apply and interpret and should not place undue time requirements on operational personnel.

Scalability--the technique should allow for the assignment of a magnitude value to the estimate of aptitude requirements.

Understandability--the scheme must be readily apparent and comprehensible to Army users.

Combatability--the scheme should be fully compatable with the Army task structure.

Comprehensiveness, generality, and flexibility--the scheme should be applicable to the full range of tasks involved in Army jobs.

Cost-effective--the taxonomy should have characteristics that permit it to be embedded within a scheme that is relatively inexpensive to employ.

Figure 5: Developmental considerations.

Eventually this project should result in a reliable and efficient technique for use during the early stages of weapon system development as an aid to the system designer and developer in assessing the aptitude requirements associated with the operation, maintenance, and support tasks implied by new equipment design. Successful adaption and use of the assessment aid will help to identify potential areas of excessive demand on human capability and performance. Also, the aid will be useful in the development of aptitude profiles similar to the MOS task derived classification scheme.

## IV. ADDITIONAL INPUTS

The long range value and utility of the project briefly described in this paper will be enhanced by the results of a number of related research efforts currently under way. Among these is the renewed interest in the investigation of the relationship between human aptitude and human performance (Christal, 1980; Imhoff and Levine, 1981). While there has been extensive effort and some success in attempting to relate pencil and paper aptitude measures to training

and ultimately to field performance, there has been very little effort devoted to the examination of aptitudes, training, and performance relationships , especially in the area of cognition. A greater understanding of these relationships will contribute greatly to the further development of task taxonomies for use in interpretation and prediction of human performance (Fleishman, 1975). Collateral efforts in the development of improved techniques for the translation of system design specifications to functional requirements, and operator and maintainer tasks will contribute to our understanding of the impact of material design characteristics on the human resources in terms of the aptitudes and skills required of those tasks.

## V. References


Army Science Board Ad Hoc Study Group on Human Issues: Office, Assistant Secretary of the Army, Washington, DC, March 1980.

Askren, W.B. Human Resources as Engineering Design Criteria. AFHRL-TR-76-1, U.S. Air Force Human Resources Laboratory, Wright-Patterson AFB, Dayton, Ohio, March 1976.

Bureau of the Census. Projections of the Population of the United States: 1977 to 2050. (Series P-25, No. 704) United States Department of Commerce, Washington, DC, July 1977.

Christal, R.E. The need for laboratory research to improve the state of the art in ability testing. Unpublished manuscript, US Air Force Human Resources Laboratory, Brooks AFB, Texas, 1981.

Eckstrand, G.A. Manpower factors in systems acquisition. Paper presented at the Aerospace Industries Association Symposium, Seattle, Wash., October 1980.

Ekstrand, G.A., Asken, W.B. and Snyder, M.T. Human resources engineering: A new challenge. Human Factors, 1967, 9, 517-520.

Fleishman, E.A. On the relation between abilities and human performance. American Psychologist, 1972, 27, 1017-1032.

Fleishman, E.A. Toward a taxonomy of human performance. American Psychologist, 1975, 30, 1127-1149.

Imhoff, D.L. and Levine, J.M. Perceptual Motor and Cognitive Performance Task Battery for Pilot Selection. AFHRL-TR-80-27, US Air Force Human Resources Laboratory, Brooks AFB, Texas, January 1981.

Kerwin, W. T. and Blanchard, G.S. Man/Machine Interface - A Growing Crisis. Army Top Problem Areas, Discussion Paper Number 2, Army Material Systems Analysis Activity, Aberdeen Proving Ground, Maryland, August 1980.

Lawrence, R.D. Force modernization; Big job, big rewards. Army. October 1979.

Mallamad, S.M., Levine, J.M. and Fleishman, E.A. Identifying ability requirements by decision flow diagrams. Human Factors, 1980, 22, 57-68.

Meister, D. Behavioral Foundations of System Development. John Wiley and Sons: New York, NY 1976.

Miller, R.B. A Method for Man-Machine Task Analysis. WADC-TR-53-137, Wright Air Development Center, Dayton, Ohio, June 1953.

National Review. Divergent views on the problems of the all volunteer armed forces with one point in common. March 6, 1981.

Rimland, B. and Larson, G.E. The Manpower Quality Decline: An Ecological Perspective. NPRDC-TN-81-4, US Navy Personnel Research and Development Center, San Diego, California, November 1980.

Siegel, A.I., Federman, P.J. and Welsand, E.H. Perceptual/Psychomotor Requirements Basic to Performance in 35 Air Force Specialties. AFHRL-TR-80-26, US Air Force Human Resources Laboratory, Brooks AFB, Texas, December 1980.

AD P001369

Rumsey, Michael G., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Thurs. P.M.)

## Race Influences on Peer Ratings in ROTC Training Platoons

This study examined the influences of rater and ratee race on peer ratings of 4604 white and 884 black cadets distributed between three regional Army ROTC training camps. Blacks and whites were found to each give consistently higher ratings to their own subgroup than to the other, a tendency which was exacerbated when the minority subgroup judgments were particularly discrepant from the platoon judgments. The possible applicability of the concept of "race-bounded" friendships to these findings is considered. The pattern of black-white differences on peer ratings paralleled the pattern of such differences on other Advanced Camp measures.



PREVIOUS PAGE IS BLANK

# RACE INFLUENCES ON PEER EVALUATIONS IN ARMY ROTC TRAINING PLATOONS

Michael G. Rumsey

US Army Research Institute
for the Behavioral and Social Sciences

Two recent literature reviews have presented a strong case for the utility of peer assessments. In their review, Lewin and Zwany (1976, p. 430) concluded: "In summary, peer evaluations are valid tools for predicting future success and are superior to all other measures available at the time of rating." Kane and Lawler (1978), reviewing a long list of studies reflecting positively on the validity of peer evaluations, suggested that the time for widespread operational use of this technique in work settings is at hand.

The growing enthusiasm for peer evaluations is understandable and, in a sense, overdue. Research on such evaluations has been conducted for many years, particularly in the military, with generally favorable results. Peer evaluation has been shown to have validity for predicting a variety of criteria, including future officer performance (Haggerty, 1963) and promotion (Downey, Medland, & Yates, 1976).

There are a number of reasons which may account for the favorable results associated with peer assessments. While other types of evaluators may be exposed to a limited sample of an individual's behavior, peers have the opportunity to observe performance in a wide variety of situations. Generally, peer evaluations also offer the opportunity to pool the observations of a substantial number of raters, an opportunity not inherent in other types of evaluations. This pooling has two fortuitous consequences: first, it operates to expand the behavioral base for the final rating and secondly, it reduces the impact of idiosyncratic rating tendencies associated with any given evaluator.

Kane and Lawler (1978, p. 555) suggested two reasons why, despite the empirical support for peer evaluation, its use in operational evaluation systems has historically been rather limited. One is the apparent confusion between peer assessment and sociometry, a procedure emphasizing personal preferences rather than evaluations of performance. The other is "the failure to recognize the need for its use." Were these the only bases for reluctance by managers to use a peer evaluation system, we might indeed expect that such reluctance would soon give way. However, I would suggest that there are a number of additional reasons, many of them associated with factors which could potentially compromise the accuracy of peer assessments. Peers typically have minimal training and experience as evaluators and may well have a personal involvement in the outcome of the evaluation process. Thus, they may be more susceptible than other evaluators to the influences of assessee characteristics which are fundamentally irrelevant to the evaluator's task. A companion paper (Rumsey, 1981) has examined the influence of one such characteristic, assessee gender; the present paper examines the influence of another, assessee race.

A number of studies (Cox & Krumboltz, 1958; DeJung & Kaplan, 1962; Mohr & Reidy, 1976) focusing on peer evaluations in predominantly white military units have provided evidence of same-race favoritism in the expression of these evaluations. Blacks tended to demonstrate a higher level of such favoritism than whites, but this finding may well be attributed to the blacks' minority status within the groups studied. DeJung and Kaplan (1962, p. 373) suggested as much, noting that the minority subgroup member, in rating others within this subgroup, might well be rating his or her "closest buddies." The majority member, in rating others in the majority subgroup, might also be rating his or her "closest buddies" but would be rating almost everyone else in the unit as well. Thus, the same-subgroup ratings given by majority members would not appear as consistently high as those given by minority members. This explanation assumes the existence of "race-bounded" friendships, a term used by Cox and Krumboltz (1958) as well as DeJung and Kaplan. We will return to a discussion of such friendships a little later.

In a study conducted by Schmidt and Johnson (1973) which examined groups composed of equal numbers of blacks and whites, no same-race favoritism was observed. Since these investigators exposed their subjects to human relations training, it was not possible to determine whether this training or the numerical racial equality was more responsible for the unbiased ratings in this study. However, another study conducted by Clore, Bray, Itkin and Murphy (1978) indicated that numerical equality may indeed be a significant factor. Here, after equal numbers of black and white children had attended summer camp together, both blacks and whites showed positive changes in their attitudes toward each other. The authors suggested that the numerical equality among black and white children, counselors and administrative staff contributed to an environment which eliminated status differences associated with race and thereby fostered the attitudinal changes which took place.

The present study examined race influences on peer nominations in the same environment as that studied by Mohr and Reidy (1976). As in this earlier study, Army ROTC cadets participated in a six week training camp, called Advanced Camp, and were organized into predominantly white platoons which typically contained a substantial minority of blacks. In the year between the Advanced Camp which was the focus of the previous study and the Advanced Camp which provided the data for the present study, however, the instructions for the peer nominations had been changed. In order to shift the focus of evaluation from personal feelings to performance, instructions were rewritten to emphasize "effectiveness" and "demonstrated contributions," whereas previously cadets were asked who they would be most and least willing to serve under. One purpose of the present study was to determine whether, given these new instructions, the previous finding of same-race favoritism would be replicated.

If such favoritism were found, this finding would activate a second purpose of this study: to examine the relationship between such favoritism and overall group disagreement. When judgments of a particular racial subgroup diverge from those of the overall group, they may do so because of same-subgroup preference or for a variety of other reasons. An exploration of whether same-subgroup preference was a major contributor to such disagreement was planned as a means of obtaining further understanding of racial influences on peer evaluations.

A final purpose of this study was to examine how peer evaluations received by blacks and whites compared with scores these subgroups received on other measures of Advanced Camp performance. Fortunately, a number of such scores were available. Although none could serve as a totally accurate representation of a cadet's performance, the combination of all measures provided a rough picture of such performance. Differences between the races obtained in peer ratings but not observed on other measures would be a possible indication that one of the subgroups was inappropriately disadvantaged by peer evaluations.

## METHOD

The present investigation involved the use of data collected in 1976 at two of the three regional ROTC Advanced Camps. Only a nominal number of blacks attended the third camp, so data from that camp were not used. Subjects were 4545 ROTC cadets, including 3690 whites and 855 blacks. Various measures of cadet performance were administered at each camp by designated officials. These included two cadet peer evaluations, for which cadets rated the top ten and bottom ten individuals in their platoon on leadership potential and team member performance. An index reflecting the agreement of both blacks and whites with the overall platoon rating was also calculated at one of the camps. Cadets were also given two scores on overall camp performance and one score on performance in a tactical exercise by designated evaluators. Scores on a performance test designed to measure the cadet's ability to apply military skills, an orienteering test and a physical fitness test were also obtained.

## RESULTS

Each of the questions examined in this study was explored through the use of multiple t-tests. As a precaution against the possibility of achieving a spurious finding of significance by this technique, the acceptable level of significance for each set of comparisons was reduced according to the number of t-tests performed. If the number of comparisons did not exceed 10, .005 was adopted as the appropriate level. If the number exceeded 10 but did not exceed 20, .0025 was the level used. It should also be noted that, although results are given for both leadership and team member peer ratings, the high correlation coefficient of .88 ($p < .001$) obtained between the two ratings across all 5598 cadets at the three Advanced Camps indicates that the two ratings are not really independent.

An examination of peer evaluation scores revealed that both blacks and whites favored their own subgroup. Mean scores given by blacks and whites to members of each of these two racial subgroups are shown in Table 1. These scores were examined in two types of comparisons. First, t-tests were used to compare each subgroup to subgroup rating with the score expected if both groups had performed at exactly the same level. The "expected" score was 2.00. Since 16 comparisons were involved, the significance level was set at .0025. All ratings were found to be significantly different from 2.00, with both blacks and whites consistently evaluating members of their own subgroup above this level and the other subgroup members below this level. For both the leadership and team member peer ratings, the highest subgroup to subgroup rating at each camp was that given by blacks to other blacks, followed by whites' ratings of other whites.

The second procedure used *t*-tests to compare ratings given within a particular racial subgroup with ratings given to that subgroup by the other. Eight comparisons were involved here, so the significance level was set at .005. At both camps, leadership and team member ratings given by whites to whites were significantly (*p*'s < .001) higher than ratings given by blacks to whites. Similarly, leadership and team member ratings given by blacks to blacks were significantly (*p*'s < .001) higher than ratings given by whites to blacks.

Given this evidence of same-subgroup favoritism by members of each race, the next set of analyses was directed at the question of how such favoritism was related to overall group disagreement. Within each platoon, an index of agreement between judgments of black raters and judgments of all raters was calculated on the basis of an accumulation of discrete comparisons between how individual blacks rated a particular cadet and how the entire platoon rated that cadet. On the basis of this index, platoons were classified as high or low in agreement. Differences between judgments rendered in high and low agreement platoons were analyzed by means of *t*-tests, with the results shown in Table 2. The significance level, based on the number of *t*-tests involved, was set at .005. Under the "Ratings to Blacks" section of the table, it can be seen that blacks assigned significantly higher same-subgroup ratings in the low agreement platoons than in the high agreement platoons. In the same section of Table 2, one finds that white ratings of blacks were just the reverse, being significantly higher in the high agreement platoons than in the low agreement platoons. A comparable pattern of divergence does not appear in the "Ratings to Whites" section of this table, where platoons are dichotomized according to agreement between white raters and all raters.

The peer rating results were then examined in the context of results on all Advanced Camp measures. Black and white scores on each measure, as shown in Table 3, were compared on the basis of *t*-tests. For six of these comparisons, where a preliminary test called into question the assumption of homogeneity of variance underlying the conventional *t*-test, a modified version of this test resulting in a more appropriate *z* statistic was used. The significance level for the total of 16 comparisons shown in this table was set at .0025. You will note a disparity between peer rating scores in Table 3 and those shown in Tables 1 and 2. The data in Table 3 are presented in terms of a score standardized to have an overall mean of 100, while Tables 1 and 2 present scores computed such that the overall mean is 2.00.

The *t*-test results indicated that black-white differences on peer ratings were not sharply divergent from differences between these subgroups on other Advanced Camp measures. If peer ratings are disregarded, whites received significantly higher scores than blacks on five out of six categories at Camp B and four of six at Camp A. If peer ratings are included, whites received significantly higher scores than blacks on seven of eight categories at Camp B and four of eight at Camp A.

## DISCUSSION

Let us now consider the implications of the findings obtained here. This study essentially replicated the Mohr and Reidy (1976) finding of same-race favoritism in peer nominations in an ROTC Advanced Camp environment and provided evidence that the earlier finding was not merely an artifact of the particular rating instructions used in that study.

The comparison of ratings in platoons where black judgments were closely in accord with consensus judgments with ratings in platoons where such agreement was minimal provided further information about the rating patterns of each subgroup. As black ratings drifted further from consensus, black ratings of blacks became more favorable and white ratings of blacks became less favorable.

In examining possible explanations for this finding, one finds the concept of "race-bounded friendships" suggested in earlier studies a reasonable place to begin. A substantial body of literature (see Byrne, 1971) supports the proposition that individuals are attracted to those perceived as similar to themselves. Race presumably operates for many as at least a clue concerning the other's similarity. Thus, there may well be an initial predisposition to prefer same-race members in establishing friendships. This predisposition may be enhanced when the two major racial subgroups are represented unequally in the overall group. Such inequality may inhibit interracial friendships by emphasizing the distinctive nature of each subgroup, heightening the salience of subgroup membership and minimizing naturally occurring interactions between majority and minority members.

Any tendency by members of either subgroup to erect racial boundaries in the process of friendship formation obviously has implications for members of the other subgroup, who are likely to perceive this tendency and reciprocate. Furthermore, there is a basis for expecting race-bounded friendships to perpetuate themselves. Wilder and Allen (1978) have found that subjects choosing membership in one of two subgroups tend to prefer information which enhances their similarity to the chosen subgroup and their dissimilarity to the other. Perceived similarity with another is likely to influence attributions concerning the other's performance (Banks, 1976) and ultimately the evaluation of that performance. Thus, members of a race-bounded friendship subgroup might well be expected to rate one another more positively than they rate members of the other subgroup, an expectation reinforced by results from a number of studies showing a positive relationship between one's friendship with a peer and one's evaluation of that peer (Hollander, 1956; Hollander & Webb, 1955; Waters & Waters, 1970). Thus, race-bounded friendships may indeed produce the type of polarization between majority and minority subgroup member judgments observed in the present study, although the viability of this explanation relative to other hypotheses remains to be tested.

Despite the observed racial influences on peer evaluations in the present study, the relative performances of blacks and whites on these measures did not appear to depart substantially from the relative performances of these subgroups on other Advanced Camp measures. While the imperfect nature of these additional measures was commented on earlier, the consistency of racial differences across measures does suggest that peer ratings did not operate to the particular disadvantage of either racial subgroup.

Nevertheless, the evidence of racial influences on peer judgments is a matter of serious concern. This finding does not contradict the clear evidence from other studies that peer assessments are valid predictors of future performance, but does present a problem that needs to be addressed if such assessments are to be used to achieve their maximum potential.

Clearly, more research is needed to enable us to fully understand the processes that impact upon black and white peer evaluations. However, it is possible even at this point to identify approaches which appear promising as means of reducing racial influences upon these evaluations. Landy and Farr (1980) have suggested that rating errors can be reduced if rating formats incorporating behavioral anchors are used, although the expected gains from this procedure are relatively modest. Rater training may also be beneficial. Such training might incorporate such topics as: a description and discussion of the rating format used, a discussion of the importance of behavior as a basis for evaluation, guidance on how to observe behavior, a discussion of racial stereotypes and their contribution to rating errors, and a discussion of how to avoid rating errors associated with selective attention and inaccurate attributions.

If race-bounded friendships are indeed primarily responsible for the results observed in this study, then approaches which fail to deal directly with such friendships are likely to have limited impact. Perhaps the most effective mechanism for reducing racial influences, where feasible, would be to modify the racial composition of the units in which ratings are given. A distribution of cadets such that, in those units in which blacks are represented, they are represented in equal numbers with white cadets, might well create an environment less conducive to the formation of race-bounded friendships than the one examined here. To the extent that racial boundaries in the development of friendship groups interfere with free interactions between blacks and whites, one might well consider such an environmental change beneficial quite apart from its anticipated positive effect on rating accuracy.

## REFERENCES


Banks, W. C. The effects of perceived similarity upon the use of reward and punishment. Journal of Experimental Social Psychology, 1976, 12, 131-138.

Byrne, D. The attraction paradigm. New York: Academic Press, 1971.

Clore, G. L., Bray, R. M., Itkin, S. M., & Murphy, P. Interracial attitudes and behavior at a summer camp. Journal of Personality and Social Psychology, 1978, 36, 107-116.

Cox, J. A., & Krumboltz, J. D. Racial bias in peer ratings of basic airmen. Sociometry, 1958, 21, 292-299.

DeJung, J. E., & Kaplan, H. Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. Journal of Applied Psychology, 1962, 46, 370-374.

Downey, R. G., Medland, F. F., & Yates, L. G. Evaluation of a peer rating system for predicting subsequent promotion of senior military officers. Journal of Applied Psychology, 1976, 61, 206-209.

Haggerty, H. R. Status report on research for the U. S. Military Academy (cadet leaders task). (Army Research Institute Technical Research Report 1133) Alexandria, Va.: Army Research Institute, 1963.

Hollander, E. P. The friendship factor in peer nominations. Personnel Psychology, 1956, 9, 435-447.

Hollander, E. P., & Webb, W. B. Leadership, followership, and friendship: An analysis of peer nominations. Journal of Abnormal and Social Psychology, 1955, 50, 163-167.

Kane, J. S., & Lawler, E. E. III. Methods of peer assessment. Psychological Bulletin, 1978, 85, 555-586.

Landy, F. J., & Farr, J. L. Performance rating. Psychological Bulletin, 1980, 87, 72-107.

Lewin, A. Y., & Zwany, A. Peer nominations: A model, literature critique and a paradigm for research. Personnel Psychology, 1976, 29, 423-447.

Mohr, E. S., & Reidy, R. F. Racial bias in peer ratings at ROTC Advanced Summer Camp, Fort Bragg, 1975. (Army Research Institute Research Memorandum 76-22) Alexandria, Va.: Army Research Institute, 1976.

Rumsey, M. G. Gender influences on peer ratings in ROTC training platoons. Paper presented at Annual Meeting of the American Psychological Association, Los Angeles, August, 1981.

Schmidt, F. L., & Johnson, R. H. Effect of race on peer ratings in an industrial situation. Journal of Applied Psychology, 1971, 57, 237-241.

Waters, L. K., & Waters, C. W. Peer nominations as predictors of short-term sales performance. Journal of Applied Psychology, 1970, 54, 42-44.

Wilder, D. A., & Allen, V. L. Group membership and preference for information about others. Personality and Social Psychology Bulletin, 1978, 4, 106-110.

ACKNOWLEDGEMENT

The creative and knowledgeable statistical analysis and consulting assistance provided by Mrs. Gail Rowan is gratefully acknowledged by the author.

TABLE 1

Peer Rating Scores by and for Blacks and Whites

| | | Ratings Given by Blacks | | Ratings Given by Whites | | |
|---|---|---|---|---|---|---|
| Rated Group | N | M* | SD | M* | SD | t** |
| **Camp A: Leadership Rating** | | | | | | |
| Blacks | 444 | 2.39 | .36 | 1.87 | .37 | 20.08 |
| Whites | 1740 | 1.91 | .33 | 2.04 | .32 | 11.18 |
| **Camp A: Team Member Rating** | | | | | | |
| Blacks | 444 | 2.42 | .39 | 1.88 | .39 | 20.88 |
| Whites | 1740 | 1.91 | .37 | 2.03 | .38 | 10.08 |
| **Camp B: Leadership Rating** | | | | | | |
| Blacks | 411 | 2.48 | .39 | 1.82 | .38 | 24.15 |
| Whites | 1950 | 1.92 | .38 | 2.04 | .40 | 9.92 |
| **Camp B: Team Member Rating** | | | | | | |
| Blacks | 411 | 2.39 | .38 | 1.79 | .36 | 23.00 |
| Whites | 1950 | 1.93 | .36 | 2.05 | .33 | 10.45 |

*All $t$ values comparing means with 2.00 significant at $p < .0025$
**All $t$ values shown significant at $p < .001$

TABLE 2

Ratings Received by Whites and Blacks
in Platoons with High and Low Agreement Indices

| | High Agreement | | | Low Agreement | | | |
|---|---|---|---|---|---|---|---|
| Rating Type | N | M | SD | N | M | SD | t |
| **Ratings to Whites** | | | | | | | |
| Black Ldr | 23 | 1.93 | .03 | 22 | 1.89 | .04 | 3.75* |
| White Ldr | 23 | 2.04 | .04 | 22 | 2.03 | .03 | 1.00 |
| Black TM | 21 | 1.92 | .04 | 23 | 1.89 | .05 | 2.62 |
| White TM | 21 | 2.03 | .04 | 23 | 2.04 | .04 | .66 |
| **Ratings to Blacks** | | | | | | | |
| Black Ldr | 24 | 2.31 | .13 | 24 | 2.50 | .16 | 4.49* |
| White Ldr | 24 | 1.93 | .13 | 24 | 1.79 | .10 | 4.18* |
| Black TM | 24 | 2.35 | .26 | 24 | 2.53 | .20 | 3.27* |
| White TM | 24 | 1.94 | .13 | 24 | 1.81 | .09 | 3.96* |

*$p < .005$

TABLE 3

Scores Received by Blacks and Whites
on all Advanced Camp Measures

| Variables | Whites N | Whites M | Whites SD | Blacks N | Blacks M | Blacks SD | t |
|---|---|---|---|---|---|---|---|
| | | | Camp A | | | | |
| SOAT | 1734 | 100.37 | 5.83 | 443 | 98.85 | 6.43 | 4.47*[a] |
| PNAT | 1740 | 100.34 | 19.40 | 444 | 99.05 | 20.64 | 1.24 |
| POAT | 1740 | 100.84 | 19.54 | 444 | 97.39 | 19.78 | 3.32* |
| PEER (TM) | 1740 | 100.59 | 19.02 | 444 | 97.82 | 21.57 | 2.47[a] |
| PEER (LDR) | 1740 | 100.36 | 19.51 | 444 | 99.04 | 19.94 | 1.21 |
| PT | 1734 | 407.39 | 39.50 | 444 | 424.06 | 37.75 | 8.01* |
| ORIENT | 1740 | 101.99 | 15.40 | 444 | 92.92 | 16.72 | 10.93* |
| MIL STAKES | 1738 | 102.91 | 18.31 | 444 | 88.53 | 20.38 | 13.57*[a] |
| | | | Camp B | | | | |
| SOAT | 1934 | 100.36 | 5.83 | 408 | 99.15 | 6.38 | 3.56*[a] |
| PNAT | 1950 | 101.25 | 19.39 | 411 | 95.83 | 20.02 | 5.11* |
| POAT | 1950 | 101.25 | 19.56 | 411 | 95.55 | 18.73 | 5.43* |
| PEER (TM) | 1950 | 101.37 | 19.25 | 411 | 93.16 | 19.53 | 7.82* |
| PEER (LDR) | 1950 | 100.82 | 19.59 | 411 | 96.08 | 18.67 | 4.51* |
| PT | 1948 | 429.02 | 41.06 | 411 | 445.31 | 32.84 | 8.71*[a] |
| ORIENT | 1950 | 101.79 | 15.34 | 411 | 93.01 | 15.47 | 10.58* |
| MIL STAKES | 1950 | 101.81 | 18.58 | 411 | 91.30 | 21.32 | 9.30*[a] |

*$p < .0025$
[a] $z$ statistic

Brief Description of Variables Presented in Table 3

SOAT: Rating of cadet performance in a one-day tactical exercise.

PNAT: Rating of cadet overall camp performance by a non-commissioned officer.

POAT: Rating of cadet overall camp performance by an officer

PEER (TM): Peer rating on team member performance.

PEER (LDR): Peer rating on leadership potential.

PT: Performance on a physical fitness test. Here, total raw score, rather than a standardized score with a mean of 100, is used.

ORIENT: Score on a timed freestyle orienteering performance test.

MIL STAKES: Score on a performance test designed to measure the cadet's ability to apply military skills.

AD P001370

# THE NAVY PERSONNEL ACCESSIONING SYSTEM

by

W. A. Sands

Navy Personnel Research and Development Center
San Diego, California 92152

## ABSTRACT

The purpose of the Navy Personnel Accessioning System (NPAS) project is to develop a computer-based system which integrates and supports four functions at the recruiting station level: (1) individualized testing, (2) vocational guidance, (3) assignment prediction, and (4) management support. The individualized testing function involves the administration, scoring, and interpretation of an adaptive aptitude screening test. Vocational guidance includes a discussion of career planning; administration, scoring, and interpretation of an interest inventory; interpretation of ASVAB results; and an overview of the Navy world of work. The assignment prediction function indicates the extent to which an individual applicant's personal characteristics (aptitudes and preferences) match the requirements of the entry-level Navy ratings. The last function, management support, involves: (1) data entry, storage, and retrieval of information on recruit applicants; (2) generation of forms employed in the enlistment process (e.g., the DD-1966); and, (3) generation of management reports.

## INTRODUCTION

The population of males between the ages of 17 and 21 is declining, and forecasts indicate that this trend will continue into the early 1990's. At the same time, technological advances in equipment have increased the need for high-quality personnel in the military services. Competition for high school graduates within this target population will become increasingly fierce. Not only will the Navy be competing with the other military services, but also with colleges and universities which, faced with declining enrollments, will be intensely recruiting this population group.

The Navy Personnel Accessioning System (NPAS) has focused on a target population consisting of male, nonprior service, enlisted applicants who seem promising to the Navy recruiter. The purpose of the NPAS project has been to develop a computer-based personnel accessioning system to integrate and support four major functions at the recruiting station level: (1) individualized testing, (2) vocational guidance, (3) assignment prediction, and (4) management support.



PREVIOUS PAGE IS BLANK

## FUNCTIONS

### Individualized Testing

The first function, individualized testing, involves the computer-based administration of a short ability test battery. The component tests will be adaptive, as distinguished from conventional tests wherein everyone takes the same test items regardless of ability level. Conventional testing is very inefficient, as items of low difficulty are wasted on high-ability persons, and items with high difficulty levels are wasted on low-ability persons. The process of adapting a test to the individual examinee is called Computerized Adaptive Testing (CAT).

The adaptive test developed under the NPAS project is called the Computerized Adaptive Screening Test (CAST). The CAST is designed to replace the Enlistment Screening Test (EST) for predicting an applicant's performance on the Armed Services Vocational Aptitude Battery (ASVAB). More specifically, the CAST is designed to predict an applicant's Armed Forces Qualification Test (AFQT) score on the ASVAB.

Three item banks were developed for the CAST: (1) arithmetic reasoning, (2) word knowledge, and (3) paragraph comprehension. These areas were chosen as being most promising for the prediction of the AFQT score.

In comparison to the EST, the CAST should exhibit a number of benefits, including: increased measurement precision, reduced testing time, improved test security, reduced clerical errors, and reduced costs.

### Vocational Guidance

The second function supported by the NPAS system is vocational guidance, an area largely ignored by the current accessioning system. The Navy, like the other military services, includes a large number of diverse jobs. Many, if not most, of these jobs are unfamiliar to the typical applicant. To assist an applicant in career planning, his interests will be measured using the Vocational Interest Career Examination (VOICE)[1]. The 245 items of VOICE (Form C) will be administered by computer. The administration will be conventional, not adaptive; i.e., each applicant will take the same items in the same sequence. Scores will be determined for 18 basic interest scales and will be presented and interpreted to the applicant using percentile scores and bar graphs. These will be shown on the Video Display Terminal (VDT), with results available to the applicant on hardcopy printout.

---

[1]The Vocational Interest Career Examination was developed by the Educational Testing Service under contract to the Air Force Human Resources Laboratory.

Computer-based vocational guidance offers a number of important advantages over current procedures, including: (1) accurate, consistent, and current information, (2) rapid access to that information, (3) self-paced progress by the applicant and (4) independence. This last advantage, independence from the recruiter's time schedule, is particularly important. While the applicant is interacting with the computer system, the recruiter can attend to other duties.

Assignment Prediction

Prediction of an applicant's assignment options represents the third major function of the NPAS system. The Personalized Recruiting for Immediate and Delayed Enlistment (PRIDE) is the Navy's person-job matching system. Until recently, this system resembled an airline reservation system. Applicants were treated dichotomously, as either "eligible" or "ineligible" for each entry-level Navy rating. The quality of a person-job match was ignored. The Classification and Assignment within PRIDE (CLASP) model was developed to improve this situation. CLASP is an optimal person-job matching model which was recently incorporated into the PRIDE system and is currently operational at the Armed Forces Examining and Entrance Stations (AFEES).

The utility function employed in the CLASP model includes five components: (1) school success, (2) aptitude/complexity, (3) Navy need/preference, (4) minority-fill, and (5) fraction-fill (Kroeker, 1979). A score is determined for each of these five components. These scores are weighted and summed to provide a weighted composite payoff for matching an applicant with each entry-level Navy rating. Using a decision index procedure (Ward, 1958; 1959), each composite payoff is transformed into an optimality indicator. This optimality indicator represents the "goodness of fit," or quality of the person-job match. Rating options for which the applicant is eligible are rank-ordered from high to low by optimality indicator. Ratings with the highest optimality indicators are offered to applicants by the Navy classifier during the interview at an AFEES location.

Since the NPAS system is designed for use at the recruiting station level, the rating options and associated optimality indicators must be predicted. The Pre-CLASP model was designed to forecast the rating options which will be offered to the applicant during the subsequent classification interview at the AFEES. In addition to the information required by the CLASP model, Pre-CLASP requires the projected AFEES arrival date and the applicant's preferred shipping month. Output data from the Pre-CLASP model include a set of rating options, a set of optimality indicators, and a set of probabilities that the ratings will be open when the applicant arrives at the AFEES classification interview. The ratings are rank-ordered by optimality indicator. Those rating options with the highest optimality indicators are displayed on the VDT for the applicant. Presentation of this information in the recruiting station environment gives the applicant time for

serious consideration of vocational alternatives. This is in marked contrast to the short, pressured classification interview at the AFEES. Presumably, this information, along with the time necessary to absorb it, will enhance the quality of the person-job match. Specifically, it should decrease the problem of unmet expectations and the resultant premature attrition. A more detailed discussion of these three person-job matching functions has been presented previously (Sands, 1980).

Management Support

Whereas the previous three functions have been aimed at the applicant, this last function is designed for the recruiter. The NPAS system supports three recruiting station management capabilities: (1) data entry, storage, and retrieval; (2) forms generation; and, (3) reports generation.

Data entry and editing for this function will be accomplished by the recruiter using a keyboard with a typewriter layout and a VDT. To facilitate the efficient use of the NPAS system, software design and development has been guided by two considerations. First, the system should be "user-friendly." This means that effective use of the NPAS system requires no background training or knowledge of computers. The second design consideration involved making the software "menu-driven" wherever possible to decrease clerical time and errors. In soliciting information from the recruiter, the question is presented on the VDT. As in a multiple-choice question, the possible response alternatives are displayed directly beneath the question and are associated with single-digit numbers. The recruiter selects the most appropriate response alternative and enters the single-digit number associated with the alternative.

Storage of information on computer-readable media (e.g., floppy diskettes) will significantly reduce physical storage requirements while, at the same time, greatly increasing the speed of information retrieval.

Forms generation represents the second management capability. The process of enlisting an applicant into the U.S. Navy requires a tremendous amount of paperwork. The Application for Enlistment - Armed Forces of the United States (DD Form 1966) is a noteworthy example. This form consists of eight pages and requires a considerable amount of clerical effort and time. For example, the current manual procedures require the recruiter to type an applicant's name and social security account number on the top of each of the eight pages of the DD Form 1966 and to enter the same information on various other enlistment forms. Automation of the enlistment kit forms will eliminate the redundant entry of data. As indicated above, the menu-driven software will lessen the recruiter's clerical burden still further.

Reports generation is the third management support capability. Operation of a Navy recruiting station entails a substantial

number of management topics ranging from applicant flows to reports on vehicle usage. The work involved in producing many of these periodic reports is largely clerical "bean-counting." The computer can produce these management reports far more rapidly and accurately than the recruiter, while freeing the recruiter's time for other activities. Moreover, since the data are stored on computer media, they are available for aggregation and summary reports at higher levels in the Navy Recruiting Command.

## SUMMARY

In summary, the NPAS system is a computer-based personnel accessioning system designed for use in the 1980's timeframe. The system integrates and supports four functions at the recruiting station level: (1) individualized testing, (2) vocational guidance, (3) assignment prediction, and (4) management support.

The NPAS system offers a number of significant benefits including more accurate and extensive screening and testing, enhanced vocational guidance, improved classification procedures, and extensive management support for the Navy recruiter.

## REFERENCES

Kroeker, L. Policy Specifying, Judgement Analysis, and Navy Personnel Assignment Procedures. San Diego, California: Proceedings of the 21st Annual Conference of the Military Testing Association. October 1979, 592-598.

Sands, W.A. The Automated Guidance for Enlisted Navy Applicants (AGENA) System. Toronto, Canada: Proceedings of the 22nd Annual Conference of the Military Testing Association. Vol. II, October, 1980.

Ward, J.H. Jr. The Counseling Assignment Problem. Psychometrika, 1958, 23, 6-17.

Ward, J.H. Jr. Use of a Decision Index in Assigning Air Force Personnel. Lackland Air Force Base, Texas: Personnel Laboratory, Wright Air Development Center, Air Research and Development Command. Technical Note WADC-TN-59-38, April 1959.

AD P001371

# HUMAN FACTORS EVALUATION OF DIVISION AIR DEFENSE GUN SYSTEMS

Gary G. Sarli and Richard J. Carter
Research Psychologists
U.S. Army Research Institute for
the Behavioral and Social Sciences
Fort Bliss Field Unit
Fort Bliss, Texas 79916

A human factors evaluation was conducted upon DIVAD Gun prototypes during Operational Test II. It was physically impossible to observe the crew-members during operations; therefore, data were gathered from each of 32 enlisted crew-members by means of five questionnaires drawn from a master set of 506 items. Despite some difficulties experienced by crew-members in responding to the questionnaires, the required data were obtained and the results were submitted to the U.S. Army Operational Test and Evaluation Agency and the DIVAD Gun Source Selection Board.



PREVIOUS PAGE
IS BLANK

# HUMAN FACTORS EVALUATION OF DIVISION AIR DEFENSE GUN SYSTEMS

The U.S. Army is developing a new air defense gun system, DIVAD Gun, to replace the self-propelled, 20mm, M163 Vulcan System. The DIVAD Gun is designed to:

1. Provide air defense for divisional maneuver elements.

2. Provide air defense for selected ciritcal assets, choke points, and convoys in the division area.

3. Deter easy access to rear areas by low altitude threats.

4. Provide effective ground fire against lightly armored vehicles and enemy personnel.

General Dynamics and Ford Aerospace have each built two prototype DIVAD Gun systems. Both types are mounted on modified M48A5 tank chassis and incorporate government furnished communications equipment, secondary armament, and nuclear, biological, and chemical (NBC) equipment. The XM246, developed by General Dynamics Corporation, uses twin 35mm Oerlikon KDA guns and a fire control based on the US Navy Phalanx, close-in weapon system, gun system. The XM247, designed by Ford Aerospace and Communications Corporation, utilizes a pair of Bofors 40mm guns. The radar-directed fire control is based on the Westinghouse F-16 aircraft radar. The turret structure for both systems contains the armor, gun mount, magazine, crew compartment, and operator controls and displays (Vereb, 1980).

An operational test was conducted on four DIVAD Guns (two of each prototype) by the U.S. Army Operational Test and Evaluation Agency (OTEA). It was performed during the time interval July to November 1980 at North McGregor Range, New Mexico and was divided into four phases: Detection/tracking, aerial live fire, ground live fire, and manuever (Houser & Donovan, 1980).

The test evaluated:

1. The operational effectiveness of the two prototypes in the areas of fire power, fire control, and total system integration.

2. The mobility and survivability of the systems under operational conditions.

3. The reliability, availability, and maintainability characteristics of the gun systems under operational conditions.

4. The adequacy of the proposed training program and personnel selection criteria.

5. The adequacy of proposed doctrine, tactics and organization for employment of the DIVAD Gun candidates under operational conditions.

6. The susceptibility/vulnerability of the prototype systems in an Electronic Countermeasure (ECM) environment.

The test was run under: (1) two system modes: Moving and stationary, (2) two visibility conditions: Day and night, (3) four levels of ECM: Noise, deception, chaff, and benign, and (4) three environments: Normal, NBC, and dust/smoke.

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) was requested by OTEA to conduct a human factors evaluation of the two systems during the operational test. The purposes of the evaluation were to determine whether the crews can perform all required tasks to accomplish the mission objectives of the DIVAD Gun and to identify man-machine interfaces negatively affecting task accomplishment.

## Method

### Subjects

Thirty-two male service members, E2 through E6, stationed at Ft Bliss, Texas participated in the operational test. They were from the 1st Battalion (Chaparral/Vulcan), 55th Air Defense Artillery. The service members were separated into eight crews: Two for each of the two Ford systems and two for each of the two General Dynamics systems. A crew consisted of a squad leader, senior gunner, driver, and ammunition handler. Prior to the start of the test, the gun system contractors gave each of their crew members New Equipment Training (NET) appropriate to his gun system. Shortly after the start of the test, one crew member was replaced. The replacement received on-site training.

### Apparatus

It had originally been planned to conduct the evaluation by observing the crew members via closed-circuit TV. However, permission to place TV cameras inside the vehicles was denied. It was then decided that written questionnaires would be a feasible method of collecting the data.

A master set of 506 items was developed, using the Questionnaire Construction Manual (Dyer, Matthews, Wright, & Yudowitch, 1976) designed by the ARI Ft Hood Field Unit as a guide. The master set consisted primarily of closed-ended questions (See Table 1). Most of these used 5-point, unipolar rating scales, although there were a few ranking items. The 5-point scales were preferred over 7 or 9-point scales for this use (Dyer, et. al., 1976, Chap VI-G, pg. 2). The 506 items, dealing with 22 subject areas (See Table 2), were used to construct five questionnaires.

### Procedure

The questionnaires were to be administered to the Ford and General Dynamics crew members in separate locations just after the crew members had participated in relevant exercises (for example, night firing).

Two types of problems were encountered. The first type of problem was caused by scheduling changes, which sometimes resulted in crew members receiving questions concerning events in which they had not yet participated. When this occurred, those questions had to be recycled into the next questionnaire.

The other type of problem concerned the questions themselves. Prior to assembling the questionnaires, all the questions had been edited by subject matter experts, both for accuracy and for reading level. However, some of the crew members still did not understand some of the questions. Also, some of the crew members, despite written and verbal instructions, reacted to the ranking items as if they had been checklists.

Following the administration of the first questionnaire, the vocabulary used in the remaining items in the master set was further simplified as required. Also, simplified versions of those items which had been misunderstood on the first questionnaire were recycled into later questionnaires. The ranking items were replaced by items asking how often the events in question had occurred.

It has been suggested (Oppenheim, 1966, pp. 85 & 86) that the order of the response alternatives on ranking items be varied so that the first alternative is neither always positive nor always negative. This advice was followed but abandoned after the crew members expressed confusion.

To further clairfy the answers to some of the questionnaire items, group discussions were held with the crew members.

## Results and Discussion

ARI succeeded in obtaining answers to all of the items in the questionnaire. These results will be published in a future paper.

Today's soldiers have a wide range of reading ability. Therefore, individual interviews are probably preferable to written questionnaires. Unfortunately, giving individual interviews is not always practical, and a certain amount of data will be lost in group interviews. If questionnaires are used and it is not possible to perform a pilot study, the questions should be edited by soldiers of the same Military Occupational Speciality, educational level, and rank as those in the target population. Also, discussion sessions should be planned with at least a sample of the target population, to clarify, and perhaps expand upon, their responses.

## References


Dyer, R. F., Matthews, J. J., Wright, C. E., & Yudowitch, K. L. Questionnaire construction manual. Ft Hood, TX: U. S. Army Research Institute for the Behavioral and Social Sciences, Ft Hood Field Unit, 1976.

Houser, B. J., & Donovan, J. Division air defense gun development/operational combined test plan (TDP-OT-582 3 WE-100-DIV-001). Falls Church, VA: U. S. Army Operational Test and Evaluation Agency, March 1980.

Oppenheim, A. N. Questionnaire design and attitude measurement. New York: Basic Books. Inc., 1966.

Vereb, T. A. The division gun program from the beginning to now. Air Defense Magazine, October-December, 1980, 26-29.

Table 1

SAMPLE QUESTIONS

012. How easy or hard is it to get out of the fire unit?

___VERY EASY
___EASY
___BORDERLINE
___HARD
___VERY HARD
___NO OPINION/DON'T KNOW

H5. Rate the quality of the view through the optical sight when DIVAD Gun is moving.

___EXCELLENT
___GOOD
___BORDERLINE
___POOR
___TERRIBLE
___NO OPINION/DON'T KNOW

024. Did your eyes get tired after watching the plasma display?

___NOT AT ALL TIRED
___A LITTLE TIRED
___TIRED
___QUITE TIRED
___EXTREMELY TIRED
___NO OPINION/DON'T KNOW

037. Which displays, gauges, dials, etc. (if any) are hard to read?

Table 2

The Number and Type of Question for Each Area

| Area Title | Question Types | | | | Total Number of Questions |
|---|---|---|---|---|---|
| | Closed-Ended | | | Open-Ended | |
| | 5-Point Scale | | Other | | |
| | Unipolar | Bipolar | | | |
| Detection | 31 | 0 | 1 | 1 | 33 |
| Identification | 23 | 0 | 1 | 5 | 29 |
| Achieving Lock | 19 | 0 | 0 | 2 | 21 |
| Reasons for Breaking Lock | 0 | 0 | 20 | 5 | 25 |
| Smoke + Blast Signatures | 8 | 0 | 0 | 0 | 8 |
| ECM (Jamming) | 11 | 0 | 1 | 1 | 13 |
| Collimation + Alignment | 6 | 0 | 0 | 2 | 8 |
| Moving DIVAD Gun | 9 | 0 | 1 | 1 | 11 |
| Ammunition Reload | 7 | 0 | 0 | 1 | 8 |
| Bite Diagnostics | 3 | 0 | 0 | 1 | 4 |
| Training | 20 | 10 | 10 | 0 | 40 |
| Ground Targets | 18 | 0 | 10 | 6 | 34 |
| Night Operations | 19 | 0 | 1 | 2 | 22 |
| Logistics | 20 | 0 | 0 | 3 | 23 |
| Human Factors | 52 | 5 | 5 | 10 | 72 |
| Self + Mutual Defense | 17 | 0 | 0 | 0 | 17 |
| NBC | 32 | 0 | 0 | 6 | 38 |
| Camouflage | 10 | 0 | 1 | 1 | 12 |
| March Order/Emplacement | 4 | 0 | 2 | 4 | 10 |
| Doctrine + Tactics | 25 | 0 | 0 | 17 | 42 |
| Organization | 7 | 0 | 1 | 8 | 16 |
| Summary | 18 | 0 | 1 | 1 | 20 |

AD P001372

# IDENTIFYING COMMON DUTIES AMONG NAVAL SKILLED TRADES

Amiel T. Sharon
Personnel Research Psychologist
United States Office of Personnel Management

A comprehensive occupational survey of 22 Naval civilian trades and crafts was undertaken to establish a data base for the development of examinations to select apprentices and to determine the job-relevance of apprentice training programs. The initial step in preparation for the survey was to identify a group of common duties by reviewing draft task inventories in each of the 22 occupations. The common duties and task inventories were incorporated into occupational survey questionnaires and reviewed by 79 subject matter experts. The questionnaires, which were completed by approximately 5000 skilled blue collar workers, also sought information about the workers' background, job hazards, relevance of classroom apprenticeship training to the job, muscular effort used on the job, tools and equipment used, and relative importance of different job tasks. The results are expected to indicate the extent to which common duties are shared by workers in different trades and suggest a job family structure that would guide examination development.



PREVIOUS PAGE IS BLANK

For want of a nail the horseshoe was lost. Because the horseshoe was lost, the horse, the rider, the battle, the war and the kingdom were lost. For want of a nail the kingdom was lost. Had the blacksmith secured the horseshoe with that critical nail, then the kingdom would have been saved. Why then, do we pay more attention to horses and warriors then to blacksmiths who secure the horseshoes?

The growing sophistication of naval equipment and technology is creating a need for a highly skilled workforce to maintain and repair ships, aircraft and weapons systems. Approximately 7000 apprentices in some 60 different trades and crafts are trained each year in Navy's civilian apprenticeship programs. These training programs, which equip the apprentice to perform the duties of a skilled craft at a journeyworker level, consist of academic and trade theory instruction as well as on-the-job training.

The selections made to Navy's apprenticeship programs are highly competitive. The Office of Personnel Management (OPM) recruits and examines tens of thousands of individuals each year for positions at naval shipyards, air rework facilities, and public works centers. Applicants are evaluated by OPM on their aptitude and interest for learning trade theory and practice, ability to follow oral directions in a shop, and reliability and dependability. If found qualified, applicants' names are placed on a register from which appointments are made.

The Navy Department and OPM have begun a joint project to develop and validate a new examination for the selection of applicants to apprentice training programs. The goal will be to develop a legally and professionally defensible examination that will update or replace the current examination that was constructed many years ago. A second, but not less important, goal of the project is to determine the relevance and adequacy of apprentice training to the job. Since the Navy Department is the largest single employer and trainer of apprentices for blue collar occupations in this country, it is interested in ensuring that its apprentice training programs impart the knowledges and skills necessary to effectively carry out the duties of a skilled trade. The third objective of the project is to identify and describe the common activities and skill requirements of Navy's trades and crafts. This goal stems from a need to make the selection and training of apprentices both cost-effective and administratively efficient. From an administrative viewpoint, it is more desirable to use a single or a limited number of examinations to select personnel to all skilled blue collar occupations rather than to use a separate examination for each occupation. Similarly, a core training program that will teach all apprentices the required common skills will increase the efficiency and cost-effectiveness of instruction.

The goals of the project could best be met by establishing an occupational data base through a comprehensive occupational survey. Such a survey provides the job-analytic data needed for examination development and training program evaluation and improvement. It was decided that the most efficient and valid way to collect the relevant job information is to survey the job incumbents in the skilled trades. Fully skilled workers who are currently employed are likely to provide the most accurate and current information about the nature of their job.

A major problem confronting the project in its initial stage was how to conduct an occupational survey, including a job task survey, of 60 different trades and crafts. Since the job of an aircraft mechanic, for example, is so different from that of a carpenter, it became evident that we could not use the same task inventory for these two apparently dissimilar occupations. We decided to avoid part of the problem by surveying only 22 of the occupations. These 22 are the key trades and crafts in which most of the workers are employed. The occupations were chosen in such a way that the three major Naval activities - - shipyards, air rework facilities, and public works centers -- would be represented. The occupations selected are listed in Appendix A.

Although the decision to survey 22 rather than 60 occupations made the project more manageable, it did not solve the problem of identifying the common duties of different jobs. A common denominator was needed to describe the variety of work activities that are performed in the craft occupations. If similar duties, skills, and operations exist in the different trades, even though the tools, materials and processes may be different, then the duties could be described by using a common set of job descriptors. How, then, could a set of common job descriptors be developed?

Fortunately, we did not have to start from scratch. Task inventories in each of the 22 target trades and crafts were already available. these inventories were prepared as part of an earlier effort by the Navy Department to develop performance appraisal instruments for blue collar civilian personnel. The inventories were developed by instructors, supervisors, and workers familiar with the jobs. Although the inventories were at that point in time only in draft form, they did provide us with the information necessary to identify the common job descriptors. The individual task inventories ranged from approximately 100 to 1000 task statements grouped into major duties. All task statements were reviewed by three OPM psychologists to identify the overlap in activities among the trades. This review resulted in the identification of 23 major job duties that are common to two or more of the 22 trades. The list of common duties was reviewed by 79 of Navy's subject-matter experts. These reviewers, who represented persons familiar with all of the target trades, expanded the list to 27 duties. The final list of common duties is presented in Appendix B.

The list of common duties was incorporated into a comprehensive occupational survey instrument that was administered to approximately 5000 job incumbents in the target occupations. Each duty was rated by the respondents on "relative time spent" performing that duty and its relative importance using five-point scales. Although the duty ratings have not yet been analyzed, I will give a brief description of the analysis plan. First, the percentage of workers performing each duty in each trade will be compared. Those duties that are performed by a majority of workers in all of the trades and are also above average in terms of "relative time spent" and "importance" may form the job analytic base for a single test or test battery that would be used to select apprentices. On the other hand, a cluster analysis of the ratings may indicate that the data can be accounted for by two or more job families that require different skills. Whether a single selection examination could be used for all targeted occupations will depend not only on the common duties shared by the various trades, but also on the common skills required for job success.

The Uniform Guidelines on Employee Selection Procedures (1978) indicate that similarity of jobs for the purpose of test validation must be shown in terms of common "work behaviors." Previous research on occupational grouping for prediction of job success suggests grouping of skilled craft occupations for test validity generalization is feasible. Long-term research on the tests of the General Aptitude Test Battery (GATB) indicates that all but one of the 22 trades targeted in this project fall within a single "Occupational Ability Pattern" (U.S. Department of Labor, 1979). These findings suggest that the same pattern of abilities is required for success in the different trades. More recent research by Schmidt, Hunter, and Pearlman (1981) indicates that aptitude tests are likely to be valid across many jobs, even when the jobs differ grossly in their task makeup. Thus, the use of a common test battery for selection to more than one craft occupation may be appropriate even if the occupations have few or no duties in common. The bottom line must be the validity of the test for predicting success in an occupation.

The list of common duties that was developed to find occupational overlaps became one part of the occupational survey questionnaire. The other parts sought information about the background of job incumbents, relevance of apprenticeship training to the job, muscular effort required, job hazards, tools and equipment used, and job tasks. I will briefly describe the contents and rationale for each of these other parts.

The biodata section sought relevant biographical information about the respondents, including work location, age, race and ethnicity, educational level, length of time in the Navy as a civilian worker, career path, and nature of prior training. These data will be used to describe the respondent sample, divide the sample into appropriate subgroups for statistical analysis and correlate certain biodata variables with responses on other parts of the questionnaire. This part of the survey questionnaire also contained questions about hazards or dangers encountered on the job, injuries received, and the frequent physical discomforts. One of the uses of these data will be to develop realistic previews of the job for prospective applicants.

Although many of the jobs surveyed in this study appear to require substantial physical strength or muscular effort, there are no current strength standards for them. The desire by the Navy Department to recruit and train women in many of the skilled blue collar occupations in which they have been traditionally underrepresented requires that physical standards be defined more precisely than in the past. Since differences in certain types of strength have been documented in previous studies (Robertson, Note 1; Wilmore, 1974), it is now important to establish physical standards that specify the physical demands of the job.

Using a taxonomy of occupationally-oriented Basic Body Efforts (BBE) developed by researchers at the Naval Personnel Research and Development Center (Robertson, Note 1), this part of the questionnaire requested information about the single most muscularly demanding task of the respondent's job. Respondents are asked to describe the specific object, tool, or control moved and what is done with it. They are then asked to indicate the type of Basic Body Effort that is applied (e.g. "Turn-lever" such as using a wrench to loosen corroded mounting bolts), the frequency of performing the task, the number of persons usually teamed together to exert the required force, and whether the

effort required is within the respondent's capabilities. The foregoing information will be helpful in establishing physical standards for different blue collar occupations as well as to identify relevant criterion measures against which tests of physical performance could be correlated.

Although physical strength is not a characteristic that can be easily trained, if at all, much of the knowledge and skill applied by the craft worker on the job on a day to day basis is learned during the apprenticeship period. In addition to on-the-job training, apprenticeship programs consist of formal classroom instruction in trade theory and academic subject matter. The need for teaching the apprentice the specific content of his or her trade is self-evident but the need for basic academic instruction, such as English and Mathematics, is less well accepted. Since there is often a lack of direct correspondance between academic learning and job tasks, a controversial issue in apprentice training is the relevance of academic instruction to the job. On one side, it is argued that basic instruction must be broad enough to provide the apprentice with the skills required to grasp new technology. Training must provide the apprentice with the basic skills for performing the job as it is today as well as how it may be tomorrow. On the other side, it is argued that the skills aquired in training should be ultimately used or applied in the course of one's work. Otherwise irrelevant learning will consume and waste valuable training time and will unfairly discriminate against those who are unable to cope with academic learning.

One section of the survey questionnaire attempted to determine the classroom subjects that are applied on the job and whether these subjects were adequately taught in the apprenticeship program. A list of 18 academic subjects was presented, and respondents indicated whether they use a particular subject occasionally, frequently, or not at all and whether the apprenticeship training in the subject was adequate. The analyses of these data will identify the subject matter that is not relevant to certain crafts and trades as well as those disciplines in which formal instruction may require added emphasis and improvement.

The final two sections of the survey questionnaire concerned the tools and equipment used in a trade and the job tasks performed. Since many tools are unique to one or a few trades, a different checklist was prepared for each trade. These checklists, which included hand tools, bench tools, power tools, and measuring devices, were used by respondents to indicate the frequency with which they used various tools. The job task list, like the tool list, was unique to each occupation. The number of task statements ranged from 55 for the Insulator trade to 372 for the Electrician trade. As was mentioned earlier, draft task inventories were available from an earlier project in the Navy and these inventories were reviewed by OPM psychologists for clarity, consistency, and objectivity. They were then reviewed further by at least four subject matter experts in each trade. The final lists of tasks encompassed all of the important and nontrivial activities that might be performed by workers on the job. The task statements were designed to represent activities that are directly observable and are at the same level of generality. Tasks that were related to each other, either because they are similar or because they tend to occur concomitantly, have been grouped under general duties.

Job incumbents completed the task inventories by indicating whether or not they perform each task, the relative amount of time spent performing the task, and the relative importance of the task. The decision to use these particular scales for rating tasks was based on the requirements of the Uniform Guidelines on Employee Selection Procedures (1978) and on previous research on task inventories. The Guidelines indicate that the "critical or important work behaviors" should be identified when conducting a job analysis for content validity. Extensive research by Christal and his associates (Christal, 1974) indicates that workers can apparently state with confidence that they spend more time with one task than with another.

Before being finalized, the occupational survey questionnaires were reviewed by job experts and pretested on a small sample of job incumbents. The job experts, who included supervisors, journeyworkers, instructors, and apprentices in each trade, reviewed the questionnaires according to instructions provided by the project staff. The review and pretesting resulted in several changes being made in the questionnaires. The final version of the questionnaires was administered to a stratified random sample of over 5000 workers at 15 Naval installations throughout the United States.

The analysis of the data collected, which is now taking place, focuses on the major objectives of the project: development of valid selection procedures and improvement of training programs. In developing valid selection procedures it is necessary to identify the critical work behaviors or tasks as the first step in determining the abilities required to perform a job successfully. The identification of common duties shared by different occupations will allow the development of common selection procedures for clusters of blue collar occupations and the improvement and standardization of the basic elements of apprentice training programs. A common examination for skilled blue collar occupations will be cost effective both in its development and administration. If found to be valid, the use of the examination will aid in the selection and training of a highly productive and skilled blue collar workforce.

## REFERENCE NOTE

1. Robertson, D.W. Development of job strength requirements. Paper presented at the XXIV International Meeting of the Institute of Management Sciences (TIMS), Honolulu, Hawaii, June 18-22, 1979.

## REFERENCES

1. Christal, R.E. The United States Air Force occupational research project. (AFHRL-TR-73-75), Lackland Air Force Base: Texas, 1974.

2. Schmidt, F.L., Hunter, J.E., & Pearlman, K. Task differences as moderators of aptitude test validity in selection: a red herring. Journal of Applied Psychology, 1981, 66, 166-185.

3. Uniform guidelines on employee selection procedures. Federal Register, 1978, 43, 38290-38315.

4. U.S. Department of Labor. Section II: Occupational aptitude pattern structure. Manual for the USES General Aptitude Test Battery. Washington, D.C.: Author, 1979

5. Wilmore, J.H. Alterations in strength, body composition and anthropometric measurements consequent to a 10-week weight training program. Medicine and Science in Sports, 1974, 6, 133-138.

## APPENDIX A

## NAVAL TRADES AND CRAFTS SURVEYED

Air-conditioning Equipment Mechanic

Aircraft Electrician

Aircraft Engine Mechanic

Aircraft Instrument Mechanic

Aircraft Mechanic

Boilermaker

Boiler Plant Operator

Carpenter

Electronic Mechanic

Electrician

Electroplater

Equipment Mechanic (formerly Marine Machinist)

Inside Machinist

Insulator

Painter

Pipefitter

Rigger

Sheetmetal Mechanic

Sheetmetal Mechanic (Aircraft)

Shipfitter

Shipwright

Welder

## APPENDIX B

## DUTIES THAT MAY BE COMMON TO TRADES AND CRAFTS

1. Read and interpret manuals, written instructions, forms, and official documents
2. Read and interpret blueprints, diagrams, or drawings
3. Read and interpret graphs, tables and other written numerical materials
4. Read and interpret visual displays such as dials, gauges, clocks, and signal lights
5. Prepare written reports
6. Use measuring devices such as rulers, calipers, voltmeters, and thermometers
7. Follow oral instructions
8. Give oral instructions
9. Inspect or test materials or equipment
10. Estimate the speed of moving parts or objects
11. Estimate the quantity of material
12. Estimate the size or weight of materials or equipment
13. Troubleshoot and repair equipment
14. Set-up and adjust equipment
15. Guide or control materials being processed
16. Assemble or disassemble parts or equipment
17. Clean parts or equipment
18. Teach or instruct trainees about job
19. Operate hand-held or portable tools (including power tools)
20. Operate bench tools and stationary machinery

21. Lead or oversee the work of others

22. Plan the best way to complete an assignment

23. Perform addition, subtraction, multiplication, or division

24. Use or calculate fractions, decimals, or percentages

25. Use algebraic formulas to make computations

26. Use geometry

27. Use trigonometry

Sharp, Bradford P. & Allen, [illegible]
Coast Guard, DC [illegible]

A Performance Management System for the USCG: Strategy for Development and Implementation

The Officer Fitness Reporting System presently being used by the US Coast Guard was begun in 1965 for the purpose of providing valid information for use in officer promotion and assignments. During its lifetime, the system served remarkably well especially when compared to the appraisal systems of other organizations. In recent years, however, the system has been subjected to accelerated inflation of marks, increased challenge and a general decline in confidence among members of the officer corps. As a result, the Coast Guards' Office of Personnel formed the Officer Performance Evaluation Study (OPES) group to research, design, and introduce a replacement.

This paper will present the strategy adopted by the OPES group in the development and implementation of a new and significantly different performance evaluation system within the Coast Guard. It will describe the strategy employed by the group in achieving the acceptance of a new personnel management philosophy by Coast Guard leadership and users of the new system. In addition, it will outline the approach adopted for handling the enormous amount of work which had to be accomplished prior to the projects' absolute completion date. Included in the discussion will be a description of the specific tasks carried out in the process.

AD P001373

# A Performance Management System for the U.S Coast Guard: Strategy for Development and Implementation

Bradford P. Sharp
Nicholas H. Allen

U.S. Coast Guard Headquarters
2100 2nd St. S.W. G-PO/OPES
Washington, DC 20593
(202) 472-4773

In February of 1980, the Coast Guard began a study to look at and, if necessary, replace its existing performance appraisal system. The study, conducted jointly with an outside contractor, was originally estimated to be a three year effort. Rapid deterioration of the existing system, however, resulted in the imposition of a deadline for implementation of a replacement system. The three year project now was to be completed in eighteen months.

This paper is a look at the development of this system as well as the strategy used to achieve its' implementation within the reduced time frame. It will provide some insight into organizational change and the manner by which it can be brought about.

## Introduction

The Coast Guard's present officer fitness reporting system was developed in 1965, and except for minor modifications, has remained unchanged since it was introduced. The system has functioned remarkably well over the years especially when compared to performance appraisal systems of other organizations which typically last 2-4 years. However, like all appraisal systems which help to determine promotion and assignment, the Coast Guard's system began to experience inflation on the appraisal form's rating scales. The continuing inflation was also combined with an increasing number of successful attacks on the promotion and appraisal process by dissatified officers through the Board for the Correction of Military Records (BCMR). These two factors proved to be the primary reasons behind the initiation of a study by the Coast Guard Chief of Personnel to examine the officer evaluation system. The Officer Performance Evaluation Study (OPES), as the effort was called, was given a mandate to "review the existing appraisal system and, if appropriate, make recommendations for change". We approached this objective by dividing it into five major areas of work which are as follows:

-Define the Coast Guard environment as it may relate to the performance appraisal of officers.
-Establish the functions the officer fitness reporting system should be expected to serve and the criteria by which to judge these functions.
-Review the current system in terms of its ability to meet the previous two tasks and, if necessary;
-Identify prototypes which accomplish these functions, meet the criteria, are compatable with the Coast Guard environment, and correct for current system short falls.
-Identify the optimum prototype.

After 12 months of work, the OPES group accomplished these five tasks identifying a need to substantially change the existing appraisal system. After successfully convincing Coast Guard leadership of this need and gaining their acceptance for our prototype design, we began the sizeable task of implementing a new appraisal system in the Coast Guard. Our implementation program and final system design was divided into four major areas of effort described in the following steps:

-Develop a training orientation program for the officer corps.
-Operationally test the prototype across a Coast Guard district
-Implement the system, from the top ranks down.
-Develop an automated data process (ADP) system to monitor the program, once implemented.

DEVELOPMENT OF THE COAST GUARD APPRAISAL SYSTEM---THE RESEARCH APPROACH
(See Figure 1)

State-of-the-Art

The Officer Performance Evaluation Study (OPES) group from the outset adopted a practical, action-oriented approach to its research. It began with an in-depth look at the state-of-the-art in performance appraisal, both theory and practice. Several military, private, and public sector organizations were contacted in an effort to study alternative performance appraisal systems. Although most of these organizations had appraisal needs very different from our own, it was valuable to identify how they dealt with their particular appraisal problems. The study group through this process was able to identify many techniques which frequently provide positive results, as well as several which frequently fail. This survey also provided the Coast Guard with a wealth of research data accumulated in other study efforts at relatively little expense.

The OPES Group

The Officer Performance Evaluation Study (OPES) Group was directed and coordinated on a full-time basis by the authors, a Coast Guard Commander and a civilian assistant. We were greatly assisted by the help of a full time outside consultant who brought considerable experience in the design of appraisal systems. In addition, working with us on a part time basis were several highly qualified officers and civilians who formed the two OPES advisory groups. Advisory Group I consisted of personnel and management experts who provided professional information, both theoretical and Coast Guard related on a day-to-day basis. Advisory Group II were upper level program managers who proved to be very important in terms of the assistance they provided in policy related matters and especially in overcoming bureaucratic obstacles which frequently plagued us in the beginning. Many times, major setbacks were avoided over matters such as shortages of money or manpower when one or more members of the group came to our assistance.

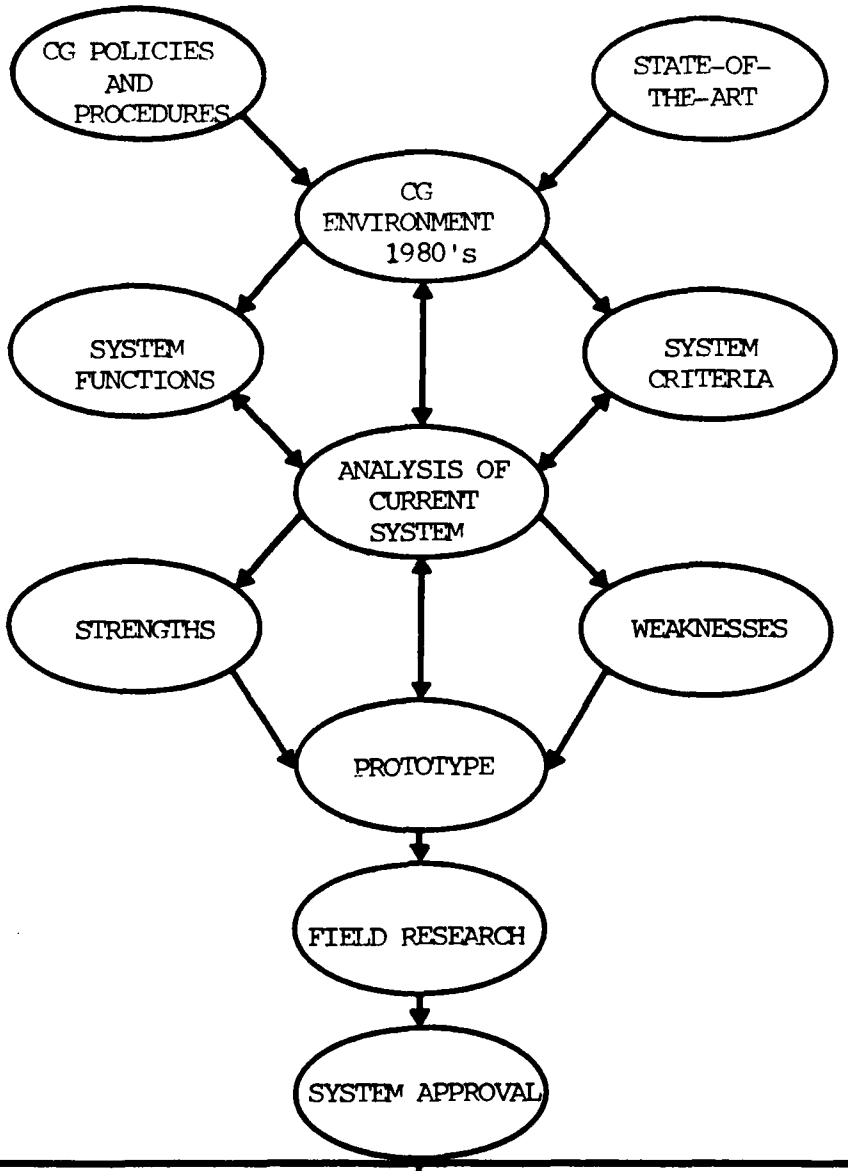
Bringing the members of these two groups into the study process and keeping them informed proved many times to be a wise decision. Besides providing us with a wealth of human knowledge and dedicated assistance, they helped to develop a broad base of top level management and expert support within the organization. This would prove to be the difference between success and failure in the later months of the study. (See figure 2)

Legal Review

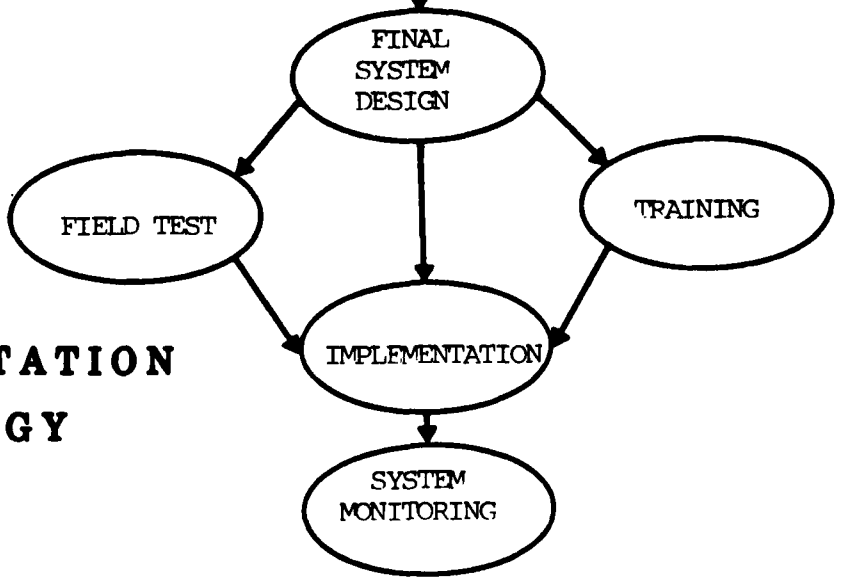
The study group conducted an in-depth review of the laws governing the management of the Coast Guard officer corps, specifically those in Title 14 of the U.S.CODE. In addition, Title 7 of the 1964 Civil Rights Act and the 1978 Uniform Guidelines were relevent since they provide organizations with a framework for designing and executing personnel actions such as performance appraisal.

The Uniform Guidelines, adopted by the government agencies concerned with enforcement of the Civil Rights legislation of Title 7 US Code, represent the most significant government work on performance appraisal to date. The basic thrust of the guidelines is to encourage employers and organizations to maintain formal, validated performance appraisal programs. Today most organizations are designing or revising their appraisal systems with the

# RESEARCH APPROACH



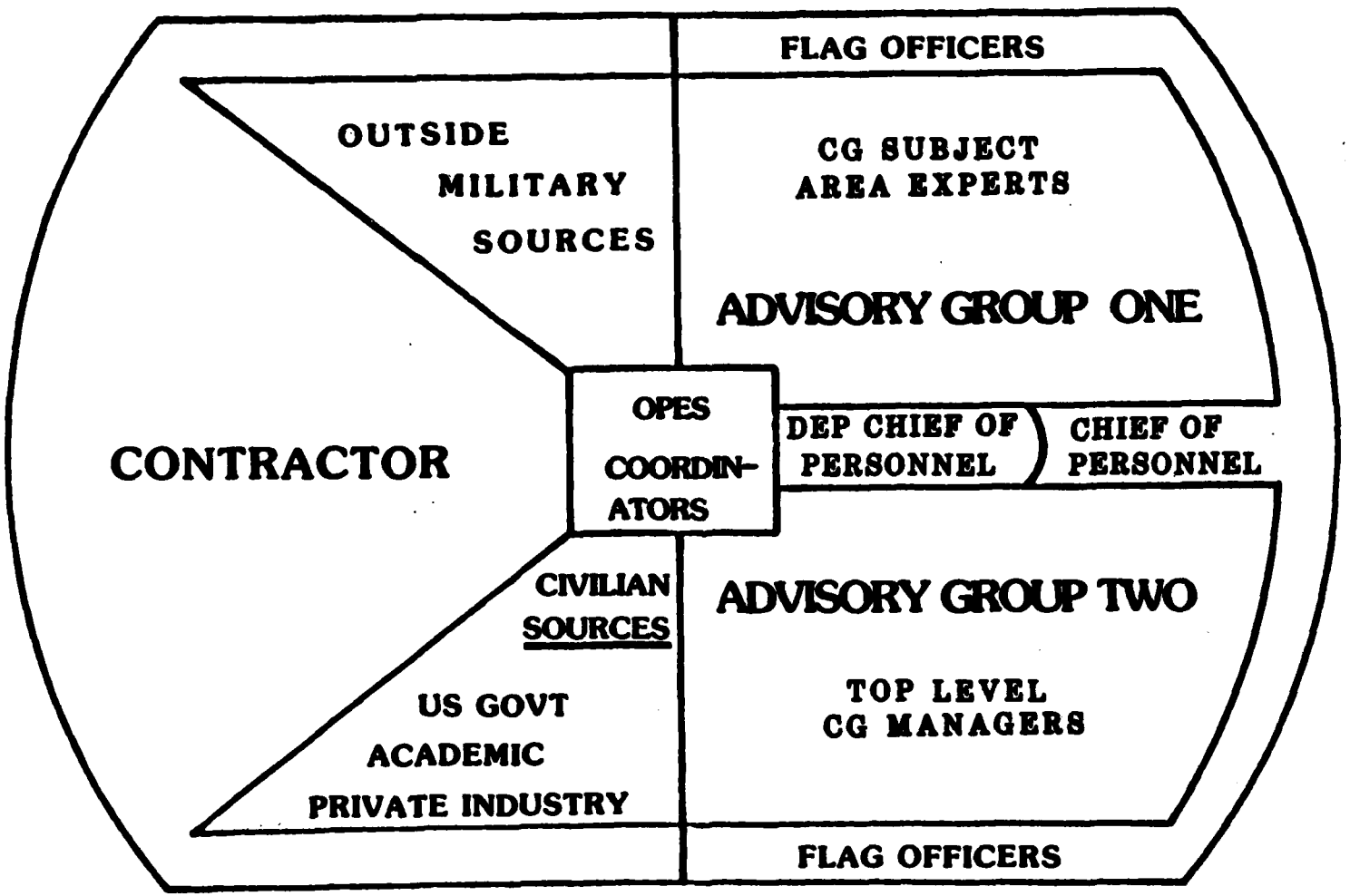

# IMPLEMENTATION STRATEGY




(Figure 1)

# OFFICER PERFORMANCE EVALUATION STUDY

# (OPES) RESOURCE ENVIRONMENT




(Figure 2)

guidelines as the starting point. Although the Coast Guard as a military service was not required to conform to the Uniform Guidelines we did not see a conflict between the military environment and establishing a performance appraisal system that would be job-related, avoid adverse impact, and be based on one of the forms of validation acceptable to the guidelines.

## Coast Guard in the 1980's

Because the success of an appraisal system depends heavily on the organization in which it will be used, it was necessary to look at the direction of the Coast Guard as it entered the 1980's. Among the issues considered by the study group were:

- -the increasing complexity of Coast Guard missions and the resulting trend toward officer specialization.
- -the increasing qualification of officers entering the service through the highly selective Coast Guard Academy and OCS programs.
- -the increasing sophistication of the officer corps in terms of their awareness of promotion and appraisal procedures and policies.
- -possible changes to the Coast Guard promotion system and other personnel policies and procedures.

We felt it was necessary to consider the environment and specifically the challenges a performance appraisal system would have to face in the next ten years. Our examination of the current fitness reporting system and any recommendations for changes would therefore be made not only in consideration of present needs but also those of the future.

## Policies and Procedures

Our investigation of the existing Coast Guard fitness reporting system began with a look at both the policies and procedures affecting, and affected by performance appraisal. These included; the roles of the various individuals in the rating chain, the personnel manual regulations, how the forms were used for promotion and assignment, handling of the forms at headquarters, and even ease of typing. It was felt that in order to really understand the performance appraisal needs of the Coast Guard, a complete knowledge of the formal and informal structure was necessary. It was a primary strategy behind the study to capture the perspectives of all people directly or indirectly affected by the fitness reporting system. We determined there were three primary categories of individuals. These groups could provide the best insight into the true state of the existing system, as well as provide any evidence supporting a need to change. The three categories were identified as the actual fitness report form users, the leadership, and the officer corps in general. A discussion of these perspectives follows.

## Policies and Procedures - Perspective 1: Headquarter's Use

Two groups were considered "Headquarters users"; selection board members responsible for both promotion and special assignment such as command, and assignment managers responsible for the routine assignment of officers. To gain the views of each group, two approaches were taken. First, a mail survey was sent to all members of the boards (approx. 90 officers) convening between 1978 and 1979. The purpose of the survey was to identify the requirements of

an appraisal form in the selection and promotion process and then to identify the strengths and weaknesses of the existing forms in light of those requirements. Secondly, interviews were conducted with sitting promotion boards immediately at the conclusion of each board process. The interviews attempted to follow up on the information gained in the mail surveys and also test alternatives for change postulated by the study group.

The second group of Headquarters users, the assignment managers, were subjected to in-depth interviews which focused on the use of the fitness report in the assignment process. The interviews sought to obtain several types of information including: the role of the fitness report in the assignment process; the strengths and weaknesses of the current system as it relates to assignment managers' duties; and recommendations for changes that would improve the effectiveness of the assignment managers in their jobs.

Policies and Procedures - Perspective 2: Leadership

The second perspective on the officer appraisal system was that of the Coast Guard leadership. In an organization such as the Coast Guard, the attitudes, perceptions and values of the leadership will have a great influence on the direction and perspectives taken by the whole of the organization. Especially in an area as sensitive as performance appraisal, there was a critical need to understand the views and opinions of the 27 admirals who comprised the leadership body and to gain their support and understanding. To do this, the study group conducted on-site personal interviews with each of the admirals to not only solicit their views but also gain their support. Each flag officer was given a briefing on the study, as well as, possible system alternatives. Their opinions and reactions were recorded and summarized providing a valuable data base for much of the work done later in the study.

Policies and Procedures - Perspective 3: Officer Corps

The third perspective was that of the officer corps as a whole. No personnel system having a significant impact on its members' career status can hope to be effective without the support of the individuals it affects. Our group hoped to gain both insight and support through a comprehensive mail survey sent to one-third of the officer corps (about 2000 officers). By allowing a large number of officers to provide input to the study not only did we gather a data base but, in addition, we created interest in the group's work. Included in the data collected were: attitudes of seniors and subordinates toward carrying out appraisal responsibilities, attitudes toward system alternatives, attitudes toward the present system, and variations in opinion according to career fields.

Integrated Analysis of the Data

The survey results, personal interviews, reviews of institutional policies, reviews of state-of-the-art methodologies and reviews of the law, enabled the OPES group to establish with reasonable confidence the organizational functions that should be served by an appraisal system in the Coast Guard. The task then was to identify a system with the proper balance of appraisal techniques which would meet the needs and expectations of the organization while not exceeding the limits of organizational acceptability. To accomplish this, the study group established the following critera against which an appraisal system should be measured.

-It should accomplish the functions identified by the organization.
-It should be accepted by the users.
-It should conform to the Uniform Guidelines.
-It should be consistent with a military organization.

Using these criteria as a guide, each section of the existing fitness report form and the corresponding policies were analyzed. The result of this analysis was an identification of the strengths and weaknesses of the fitness reporting system.

Field Research

With the identification of the specific strengths and weaknesses of the existing system, efforts to build a replacement system began. The approach used in the development of the initial prototypes was to build upon the strengths of the current system while selecting state-of-the-art, organizationally realistic, techniques to replace the weaknesses. Choosing from several options, the study group eventually focused on one basic prototype model while maintaining a number of alternative techniques.

In order to gain additional information on the feasibility of the prototype and some of its various options, a field research excercise was developed. The test was designed to provide information on the success of the model under real Coast Guard operating conditions and to determine if the officers tested possessed the management and performance appraisal skills necessary to carry out the requirements of the models. Conducted in the Fifth Coast Guard District (Portsmouth, VA), the activity included approximately 110 officers involved in a variety of Coast Guard activities. The participants experienced a modified appraisal period using the test system in the recording of actual performance. Following the exercise, two questionnaires were conducted, one addressing the merits of the system itself, the other assessing the officers' management skills.

The field research activity proved to be very valuable in terms of reducing the uncertainty that the group felt for its design work. It also provided an assessment of the training needs that would have to be met if the prototypes "process approach " to appraisal were implemented in the officer corps. The test confirmed our group's belief that performance appraisal training would be necessary with the introduction of the new system.

System Recommendation: Conclusion of Research Phase

In April of 1981, fourteen months after the beginning of the OPES project, the Commandant of the Coast Guard was presented with a design for a new officer appraisal system and a strategy to bring about the system's implementation. Following a lengthy briefing and discussion, the system and plan for its introduction was accepted. This event concluded the first phase of the OPES study. The period up to the Commandant's acceptance of the precepts of the system, had been essentially a research and development phase. The next phase however, would be significantly different. Having designed a new system which would meet the criteria we had set earlier, we changed our focus to preparing the organization for what we began to see as a major organizational intervention. The OPES group now had to develop the policies and procedures to achieve the system functions yet not make it organizationally unworkable.

DEVELOPMENT OF THE COAST GUARD APPRAISAL SYSTEM ---- IMPLEMENTATION STRATEGY

## Final System Design

The final decision on many of the accompanying features, policies and procedures, took months to complete. New personnel manual regulations had to be drafted to reflect the restructured policies and procedures. The forms were examined for issues such as color, layout, ease of typing and administrative handling. Ancillary issues such as educational evaluations and non-military raters had to be addressed. Even the wording on the forms had to be carefully chosen. To assist in this last task a panel of eleven Coast Guard officers representing the various specialties of the officer corps was convened at Headquarters. The group was used to establish the performance standards and organizational values against which all Coast Guard officers should be evaluated, as well as the terminology to be used in all rating scales on the form. Finally, with the forms designed and the regulations completed in first draft, a full test could begin.

## Field Test

Conducted in the 8th Coast Guard District (New Orleans) and affecting over 300 active duty and reserve officers, the test presented an opportunity to observe the total system prior to actual implementation. Although the major concepts of the system were known to be workable, we felt there were still some administrative and user problems we were unaware of. The field test was an attempt to identify these problems, and examine how the system would be received when introduced in a Coast Guard district. Overall, the exercise proved a success. The attitude of most officers who participated in the test was positive and, aside from the additional time they felt the new system would require, most were glad the change was being made. A battery of five questionnaires was given at various stages of the test period soliciting participants' attitudes towards the system and the confidence level felt in carrying out the various appraisal period requirements. The surveys were reinforced by personal interviews conducted at the test's conclusion.

## Training

Throughout the OPES effort we felt training would be integral to the introduction of any new system. Evidence in support of this intuition was replicated in the various questionnaires, interviews, tests, and reviews of how well Coast Guard officers were carrying out performance appraisal. Consequently, research on possible training options was begun early in the study. Our work showed that theories on training managers in performance appraisal skills are abundant but usually are unproven in large organizations. Identifying an optimum solution for the Coast Guard in terms of both cost and program effectiveness was difficult. Ultimately, a combination orientation in the new system and brief introduction to performance appraisal skills was chosen. The one-day session was designed with two objectives; first, to introduce the new forms and procedures and second, to provide an introduction to proper performance behavior through video observation and classroom practice. Through the use of several

videotaped scenarios, officers observe the modeling of skills needed in effective performance counseling and appraisal. The hope is that the 8 hour session will provide some familiarity with each technique and possibly provide some competency in the skill. The program was also designed from the standpoint that it will be the best opportunity we have to sell the officer corps on the new system. It emphasizes heavily the benefits which may be accrued if the system is used and, by example, it tries to eliminate the misconceptions many officers might have. The training which has not begun at the time of this writing will be carried out within each district by senior officers chosen for their credibility within the corps. All officers will receive the training prior to participation in the new system.

## Implementation

On January 1st 1982, the Coast Guard will begin under the new Officer Performance Management System (OPMS). The name was changed from fitness reporting system to remove any association with fitness for duty that the old term might imply. There were two concerns for implementation of the new system; that it begin first with the senior officers and work down the rank structure and; that it place as little initial burden on the Coast Guard as necessary. To accomplishing this, the system will begin with the top three grades of officers on 1 January the remaining grades to begin with the start of their normal spring rating period. This approach not only insures that senior officers are familiar with the OPMS system but that they will have the experience to provide guidance to their subordinates.

In order to facilitate an officer's understanding of the new rating system and provide for its' ease of use, a "how to do it" guide to the OPMS is also being developed and will be supplied to each Coast Guard officer. This guide, written in an easily understandable format, is intended to address all features of the system and discuss major problems and issues. It will also help to reinforce the learning received in the training program.

## System Discipline

The months immediately following introduction of the OPMS will be critical to its success or failure. There must be a complete break from the old system to the new and the requirements of the new system must be enforced. Those officers who retain their bad habits from the old system must be identified and educated. Those who continue to inflate marks must be held accountable for their failure to uphold a management system which is vital to the organization's accomplishment of its missions.

One of the techniques which will be used to assist in the maintenance of the new system will be an automated data process (ADP) system. Information from each appraisal form, called the Officer Performance Report (OPR), will be computerized at Coast Guard headquarters. In addition to providing general data on the system, such as overall trends, geographical marking differences, etc., the ADP system will also provide data on the marking habits of each officer. This marking data will be used to establish the credibility of reports written by each officer during his or her career. The intent is to reduce the value of reports written by officers who consistently inflate their

marks while at the same time protecting officers who rate their subordinates accurately. Promotion boards will receive information that will allow them to review each report in light of the rating history of the individual writing the report. Promotion boards will also receive information on a officer's rating history when that individual is up for promotion, thus providing incentive for an officer not to inflate subordinate evaluations. Through this process, it will be possible to identify those officers failing to comply with the system and provide necessary inducements; such as personal letters to the individual and letters through the chain of command, to bring about compliance.

## Conclusion

If the Officer Performance Management System is successfully implemented on 1 January 1982 and, more significantly, if it is accepted by the officer corps, it will be attributable to two factors. The first is that it was developed using a systems approach. Unlike what is all to often the case in many large organizations, the new appraisal system will not be the product of a few people working in isolation to develop a new policy. It was the result of an intensive research effort which looked at many issues relating to performance appraisal both directly and indirectly. Information was collected from many organizations, both public and private, to gain as much research data on performance appraisal as possible. Dozens of performance appraisal techniques were examined but always while keeping the realities of the Coast Guard organization in mind.

People provided a great deal of our information. Interviews and discussion were held with hundreds of people both inside and outside the Coast Guard. Perspectives which varied from university experts to newly commissioned Ensigns were considered. One-third of the officer corps was surveyed in a general questionnaire and hundreds more participated in operational tests or other surveys during the project. One-on-one or small group interviews were carried out with all users of the system including the typists. In short, this was a system designed not only to meet the needs of the Coast Guard organization but also the needs of its people in whatever role they play.

The second factor to which we would have to attribute any eventual success, would be our decision to keep as many people as possible informed on the progress of the study. By providing information and soliciting feedback we gained ownership and the support of many individuals holding critical positions in the organization. Involving these senior managers early we were able to identify areas of concern at a time when we had relative flexibility especially in terms of time. As the project progressed and the system began to take shape along with the realization that we were now heading toward a major organizational change, these managers provided OPES with a support base. Having eliminated most of their anxieties through education, and having sought their input in many of the decisions, we created on their parts understanding if not actual ownership. These alliances proved to be, during difficult times later in the study, one of the major reasons we were able to finish the 36 month effort in half the time.

This paper reflects a relatively small aspect of the new Officer Performance Management System, that of its development. Its intent was to describe the development of what will be a major management and organizational change in a military organization. The real story will begin once the system is actually implemented within the officer corps. Although we recognize it is not the "perfect system" we believe it is a significant improvement over our current system and will offer many organizational and individual benefits. It is the willingness of the officer corps to accept this change that now remains to be seen.

AD P001374

Shields, Joyce, Hanser, Lawrence, Williams, Edward, CPT and Popelka, Beverly,
U.S. Army Research Institute for the Behavioral and Social Sciences,
Alexandria, Virginia. (Wed. PM)

## Pilot Research for Validation of ASVAB and Enlistment Standards Against Performance on the Job

The Army Research Institute (ARI) conducted an initial pilot research project in the 193rd Infantry Bridage, Panama to determine the feasibility of validating the Armed Services Vocational Aptitude Battery (ASVAB) and enlistment standards against performance on the job. This report discusses some preliminary analyses of these data. The preliminary analyses focus on two areas. The first area deals with the definition of a "successful soldier," and the extent to which commanding officers and NCO's agreed on the qualities of a 'successful soldier.' The second area deals with the relations among ASVAB 5/6/7, SQTs, and selected preliminary measures of job performance. The result. suggest that the three most important factors in overall soldier performance as indicated by supervisor consist of job performance, troop responsiblity, and discipline. The relationship between ASVAB and existing measures of job performance (e.g., SQT scores, awards, honor graduate, and letter of appreciation) are also discussed.



PREVIOUS PAGE IS BLANK

# Pilot Research for Validation of ASVAB and Enlistment Standards Against Performance on the Job

Joyce L. Shields, Lawrence M. Hanser, Edward W. Williams, and Beverly A. Popelka

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

In January 1976, the Armed Services Vocational Aptitude Battery (ASVAB) was introduced as the single DOD selection and classification battery. The ASVAB provides an Armed Forces Qualification Test (AFQT) score, which consists of the word knowledge, arithmetic reasoning, and space perception subtests. This test is the basic DOD enlistment test required by Congress as a means of screening applicants for overall trainability and English language proficiency. The remaining components in the ASVAB were derived from the individual service classification test batteries, and are used for the differential assignment of volunteers to specific inservice Technical training courses. In all, ASVAB (Forms 5, 6, and 7) contains 13 subtests, These subtests along with a brief description of each are listed in Table 1.

For Army use, ASVAB subtests are further combined into nine Aptitude Area Composites. Minimum scores on these composites are used as a prerequisite for entering skill training programs. Successful completion of the training program results in the award of a Military Occupational Specialty (MOS). For example, one Aptitude Area Composite is labeled CO for Combat and is used to classify recruits into Infantry and Armor specialties; another composite is labeled EL, for Electronics Repair, and is used for all electronics repair specialties in the signal and air defense fields (Maier & Grafton, 1981). The Aptitude Area Composites along with the types of skill specialties for which the composites serve as prerequisites are shown in Table 2.

## Miscalibration of ASVAB Scoring Table

Subsequent studies revealed that a miscalibration occurred in the ASVAB 5/6/7 scoring table during its development. As a result, the Services enlisted recruits who would have been turned away, because of low test scores if the ASVAB had been calibrated correctly.

The calibration problem and congressional interest as to its impact prompted the issuance if a memorandum dated 11 September 1980 from Robert B. Pirie Jr., directing all services to pursue a long range systematic program to validate ASVAB and enlistment standards against performance on the job. The Army Research Institute (ARI) conducted an initial pilot research project in the 193rd Infantry Brigade, Panama to determine the feasibility of validating enlistment standards

---

The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U. S. Army Research Institute or the Department of the Army.

against performance on the job. This report discusses some preliminary analyses of these data. Our preliminary analyses focuses on two areas. The first area deals with the definition of a "successful soldier", and the extent to which Commanding Officers and NCO's agreed on the qualities of a "successful soldier". The second area deals with the relations among ASVAB 6/7, SQTs and selected preliminary measures of job performance.

Agreement Between Officers and NCOs on the Importance of Rating Factors

A total of 526 first-term soldiers participated in the research. In addition, 26 Company Commanders and 26 First Sergeants, from the units representing the tested troops, along with a total of 203 first line supervisors (squad leaders) completed job performance ratings and rankings on these same first-term soldiers. Combat units representing the Infantry; and support units representing Military Intelligence, Military Police, Transportation, and Medical specialties were included in the study.

As part of a special data collection, ARI researchers investigated (1) the extent to which Company Commanders and their First Sergeants agree on what it means to be a successful or unsuccessful soldier, and (2) more importantly to find out which of these indicators are considered to be the most important to Commanders and First Sergeants presently in the field. Three types of data were collected from Company Commanders and First Sergeants to focus on this issue:

1) Each was asked to rate 49 factors which could be used to measure the quality of a soldier's job performance. Ratings were on a scale of importance as indicators of job performance.

2) Each was asked to independently choose the ten most successful and the ten least successful soldiers from among first-term troops in his company.

3) Each was asked to list several observable behaviors exhibited by these soldiers which would indicate why a soldier had been chosen as one of the most or least successful.

Average ratings of the 49 individual job performance factors are attached as Appendix A. Averages were calculated across all raters as well as separately for Company Commanders and First Sergeants. Note that it appears that First Sergeants generally tended to rate items as more important than Company Commanders. This probably indicates a different response set rather than true differences. However, it is interesting to note that this trend is reversed for ratings of Discipline Factors. First Sergeants generally considered these factors as less important than Company Commanders.

In Appendix B, the 49 factors are grouped into those generally considered 'Extremely Important', 'Very Important', 'Important', and 'Somewhat Important.' These groupings provide a broad overview of what is considered important or not important by command personnel of the 193rd Inf Bde, Panama for evaluating job performance.

Ratings of the 49 factors were also checked for agreement both within companies (Commanding Officer and First Sergeant) as well as across all respondents. The range of agreement between a CO and his 1SG varied from a low of -.08 to a high of .60. Of 23 companies eligible for this analysis, acceptable agreement between Commanders and First Sergeants occurred in 13 companies (.33+), in our judgment. The average level of agreement across all 23 companies responding is .28.

Company Commanders and First Sergeants were also asked to select their 10 most and 10 least successful soldiers and to list specific characteristics of accomplishments used in selecting them. Although responses were received in many content areas (e.g., job performance, troop responsibilities, relationship with others, respect for authority, motivation, and personal behavior), the major areas of agreement between Company Commanders and First Sergeants were in the content areas of job performance, troop responsibilities, and personal behavior/discipline. Table 3 summarizes the characteristics and accomplishments that were common to both the Commander's and First Sergeant's lists of the 10 most and 10 least successful soldiers.

## Preliminary Conclusions

These data provide some interesting suggestions, both for practical application as well as future research. It's interesting to note that the three factors identified as "extremely important" for indicating quality of job performance (Appendix B): (1) performs job satisfactorily; (2) accepts responsibility; and (3) resists authority, correspond to the major categories within which comments from Company Commanders and First Sergeants fall; that is, job performance, troop responsiblity, and discipline. Thus, these data serve to corroborate one another.

The moderate level of agreement among Company Commanders and First Sergeants suggests that while there is a kernel of agreement regarding what is meant by good job performance, there is room for improvement. This baseline level of agreement is sufficiently high, however, to cause us to be optimistic about the possible future utility of supervisory ratings of job performance. As indicated previously, it is felt that an organized effort to clarify these concepts among supervisory personnel would have a direct impact on improving job performance.

Interesting questions which arise from the data include:

1. How do military personnel arrive at a concept of good job performance?

2. Why are letters of appreciation not considered more important as indicators of job performance?

3. How does one operationalize concepts such as "accepts responsiblity" and "resists authority"?

## Relationship Among ASVAB 6/7, SQTs, and Selected Measures of Job Performance

The second phase of the preliminary analysis focused on the relations among ASVAB 6/7, SQTs, and selected measures of job performance. The central problem of validating enlistment standards and tests against actual job performance is that no measure of job performance exists in the Army. With the absence of such a measure we attempted to estimate the quality of an individual's performance by assessing various indicators that have some logical, although imperfect relationship to job performance. The indicators used included (1) Skill Qualification Test Scores, (2) Number of awards, (3) Number of additional military courses completed, (4) Number of letter of appreciation, (5) Number of Article 15's (6) Honor graduate status in training schools, and (7) Peer and supervisory ratings and rankings.

### Analysis

The preliminary analyses consisted of computing simple correlations among Aptitude Area Composite scores and job performance measures.

Correlations were computed for 11B's and 95B's separately, as well as for the total sample. Correlation matrices were not computed for samples having less than 50 soldiers. Thus, 63B's, 64C's and 91B's were not analyzed separately; however, these MOSs were included in the analysis of the total sample. Finally, correlation coefficients were computed to determine the agreement between platoon sergeants on first-term soldiers' job performance. For this analysis the platoon leader's and platoon sergeant's rankings were correlated.

## Results and Discussion

Two major issues were considered: (1) the meaning and quality of the various criterion measures, and (2) the relationships between scores on ASVAB 6/7 and criterion measures. The results of the analyses appear in Table 4 through 9.

Table 4 shows the extent of agreement between platoon leader and platoon sergeant rankings of first term soldiers in six platoons where sufficient data existed to perform this analysis. The average agreement was $\underline{r} = 0.78$. This result indicates that there is reasonable consensual agreement between platoon leaders and platoon sergeants with regard to which soldiers are "good" soldiers and which soldiers are "poor" soldiers.

Tables 5, 6, and 7 are identical except for the sample on which the correlations were calculated. Table 5 includes the total sample from Panama; Table 6 includes only soldiers in MOS 11B; Table 7 includes only soldiers in MOS 95B. There are a number of substantial and interesting relations found in Table 5. For example soldiers who perform better on SQTs also tend to have received more awards ($\underline{r} = .43$), and to have completed more additional military training courses

such as airborne school ($r$ = .34). It may be that these soldiers perform better on SQTs because they are better trained, or perhaps their higher scores on SQTs and enrollment in additional military training are both the result of a higher level of motivation. Note that subjective ranking tends to be correlated with behavior ($r$ = .27), such as SQT performance, which could be observed by platoon leaders. This suggests that these subjective rankings are not merely a reflection of popularity, but probably are grounded in actual performance as well. Similar results appear in Table 6 for the 11B MOS. For example, platoon leaders apparently are cognizant of Article 15's, and rank lower a soldier who has received them ($r$ = .33).

In all three Tables, a significant relation exists between peer and platoon leader rankings. It is difficult to determine, from these data, the basis on which peer rankings were given, however, their correlations with platoon leader rankings indicate that they too are more than the result of a popularity contest. The paucity of significant relations for the 95B sample (Table 7) is inexplicable. It may be a reflection of the smaller sample size, or perhaps a reflection of true differences between the 95B and 11B samples. Tables 8 and 9 contain the correlations of the various criterion measures with Aptitude Area Composite scores from ASVAB 6/7. Before discussing the content of the Tables a few points need to be addressed.

In 1980 an error in the calibration of ASVAB 6/7 scores was discovered. This error had the impact of substantially lowering enlistment standards. That is, recruits were enlisted who would not have qualified had the ASVAB 6/7 been calibrated correctly. Analyses reported here are based on a recalibration of ASVAB 6/7 to the correct level. In the course of the analyses, it was discovered that, of the soldiers tested in Panama, sixty-four 11Bs and thirty-one 95Bs would not have been qualified for reenlistment under the corrected calibration of ASVAB 6/7. The impact of this result is not addressed as part of this report.

The soldiers who participated in the data collection in Panama initially were qualified for enlistment on the basis of their ASVAB 6/7 scores. The measures and standard deviations for each Aptitude Area Composite in an unselected population are 100 and 20, respectively. Because entry into the 11B MOS is on the basis of scores on the CO Aptitude Area Composite, explicit selection has occurred on this variable. By explicit selection, we mean that recruits were classified into an MOS based on their score on a specific composite (a minimum score of 90 was required on CO for enlistment in 11B, and a minimum score of 100 was required on ST for enlistment in 95B). This results in a restricted range of scores for the selected group on the explicit selection variable.

The same restriction has occurred for 95B and ST. One indication of this is that CO for 11Bs and ST for 95Bs have the smallest standard deviation within each group. As a result, all correlations with CO for 11Bs and ST for 95Bs have been corrected for the effects of explicit selection on one variable by a formula provided in Lord and Novick (1968). Since the Aptitude Area Composite scores are correlated among themselves, correlations with other composites are also likely to be slightly depressed.

The descriptive information provided by the means and standard deviations in Tables 8 and 9 provide interesting insight. The standard deviations indicate that, as a group, soldiers in the 11B MOS have more widely variable ability as measured by the ASVAB 6/7 than soldiers in the 95B MOS. However, note that soldiers in the 95B MOS have a generally higher level of ability. In fact the mean difference on composites ranges from approximately 6 to 16 points. One interesting result is that on the basis of mean composite scores, the soldiers in the 95B MOS are more qualified to be in the Infantry than the 11Bs (96.34 vs. 90.21). The Army's current differential classification system is demonstrated here, in that even though the 95Bs would have qualified for entry into the 11B MOS training, they were also qualified for entry into 95B MOS training; which requires a higher Aptitude Area Composite Score on ST than 11B MOS requires in CO.

Turning next to the correlations themselves, the first result is the differing pattern of correlations for the 11B and 95B samples. Again, this could be either an artifact of sample size or a reflection of true differences. For both groups the correlations between SQTs and the ASVAB composites are large, ranging from $\underline{r}$ - .30 to $\underline{r}$ = .63. The validity of the Army's classification system is supported partially by the result that CO for 11Bs and ST for 95Bs show the highest correlation with SQTs, respectively. Further, even if all composites are corrected for restriction in range the above is still true. A final comparison between the 11B and 95B data shows that peer rank exhibits reasonably high correlations with ASVAB composites for 95Bs, but no correlations for 11B. The complete lack of correlation in the 11B sample at the very least indicates that the criterion measures used by enlisted personnel to rank on the job performance of peers differs from the 95B to 11B group.

While the remaining correlations are somewhat smaller they may be worth considering. The number of small but significant correlations between the Composites and awards, additional military courses, and Honor graduate status indicates that ASVAB is predictive of other indicators of job performance.

Finally, while these data seem to indicate a relationship between ASVAB scores and job performance, it is still too early to determine if it will be feasible to set enlistment standards based on job performance. Much work remains to be done in refining existing measures of performance or developing composites of existing measures of performance for occupational specialties within the broad occupational areas we have examined.

Table 1

Content - ASVAB Forms 5, 6, and 7

| Test | No. of Items | Time (Minutes) | Test Descriptions |
|---|---|---|---|
| General Information (GI) | 15 | 07 | A test on knowledge of geography, sports, history, automobiles. |
| Numerical Operations (NO) | 50 | 03 | A speed test of the four arithmetic operations- addition, subtraction, multiplication, division. |
| Attention to Detail (AD) | 30 | 05 | A test of clerical speed and accuracy by counting the number of "C"s embedded in a series of "O"s. Involves knowledge of word meaning. |
| Word Knowledge (WK)* | 30 | 10 | A test of knowledge of word meanings. |
| Arithmetic Reasoning (AR)* | 20 | 20 | A test of reasoning and arithmetic processes. |
| Space Perception (SP)* | 20 | 12 | A test which involves the selection of three dimensional figures which are formed by folding the pattern. |
| Mathematics Knowledge (MK) | 20 | 20 | A test of knowledge and skills in algebra, geometry, and fractions. |
| Electronic Information (EI) | 30 | 15 | A test of knowledge of elementary principles of electricity and electronics. |

Table 1 continued

| | | | |
|---|---|---|---|
| Mechanical Comprehension (MC) | 20 | 15 | A test involving mechanical principles such as gears, pulleys, and hydraulics. |
| General Science (GS) | 20 | 10 | A test involving knowledge of physical and biological sciences. |
| Shop Information (SI) | 20 | 08 | A test involving knowledge of shop procedures and the use of tools. |
| Automotive Information (AI) | 20 | 10 | A test involving knowledge of auto repairs and recognition of symptoms of various malfunctions. |
| TOTAL | 295 | 135 | |

* Scores on these three subtests are added together to provide AFQT scores.

Note: The Army Classification Inventory (87 items and about 20 minutes in time) is administered along with Form 6 and 7 as part of the operational testing procedure.

Table 2

Aptitude Area Composites (6 & 7)
and Prerequisite for Major Groups of Army MOS

| Area Aptitude Composite | ASVAB 6/7 Subtests | Military Occupational Specialties (MOS) |
|---|---|---|
| CO (Combat) | AR+SI+SP+AD+CC | Infantry, Armor, Combat Engineer |
| FA (Field Artillery) | AR+GI+MK+EI+CA | Field Cannon and Rocket Artillery |
| EL (Electronics Repair) | AR+EI+SI+MC+CE | Missile Repair, Air Defense Repair, Electronics Repair, Fixed Plant Communications Repair |
| OF (Operators & Food) | GI+AI+CA | Missile Crewmen, Air Defense Crewmen, Driver, Food Services |
| SC (Surveillance & Communications) | AR+WK+MC+SP | Target Acquisition and Combat Surveillance, Communications Operations |
| MM (Motor Maintenance) | ME+EI+SI+AI+CM | Mechanical and Aircraft Maintenance, Rails |
| GM (General Maintenance) | AR+GSB+MC+AI | Construction and Utilities, Chemical, Marine Petro. |
| CL (Clerical) | AR+WK+AD+CA | Administrative, Finance, Supply |
| ST (Skilled Technical) | AR+MK+GSB | Medical, Military Policeman, Intelligence, Data Processing, Air Control, Topography and Printing, Information and Audio Visual |
| GT (General Technical) | AR+WK | Not currently used for classification into a particular MOS |

Table 3

LIST OF CHARACTERISTICS AND ACCOMPLISHMENTS BY
COMPANY COMMANDERS AND FIRST SERGEANTS

| | Most Successful | Least Successful |
|---|---|---|
| JOB PERFORMANCE: | | |
| | Soldiers High on SQT | Not Prompt |
| | Soldier of the Month | Low SQT Scores |
| | Assignment To Duty Position Authorized at Higher Grade | Poor Job Performance |
| | Overall Duty Performance (In Garrison and Field) | |
| | Selection to Bde Machine Gun Team | |
| TROOP RESPONSIBILITIES | | |
| | Needs Little Supervision | Needs Constant Supervision |
| | Accepts Responsibilities | Does Not Accept Responsibilities |
| RESPECT FOR AUTHORITY | | |
| | | Attitude |
| | | Disrespect for Authority |
| MOTIVATION | | |
| | Initiative | Lacks Motivation |
| | Self Motivation | |
| PERSONAL BEHAVIOR/DISCIPLINE | | |
| | Has Good Uniform Appearance | Drug Abuse |
| | Has Good Moral and Personal Behavior | Article 15's |
| | | Counseling Statements |
| | | AWOL |
| | | Has Poor Uniform Appearance |
| | | Physical Condition |

Table 4

Agreement Between Platoon Leaders and Platoon Sergeants from Six Platoons on First-Term Soldiers' Job Performance

| Number of First-Term Soldiers Being Rated | Correlation between platoon leader rankings and platoon sergeants rankings on first-term soldiers job Performance. |
|---|---|
| 10 | .80 |
| 13 | .50 |
| 14 | .72 |
| 10 | .87 |
| 8 | .96 |
| 11 | .80 |

Table 5

(Total Panama Sample)

Means, Standard Deviations, and Correlation Coefficients of All Criterion Measures in a Sample of 526 First-Term Soldiers

| Variables | Mean | S.D. | Var. No. | Correlation Coefficients 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Awards | .14 | .38 | 1 | | | | | | | | | |
| Mil. Crs. | .24 | .55 | 2 | .56** | | | | | | | | |
| Civ. Crs. | .03 | .22 | 3 | .18** | | | | | | | | |
| SQT | 68.64 | 14.85 | 4 | .43** | .34** | .15* | | | | | | |
| Lett. App. | .81 | 1.98 | 5 | | | (237)<br>.15* | | | | | | |
| Hon. Grd. | .07 | .28 | 6 | | | (185)<br>.37** | | (181)<br>.60** | | | | |
| Art. 15 | .61 | .86 | 7 | | | | | | | | | |
| Peer Rnk. | 7.15 | 3.32 | 8 | | | | | | | | | |
| Plt. Ldr. Rnk. | 7.14 | 4.64 | 9 | | | | (64)<br>.27* | | (76)<br>.38** | (96)<br>.37** | (182)<br>.61** | |

NOTES:

- Blank cells ≥ .05
- * ≤ .05
- ** ≤ .01
- () = sample size
- Correlation coefficients were not corrected for restriction in range.

Table 6

Means, Standard Deviations and Correlation Coefficients of All Criterion Measures in a Sample of 294 First-Term Soldiers in MOS 11B

| Variables | Mean | S.D. | Var. No. | Correlation Coefficients 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Awards | .23 | .48 | 1 | | | | | | | | | |
| Mil. Crs. | .32 | .58 | 2 | (294) .63** | | | | | | | | |
| Civ. Crs. | .01 | .11 | 3 | (294) .13* | (287) .15* | (157) .20* | | | | | | |
| SQT | 72.94 | 12.98 | 4 | (157) .48** | (157) .37** | (154) .18* | | | | | | |
| Lett. App. | .64 | 1.85 | 5 | | | (127) .34** | (79) .27* | | | | | |
| Hon. Grad. | .03 | .175 | 6 | | | | (127) .70** | | | | | |
| Art. 15 | .62 | .88 | 7 | | | | | | | | | |
| Peer Rnk. | 6.59 | 3.11 | 8 | | | | | | | | | |
| Plt. Ldr. Rnk. | 5.92 | 3.68 | 9 | | | | | | | (69) .33** | (98) .61** | |

NOTES:

- Blank cells $> .05$
- * $\leq .05$
- ** $\leq .01$
- () = sample size
- Correlation coefficients were not corrected for restriction in range.

Table 7

Means, Standard Deviations, and Correlation Coefficients of All Criterion Measures in a Sample of 126 First-Term Soldiers in MOS 95B

| Variables | Mean | S.D. | Var. No. | Correlation Coefficients 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Awards | .008 | .09 | 1 | | | | | | | | | |
| Mil. Crs. | .05 | .25 | 2 | | | | | | | | | |
| Civ. Crs. | .008 | .09 | 3 | | | | | | | | | |
| SQT | 62.21 | 11.29 | 4 | | | | | | | | | |
| Lett. App. | .71 | 1.27 | 5 | | | | | | | | | |
| Hon. Grad. | .17 | .46 | 6 | | | | | | | | | |
| Art. 15 | .67 | .97 | 7 | | | | | | | | | |
| Peer Rnk. | 8.98 | 3.71 | 8 | | | | | | | | | |
| Plt. Ldr. Rnk. | 9.59 | 5.23 | 9 | | | | | | | | (53) .55** | |

NOTES:

- Blank cells ≤ .05
- * ≤ .05
- ** ≤ .01
- () = sample size
- Correlation coefficients were not corrected for restriction in range.

Table 8

Means, Standard Deviations, and Correlations Coefficients Between ASVAB Composites Scores and Performance Measures in a Sample of 294 First-Term Soldiers in MOS 11B

| | | | Correlation Coefficients | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Variables | Mean | S.D. | SQT | Awards | Mil. Crs. | Civ. Crs. | Lett. App. | Hon. Grd. | Art. 15 | Peer Rnk. | Plt. Rnk. |
| CO (Combat)[a] | 90.21 | 12.95 | .61** | .30** | .15* | .21* | .21* | .24* | | | |
| FA (Field Artillery) | 89.00 | 15.94 | .46** | .21** | .13* | .18* | | | | | -.24* |
| EL (Electronic Repair) | 89.35 | 14.72 | .45** | .20** | | .17** | | .21* | | | -.22* |
| OF (Operators & Food) | 88.75 | 16.91 | .41** | .23** | .13* | .12* | | | | | |
| SC (Surveill & Comm) | 87.78 | 14.91 | .42** | .23** | .15* | .23* | | .23* | | | |
| MM (mechanic Maint) | 89.00 | 15.28 | .46** | .20** | | .17** | | .23* | | | |
| GM (General Maint) | 87.30 | 15.65 | .46** | .25** | | .20** | | .29** | | | |
| CL (CLerical) | 87.38 | 15.51 | .30** | .16** | | .16** | | | | | |
| ST (Skilled Tech) | 88.10 | 15.94 | .37** | .20** | | .24** | | .29** | | | -.22* |
| GT (General Tech) | 86.85 | 16.07 | .33** | .18** | | .20** | | | | | |
| N[b] = | | | 149 | 283 | 283 | 283 | | 124 | | | 105 |

NOTES:

- Blank cells ≥ .05
- * ≥ .05
- ** ≥ .01

- [a] Correlation coefficients were not corrected for restrictions in range.
- [b] N's do not total 294 because of missing data.

Table 9

Means, Standard Deviations, and Correlation Coefficients of ASVAB Composite Scores and Performance Measures in a Sample of 126 First-Term Soldiers in MOS 95B

| Variables | Mean | S.D. | Correlation Coefficients SQT | Award | Mil. Crs. | Civ. Crs. | Lett. App. | Hon. Grd. | Art. 15 | Peer Rnk. | Plt. Rnk. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CO (Combat) | 96.34 | 14.84 | | | | | | | | -.24* | |
| FA (Field Artillery) | 103.40 | 12.03 | .37** | | | | | | | -.31** | |
| EL (Electronic Repair) | 97.74 | 13.99 | .31* | | | | | | | -.26* | |
| OF (Operators & Food) | 97.48 | 14.77 | .30* | | | | | | | -.42** | |
| SC (Surveill & Comm) | 97.51 | 13.87 | .31* | | | | | | | | |
| MM (Mechanic Maint) | 97.22 | 13.97 | .34* | | | | | | | -.31** | |
| GM (General Maint) | 98.15 | 12.86 | .44** | | | | | | | -.28* | |
| CL (Clerical) | 100.98 | 13.22 | | | | | | | | -.33** | |
| ST (Skilled Tech)[a] | 104.45 | 10.53 | .63** | | .24* | | | | | | |
| GT (General Tech) | 100.90 | 12.72 | .37** | | | | | | | -.26* | |
| N[b] = | | | 51 | | | | | | | 51 | |

NOTES:

- Blank cells > .05
- * ≤ .05
- ** ≤ .01
- [a] Correlation coefficients were corrected for restriction in range
- [b] N's do not total 126 because of missing data

## REFERENCES


Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Massachusetts: Addison-Wesley, 1968, pp. 140-144.

Maier, M. A., & Grafton, F. C., Aptitude Composites for ASVAB 8, 9, and 10. ARI Technical Paper 1308. Washington, DC: Army Research Institute for the Behavioral and Social Sciences, May 1981.

Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics). Aptitude Testing of Recruits (Unclassified), July 1980.

Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics). Aptitude Testing of Recruits (Unclassified) September 1980.

c. List some specific characteristics or problems which you have observed with these ten least successful first tour personnel which led you to select them. For example: One fell asleep while on guard duty.

_____

_____

_____

_____

_____

_____

Step 3. Please rate each of the following factors as to how critical you feel each factor is in indicating the quality of an individual's overall performance.

Importance Scale

5 = Extremely Important
4 = Very Important
3 = Important
2 = Somewhat Important
1 = Not Important At All

| CO | Overall Means Importance Scale | 1SG | Performance On-The-Job |
|---|---|---|---|
| 2.71 | 2.88 | 3.04 | Current Grade |
| 3.50 | 3.75 | 4.00 | Assignment to Duty Position Authorized at Higher Grade |
| 1.92 | 2.10 | 2.29 | Time from E1 to E2 |
| 2.17 | 2.33 | 2.50 | Time from E2 to E3 |
| 2.79 | 2.83 | 2.88 | Time from E3 to E4 |
| 3.50 | 3.50 | 3.50 | Time from E4 to E5 |
| 3.83 | 3.83 | 3.83 | SQT Score |
| 4.17 | 4.21 | 4.25 | Reduction(s) for Inefficiency |
| 4.00 | 4.10 | 4.21 | Letter(s) Counseling for Inefficiency |

## Importance Scale

5 = Extremely Important
4 = Very Important
3 = Important
2 = Somewhat Important
1 = Not Important At All

## Other

| Importance Scale | | | |
|---|---|---|---|
| | | | **Presence of SMOS** |
| 3.00 | 3.19 | 3.38 | Acquired OJT |
| 3.46 | 3.54 | 3.63 | Graduate MOS Awarding School |
| | | | **Awards and Decorations** |
| 3.92 | 3.81 | 3.71 | Meritorious Service Medal |
| 3.63 | 3.67 | 3.71 | Army Commendation Medal (Merit) |
| 2.33 | 2.88 | 3.42 | Good Conduct Award |
| 3.00 | 3.38 | 3.75 | Letters of Recognition (Appreciation) |

## Discipline

| | | | |
|---|---|---|---|
| | | | **Court Martial(s)** |
| 4.67 | 4.23 | 3.79 | General |
| 4.79 | 4.48 | 4.17 | Special |
| 4.54 | 4.04 | 3.54 | Summary |
| 3.88 | 3.83 | 3.79 | Non-Judicial Punishments (Art. 15) |
| 3.96 | 3.94 | 3.92 | Days AWOL |
| 4.58 | 4.50 | 4.42 | Times AWOL |

## Importance Scale

5 = Extremely Important
4 = Very Important
3 = Important
2 = Somewhat Important
1 = Not Important At All

## Administrative

| Importance Scale | | | |
|---|---|---|---|
| | | | **Highest Security Clearance Attained** |
| 3.67 | 3.69 | 3.71 | Top Secret |
| 3.21 | 3.52 | 3.83 | Secret |
| 2.83 | 3.08 | 3.33 | Confidential |
| 3.63 | 3.85 | 4.08 | Letter(s) of Indebtedness |
| 2.83 | 3.23 | 3.63 | Auto Accidents Held at Fault |
| 4.38 | 4.44 | 4.50 | Bar to Reenlistment |
| | | | **Self-Improvement Efforts** |
| 4.08 | 4.23 | 4.38 | Educational Level Attained |
| 3.88 | 4.17 | 4.46 | Enrolled in GED Program |
| 3.79 | 4.02 | 4.26 | Enrolled in Non-Resident Military Education Program |
| 3.54 | 3.81 | 4.08 | Special Training |
| 3.58 | 3.83 | 4.08 | Schools Attended |

## Importance Scale

5 = Extremely Important
4 = Very Important
3 = Important
2 = Somewhat Important
1 = Not Important At All

## Miscellaneous

| Importance Scale | | | |
|---|---|---|---|
| 3.96 | 3.94 | 3.92 | Physical Profile |
| 3.09 | 3.34 | 3.58 | Peer Ratings (if they were systematically collected) |

Please list any additional variables we might have omitted

___ _____

___ _____

___ _____

___ _____

___ _____

GO ON TO THE NEXT PAGE.

## IMPORTANCE OF PERFORMANCE RATINGS

How important to you are each of the following behaviors in evaluating the enlistee.

### Importance Scale

5 = Extremely Important
4 = Very Important
3 = Important
2 = Somewhat Important
1 = Not Important At All

| | | | |
|---|---|---|---|
| 4.58 | 4.67 | 4.75 | 1. Performs job satisfactorily. |
| 3.71 | 3.69 | 3.67 | 2. Has difficulty communicating effectively. |
| 4.50 | 4.48 | 4.46 | 3. Sets a good example for co-workers. |
| 4.71 | 4.63 | 4.54 | 4. Accepts responsibility. |
| 4.63 | 4.54 | 4.46 | 5. Resists authority. |
| 3.83 | 4.15 | 4.46 | 6. Shows good judgement in expressing opinions. |
| 4.67 | 4.48 | 4.29 | 7. Needs constant supervision. |
| 3.96 | 4.06 | 4.17 | 8. Conforms to Army appearance standards. |
| 3.79 | 3.73 | 3.67 | 9. Has been a disciplinary problem within the last six months. |
| 4.08 | 4.23 | 4.38 | 10. Works well with others. |
| 3.88 | 3.94 | 4.00 | 11. Possesses capacity to acquire knowledge/group concepts. |
| 4.08 | 4.10 | 4.13 | 12. Demonstrates appropriate knowledge and expertise in assigned tasks. |
| 3.88 | 4.00 | 4.13 | 13. Maintains appropriate level of physical fitness. |
| 4.46 | 4.46 | 4.46 | 14. Performs well under physical and mental stress. |
| 4.08 | 4.10 | 4.13 | 15. Was absent without leave (AWOL). |

THANK YOU FOR YOUR COOPERATION.

## Appendix B

**Factors identified as 'Extremely Important' for indicating the quality of an individual's overall job performance.**

1. **Performs job satisfactorily.**
2. **Accepts responsibility.**
3. **Resists authority.**

Factors identified as 'Very Important' for indicating the quality of an individual's overall job performance.

1. Times AWOL
2. Special Court Martial
3. Sets a good example for co-workers.
4. Needs constant supervision.
5. Perfcrms well under physical and mental stress.
6. Bar to reenlistment.
7. General Court Martial
8. Educational level attained.
9. Works well with others.
10. Reductions for inefficiency.
11. Enrolled in GED program.
12. Shows good judgment in expressing opinions.
13. Letters counseling for inefficiency.
14. Demonstrates appropriate knowledge and expertise in assigned tasks.
15. Was AWOL.
16. Conforms to Army appearance standards.
17. Summary Court Martial.
18. Enrolled in non-resident Military Education Program.
19. Maintains appropriate level of physical fitness.
20. Days AWOL.
21. Possesses capacity to acquire knowledge/group concepts.
22. Physical profile.
23. Letters of indebtedness.
24. Schools attended.

AD P001375

# An Examination of the Group Differences Aspect of the Construct Validity of the Organizational Assessment Package

Major Lawrence O. Short
Lt Col David A. Wilkerson

Directorate of Research and Analysis
Leadership and Management Development Center
Maxwell AFB AL 36116

The Organizational Assessment Package (OAP) is currently being revised by Leadership and Management Development Center researchers. A part of this revision is reexamining the validity of the survey instrument in the light of data and experience gained from two years of field use. Specifically the study concerned the group differences aspect of construct validity. Support is offered for OAP construct validity as differences by OAP factor across major functional area groupings seemed consistent and strong. The differences also held across logical groupings of factors. Results were more equivocal, however, in testing hypotheses concerning specific pairs comparisons within factors. Discussion is offered regarding results, and conclusions and recommendations for additional validity research are presented.

PREVIOUS PAGE IS BLANK

# An Examination of the Group Differences Aspect of the Construct Validity of the Organizational Assessment Package[1]

The validity of a survey instrument is often defined as the extent to which the instrument measures what it intends to measure (Carmines and Zeller, 1979). Since validity is also often considered the most essential feature of an instrument, its adequate and accurate determination generally receives a high priority in instrument development, evaluation and, if necessary, redesign.

Despite its importance, however, validity is often complex to deal with in that it is a multi-faceted concept (Stanley and Hopkins, 1972), which requires continual monitoring and updating throughout the life of an instrument (Nunnally and Durham, 1975). The type of validity studied must also be matched to the purposes and goals of the validity study. This point is well made by Cronbach and Meehl (1955) who differentiate among types of validity by showing that each type involves a slightly different emphasis on the criterion measure. One particular type of validity, construct validity, is most appropriate for use when no definite, concrete, specific, or fully valid criterion measure is available. This type of validity seemed most appropriate in relation to the Organizational Assessment Package (OAP), the survey instrument which is the subject of this study.

Considering Cronbach and Meehl's (1955) subtypes of construct validity, some evidence relating to OAP construct validity is already available. Short and Hamilton (1981) presented evidence essentially supporting the internal structure and change-over-occasions aspects of construct validity. Prior to this study OAP factors were expected to be internally consistent as assessed by a Cronbach's alpha procedure and were expected to retain significant test-retest correlations across both five week and six month time intervals. It was further expected that the six month correlations would be lower than those for the five week interval because of both the longer interval and the necessity that factors be sensitive to actual organizational changes rather than being artificially rigid. These expectations were confirmed with the exception of some of the two or three item factors that seemed to lack either internal consistency or stability. A factor by factor revision of both the OAP and its supporting hardware and software systems after two years of field use and experience is currently in progress, and the authors recommended strengthening or deleting these factors as a part of this revision.

The purpose of this study was, therefore, twofold. After having examined two of Cronbach and Meehl's construct validity subtypes, the first purpose was to examine a third subtype, group differences. The second purpose was to continue gathering information to point the way toward the revision of the survey instrument.

---

[1] Because of space limitations, some details of the study were discussed more briefly than would otherwise be the case. Additional information is available from the authors upon request.

## Method

### Instrumentation

The OAP is a 109 question survey designed jointly by the Air Force Human Resources Laboratory and the Leadership and Management Development Center to aid the LMDC in its mission to: (a) provide management consulting services to Air Force commanders upon request, (b) to provide leadership and management training, and (c) to conduct research on Air Force systemic issues with information within the accumulated data base. Supporting developmental research for the OAP is provided in Hendrix and Halverson (1979a; 1979b) and Hendrix (1979).

Administration of the survey is the first step in the consultation process. The survey is given to a stratified random sample of the organization to which LMDC has been invited. The results of the survey are an important feature in the assessment of the organization. The results are handled in a confidential manner between LMDC and the client. After approximately five to six weeks for analysis, feedback of data is then provided to commanders and supervisors within the organization.

When specific problems are encountered, a consultant and supervisor develop a management action plan designed to resolve the problem at that level of the organization. Within six to nine months, the consulting team returns to readminister the survey instrument as a means to help assess the impact of the consulting process.

The data from each OAP administration effort are stored in a cumulative data base currently containing over 100,000 records for research purposes. These data are aggregated by work group codes developed for this instrument. The data may be recalled by demographics such as personnel category, age, sex, Air Force Specialty Code (AFSC), pay grade, time in service, and educational level. Through factor analysis, the 93 attitudinal items are combined into factors which cover job content, job interferences, and various types of supervisory and organizational areas. OAP factor numbers and names are presented in Figure 1.

### Sample and Procedure

In order to study group differences, officer, enlisted and civilian responses to the pre-intervention OAP were drawn from the data base and aggregated by major functional area groupings. These groupings were wing staff, group staff, resources, maintenance, operations, medical, missiles, communications (abbreviated "comm" in the tables to save space) and unique, a category containing people in organizations with orientations that were scientific, technical, research and development or special missions. The number of persons in each functional area grouping by factor is shown in Table 1. The number of people in each area may vary by factor because of individual response patterns to the instrument. Capital letters are used in Table 1 for number and standard deviation to denote the fact that these figures represent the total number of responses available in the data base by factor and functional area grouping.

| Factor Number | Factor Name |
|---|---|
| 800 | Skill Variety |
| 801 | Task Identity |
| 802 | Task Significance |
| 804 | Job Feedback |
| 805 | Work Support |
| 806 | Need for Enrichment (Job Desires) |
| 810 | Job Performance Goals |
| 811 | Pride |
| 812 | Task Characteristics |
| 813 | Task Autonomy |
| 814 | Work Repetition |
| 816 | Desired Repetitive Easy Tasks |
| 817 | Advancement/Recognition |
| 818 | Management-Supervision |
| 819 | Supervisory Communications Climate |
| 820 | Organizational Communications Climate |
| 821 | Perceived Productivity (Work Group Effectiveness) |
| 822 | Job Satisfaction |
| 823 | *Job Related Training* |
| 824 | General Organizational Climate |

Figure 1. OAP Factor Numbers and Names

Hypotheses tested were stated at three levels. First, it was anticipated that all OAP factors would be sensitive enough that between group variance would exceed within group or error variance resulting in significant F-ratios for each factor across functional area groupings. Corresponding null hypotheses of no differences among functional areas were stated for every factor. These hypotheses were tested by use of one-way analyses of variance by factor across functional area groupings. Because of large Ns, a significance level of $p < .001$ was desirable to help assure practically significant as well as statistically significant differences. A nominal significance level of $p < .05$ was required for all omnibus F-test and follow-on multiple comparison procedures.

Second, based in part on the work of Conlon (1980) and in part on consultants' observations of task, climate, productivity and leadership patterns Air Force wide, it was expected that factors dealing with perceptions of task would show the widest variation across functional areas. Similarly, it was expected that perceptions dealing with leadership function and style would be most consistent and show the least variation across functional area groupings. Since perceptions of climate as defined in the OAP may be related to perceptions of task, it was expected that climate factors would show variations

second only to task. Finally, it was expected that perceived productivity, dependent to a degree on all the other three, would show more variation than leadership factors but less variation than task or climate factors. These logical factor groupings, the factors that compose them and the hypothesized direction of differences are summarized as follows:

| Perceptions of Task (OAP Factors 806, 812, 813) | > | Perceptions of Climate (OAP Factors 820, 824) | > | Perceptions of Productivity (OAP Factor 821) | > | Perceptions of Leadership (OAP Factors 818, 819) |
|---|---|---|---|---|---|---|

The corresponding "null hypothesis" of no differences among the four groups was tested only by observation of F-ratios for each of the factors. For convenience, ratios were averaged where appropriate. While not a stringent statistical procedure, this process did help provide a clearer presentation of results.

Finally, specific pairwise differences between groups and direction of differences by factor across functional area groupings were hypothesized where information was available upon which to base such hypotheses. All pairs comparison contrasts were performed among the functional area subgroups from each factor following the omnibus F-tests using a Student-Newman-Keuls procedure ($p < .05$). Space limitations prohibit the complete display of the results of testing over 60 null hypotheses. A general statement of results is presented in the next section, however.

## Results

By reference to Table 1, all null hypotheses across functional area groupings by factor were rejected as all omnibus F-tests were significant beyond the .001 level. (It should be noted that Factors 800, 801, 802 and 804 were omitted from the table. These are not true factors, but are composed of items already contained in Factor 812.) Additionally, by reference to Table 1, the second "null hypothesis" of no differences among perceptions of task, climate, productivity and leadership across functional area groups was also rejected. Differences in observed F-ratios in the hypothesized direction did appear. To simplify seeing this relationship, Table 2 is extracted from Table 1 and presents only observed F-ratios and their "averages" for comparison of magnitude and directionality.

Finally, testing of the specific group differences following the F-tests showed mixed results. Hypotheses were made taking both significance of difference and direction into account. Considering both, less than half of the specific corresponding null hypotheses were rejected. In some cases, results were in the expected direction but non-significant, while other functional area groups simply did not behave as expected.

Table 1

OAP Factor Means, Standard Deviations and Analysis of Variance Results by Functional Area Subgroups

| OAP Factors | | Functional Area Subgroups | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Wing | Group | Resources | Maintenance | Operations | Medical | Missiles | Comm | Unique | Total Sample | F-Ratio |
| 805 | N | 3426 | 18254 | 7709 | 12604 | 3360 | 4147 | 1417 | 1556 | 13366 | 65839 | 56.86 |
| | SD | 1.21 | 1.18 | 1.18 | 1.12 | 1.12 | 1.16 | 1.10 | 1.09 | 1.15 | 1.16 | (p<.001) |
| | $\bar{x}$ | 4.45 | 4.42 | 4.59 | 4.61 | 4.48 | 4.42 | 4.77 | 4.70 | 4.58 | 4.53 | |
| 806 | N | 3493 | 18582 | 7794 | 12746 | 3467 | 4168 | 1448 | 1585 | 13664 | 66947 | 164.77 |
| | SD | 1.32 | 1.41 | 1.29 | 1.29 | 1.09 | 1.16 | 1.17 | 1.18 | 1.17 | 1.29 | (p<.001) |
| | $\bar{x}$ | 5.58 | 5.35 | 5.47 | 5.35 | 5.77 | 5.82 | 5.61 | 5.55 | 5.73 | 5.51 | |
| 810 | N | 3487 | 18584 | 7828 | 12760 | 3457 | 4158 | 1451 | 1584 | 13585 | 66894 | 41.24 |
| | SD | 1.09 | 1.08 | 1.00 | .98 | .96 | .99 | .99 | .98 | 1.07 | 1.04 | (p<.001) |
| | $\bar{x}$ | 4.75 | 4.66 | 4.78 | 4.72 | 4.86 | 4.88 | 4.73 | 4.80 | 4.62 | 4.71 | |
| 811 | N | 3593 | 18896 | 7945 | 12872 | 3507 | 4230 | 1468 | 1612 | 13792 | 67915 | 70.63 |
| | SD | 1.71 | 1.82 | 1.70 | 1.71 | 1.58 | 1.59 | 1.64 | 1.61 | 1.69 | 1.73 | (p<.001) |
| | $\bar{x}$ | 5.01 | 4.79 | 4.86 | 4.73 | 5.21 | 5.27 | 4.86 | 5.05 | 4.97 | 4.89 | |
| 812 | N | 3440 | 18095 | 7754 | 12633 | 3435 | 4136 | 1434 | 1562 | 13525 | 66014 | 128.53 |
| | SD | 1.18 | 1.20 | 1.02 | 1.01 | .97 | .99 | 1.03 | .98 | 1.06 | 1.09 | (p<.001) |
| | $\bar{x}$ | 5.11 | 4.86 | 5.04 | 4.97 | 5.19 | 5.36 | 4.96 | 5.12 | 5.08 | 5.02 | |
| 813 | N | 3525 | 18584 | 7848 | 12694 | 3461 | 4175 | 1440 | 1582 | 13656 | 66965 | 163.46 |
| | SD | 1.52 | 1.57 | 1.41 | 1.43 | 1.34 | 1.44 | 1.42 | 1.34 | 1.47 | 1.49 | (p<.001) |
| | $\bar{x}$ | 4.34 | 3.88 | 4.16 | 3.65 | 3.95 | 4.18 | 3.83 | 3.96 | 4.14 | 3.97 | |
| 814 | N | 3573 | 18896 | 7932 | 12856 | 3498 | 4218 | 1456 | 1588 | 13764 | 67781 | 130.05 |
| | SD | 1.50 | 1.46 | 1.43 | 1.39 | 1.36 | 1.35 | 1.46 | 1.47 | 1.49 | 1.45 | (p<.001) |
| | $\bar{x}$ | 4.81 | 5.10 | 5.02 | 5.03 | 4.85 | 5.23 | 4.76 | 4.73 | 4.67 | 4.95 | |

Table 1, Cont.

Functional Area Subgroups

| OAP Factors | | Wing | Group | Resources | Maintenance | Operations | Medical | Missiles | Comm | Unique | Total Sample | F-Ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 816 | N | 3512 | 18627 | 7784 | 12708 | 3421 | 4139 | 1439 | 1567 | 13553 | 66750 | 148.29 |
| | SD | 1.48 | 1.51 | 1.41 | 1.39 | 1.26 | 1.39 | 1.29 | 1.33 | 1.34 | 1.43 | (p<.001) |
| | $\bar{x}$ | 3.14 | 3.32 | 3.16 | 3.18 | 2.73 | 3.05 | 2.90 | 3.06 | 2.85 | 3.11 | |
| 817 | N | 3398 | 18186 | 7662 | 12567 | 3420 | 4013 | 1433 | 1556 | 13273 | 65508 | 24.67 |
| | SD | 1.34 | 1.32 | 1.30 | 1.21 | 1.15 | 1.27 | 1.21 | 1.22 | 1.30 | 1.28 | (p<.001) |
| | $\bar{x}$ | 4.11 | 4.04 | 4.17 | 4.06 | 4.24 | 4.13 | 4.32 | 4.29 | 4.10 | 4.10 | |
| 818 | N | 3409 | 18317 | 7732 | 12594 | 3377 | 4076 | 1413 | 1562 | 13324 | 65804 | 34.52 |
| | SD | 1.63 | 1.65 | 1.63 | 1.57 | 1.39 | 1.62 | 1.45 | 1.53 | 1.55 | 1.59 | (p<.001) |
| | $\bar{x}$ | 4.92 | 4.92 | 4.92 | 4.78 | 5.18 | 4.87 | 5.01 | 5.01 | 5.04 | 4.93 | |
| 819 | N | 3331 | 18123 | 7606 | 12530 | 3286 | 4013 | 1414 | 1547 | 13105 | 64955 | 37.10 |
| | SD | 1.69 | 1.70 | 1.68 | 1.61 | 1.49 | 1.65 | 1.52 | 1.59 | 1.61 | 1.64 | (p<.001) |
| | $\bar{x}$ | 4.48 | 4.52 | 4.53 | 4.34 | 4.74 | 4.48 | 4.65 | 4.67 | 4.63 | 4.52 | |
| 820 | N | 3324 | 18031 | 7574 | 12454 | 3388 | 4023 | 1422 | 1542 | 14267 | 66025 | 89.37 |
| | SD | 1.47 | 1.47 | 1.40 | 1.33 | 1.27 | 1.35 | 1.26 | 1.27 | 1.41 | 1.41 | (p<.001) |
| | $\bar{x}$ | 4.53 | 4.35 | 4.58 | 4.27 | 4.82 | 4.45 | 4.76 | 4.62 | 4.41 | 4.43 | |
| 821 | N | 3437 | 18314 | 7779 | 12695 | 3413 | 4103 | 1445 | 1558 | 13450 | 66194 | 66.68 |
| | SD | 1.32 | 1.39 | 1.26 | 1.24 | 1.08 | 1.24 | 1.14 | 1.26 | 1.26 | 1.29 | (p<.001) |
| | $\bar{x}$ | 5.50 | 5.35 | 5.58 | 5.45 | 5.80 | 5.55 | 5.65 | 5.46 | 5.56 | 5.94 | |
| 822 | N | 3101 | 16829 | 6894 | 11385 | 3137 | 3750 | 1347 | 1407 | 12193 | 60043 | 102.68 |
| | SD | 1.34 | 1.39 | 1.27 | 1.29 | 1.23 | 1.16 | 1.24 | 1.17 | 1.26 | 1.31 | (p<.001) |
| | $\bar{x}$ | 5.09 | 4.88 | 5.07 | 4.74 | 4.90 | 5.28 | 4.89 | 5.08 | 5.07 | 4.96 | |
| 823 | N | 3096 | 17228 | 7294 | 12188 | 2955 | 3605 | 1325 | 1486 | 11936 | 61113 | 39.82 |
| | SD | 1.71 | 1.68 | 1.66 | 1.61 | 1.50 | 1.59 | 1.48 | 1.58 | 1.64 | 1.64 | (p<.001) |
| | $\bar{x}$ | 4.31 | 4.34 | 4.42 | 4.33 | 4.78 | 4.63 | 4.62 | 4.53 | 4.41 | 4.41 | |
| 824 | N | 3358 | 17993 | 7596 | 12405 | 3392 | 4040 | 1419 | 1559 | 14477 | 66239 | 129.91 |
| | SD | 1.52 | 1.54 | 1.50 | 1.42 | 1.32 | 1.41 | 1.34 | 1.35 | 1.45 | 1.48 | (p<.001) |
| | $\bar{x}$ | 4.73 | 4.41 | 4.57 | 4.27 | 5.02 | 4.48 | 4.87 | 4.72 | 4.55 | 4.50 | |

Table 2

Observed F-Ratios for Perceptions of Task, Climate, Productivity and Leadership*

| Task | Climate | Productivity | Leadership |
|---|---|---|---|
| 164.77 (806) | 89.37 (820) | 67.68 (821) | 31.52 (818) |
| 128.53 (812) | 129.91 (824) | | 37.10 (819) |
| 163.46 (813) | | | |
| 152.25 ($\bar{x}$) > | 109.64 ($\bar{x}$) > | 67.68 ($\bar{x}$) > | 34.31 ($\bar{x}$) |

* Each factor number is in parenthesis following its observed F-ratio.

## Discussion, Conclusions and Recommendations

By way of discussion, results of the omnibus F-tests by factor across functional area groups were consistent with expectations. Also, looking at predictions based on consulting experience and prior OAP research, differences in the expected directions did exist across functional area groups by groups of factors representing perceptions of task, climate, productivity, and leadership. Concerning specific pairwise comparisons, however, results were more equivocal and warrant some further consideration.

Three possible reasons for the pairwise comparison results seem possible. First, it may not be reasonable to expect a survey instrument to predict down to such a level of specificity. Other research of this type stops at the omnibus F-level without hypothesizing follow up comparisons (see, for example, Sims, Szilagyi, and Keller, 1976; Turney and Cohen, 1976). Even within highly diversified functional area groupings, results of this study also showed consistent and stable differences at the omnibus F-level.

Second, the locations contained in the functional area groupings may be a problem. Some groups contain people in a wide variety of locations, while other groups do not. In fact, some of the groups with the smallest numbers may represent data from only a very few locations. It is possible, then, that perceptions measured are actually situation specific rather than functional area specific. No selection method was employed to establish the groups; they were taken intact from the OAP data base. The result may be groups that don't show expected results because people in those few locations may score higher or lower on certain factors than would be the case if a wider variety of locations were available in that particular functional area group.

Finally, it may be that the OAP factors lack the construct validity to support these specific hypotheses. More to the point, it seems that the two or three item OAP factors may lack the validity necessary to support specific hypotheses. Reference to factors containing six to eight items shows results much more stable and predictable than was the case with two and three item

factors. This is consistent with the findings of Short and Hamilton (1981), who concluded that these smaller factors often showed problems with internal consistency, stability or both. These findings, of course, support the necessity to revise, delete or strengthen these few factors.

In the final analysis, it seems likely that all three reasons played a part in observed results. It does seem possible, however, to make such specific hypotheses and to expect an instrument to make corresponding discriminations. This belief forms a major part of the revision goal--to sharpen an already good instrument to make it even more precise and efficient.

It is reasonable to conclude, then, that the construct validity of the OAP continues to be demonstrated. More work remains to be done, however, on the subject. For example, factor stability is a major issue in determining construct validity (Anastasi, 1976). Factor analyses have been run across these same functional area groupings and also across demographic categories of personnel category (officer, enlisted, civilian), gender and ethnic group. While some change in factor structure is again indicated, the stability and consistency of this most recent factor solution is striking. Factor congruency coefficients (Harman, 1967) are being calculated and should be published in early 1982. In addition, a study using the multitrait-multimethod approach to convergent and discriminant validity (Campbell and Fiske, 1957) seems appropriate and is currently being planned.

As emphasized in the Standards for Educational and Psychological Tests (APA, 1974), "Evidence of construct validity is not found in a single study; rather judgements of construct validity are based upon an accumulation of research results" (p. 30). This statement is also true of the current study. Nevertheless, reevaluation research regarding the OAP continues to provide solid evidence for the validity of most factors as well as blueprints for the revisions of the few that can be made stronger, more specific, and more efficient.

## References

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Anastasi, A. Psychological testing. (4th Ed.) New York: Macmillan, 1976.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Carmines, E. G., & Zeller, R. A. Reliability and validity assessment. Beverly Hills, CA: Sage, 1979.

Conlon, E. J. Investigations of behavioral consultation in the Air Force. In USAF Summer Faculty Research Program Research Reports, Vol 1. Washington, D.C.: The SCEEE Press, 1980. Pp. 14-1 to 14-75.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Harman, H. H. Modern factor analysis. Chicago: University of Chicago Press, 1967.

Hendrix, W. H. Organizational assessment indices of effectiveness (AFHRL-TR-79-46). Brooks AFB, TX: Air Force Human Resources Laboratory, 1979.

Hendrix, W. H., & Halverson, V. B. Organizational survey assessment package for Air Force organizations (AFHRL-TR-78-93). Brooks AFB, TX: Air Force Human Resources Laboratory, 1979. (a)

Hendrix, W. H., & Halverson, V. B. Situational factor identification in Air Force organizations (AFHRL-TR-79-10). Brooks AFB, TX: Air Force Human Resources Laboratory, 1979. (b)

Nunally, J. C., & Durham, R. L. Validity, reliability, and special problems of measurement in evaluation research. In E. L. Struening and M. Guttentag (Eds.), Handbook of evaluation research, Vol 1. Beverly Hills, CA: Sage, 1975.

Short, L. O., & Hamilton, K. L. An examination of the reliability of the Organizational Assessment Package (LMDC-TR-81-2). Maxwell AFB, AL: Leadership and Management Development Center, 1981.

Sims, H. P., Jr., Szilagyi, A. D., & Keller, R. T. The measurement of job characteristics. Academy of Management Journal, 1976, 19, 195-212.

Stanley, J. C., & Hopkins, K. D. Educational and psychological measurement and evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Turney, J. R., & Cohen, S. L. The development of a work environment questionnaire for the identification of organizational problem areas in specific Army work settings (Technical Paper 275). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1976.

AD P001376

# ATTRITION: CAUSALITY, EXPLANATION, AND LEVEL OF ANALYSIS

Guy L. Siebold
US Army Research Institute for the Behavioral & Social Sciences

SUMMARY: A substantial amount of research has been and is being conducted on attrition. The time has come to be more rigorous in the conceptualization and terminology used in this research. The term "attrition" for example has been loosely used. Therefore, the author suggests a basic generic definition which distinguishes it from terms like premature separation. Similarly, "cause of attrition" has been loosely used to describe certain demographic variables which only explain variance and to describe reasons for separation which only represent marks on documents or question responses. The amounts and rates of attrition are aggregate variables. Causal analysis requires these aggregate variables to be investigated within smaller, more meaningful personnel groupings. The author suggests that proper causal analysis and the development of successful counter-attrition programs and procedures require investigations at the individual, organizational, and manpower levels. For example, premature separations may be considered a symptom of causal disorders at the organizational level. Also, manpower attrition policy acts as a constraint with resource consumption impacts on training, discipline, and other arenas. To be effective, cross-national research must pay attention to the foregoing. Finally, the author suggests more effort be directed toward research on building primary group relationships as a long term and pervasive counter-attrition strategy.

## INTRODUCTION

The rate of military attrition is higher than had been anticipated for the AVF. The cost, likewise, has been heavy (Comptroller General, 1980). In recognition of the high rate and consequent costs, the military has supported an increased amount of research in attrition over the last few years (cf. Sinaiko, 1977). Much has been learned from the research about AVF patterns of attrition, about individual and organizational variable correlates, and about many issues surrounding premature separation such as how it is perceived by unit managers (Goodstadt and Romanczuk, 1980) and how it affects the separated individuals (Dodd et al., 1975). Some counter-attrition pilot research projects have also been completed with promising results (e.g., Randolph et al., 1980).

Because each prematurely separated service member (SM) represents a substantial organizational and financial loss, special emphasis has been given to the development of methods to screen out recruit candidates who present a high risk of not completing their contractual tour of duty (e.g., Lockman and Lurie, 1980; Albert, 1980; Seeley et al., 1978; Sands, 1977). While this research is still in progress, its current models would still reject too many SMs who would actually finish their tours and accept too many that would not, considering present manpower pools. However, the screening models do identify high risk categories

of recruit candidates who can be monitored closely for needed assistance.

Since the AVF attrition research is rapidly approaching a quasi-paradigmatic stage, it is important to denote where the research needs more rigor and to fill in gaps. In part through literature reviews, scientists have already started this task (e.g., Goodstadt and Yedlin, 1979; Fox, 1979; Hand et al., 1977). This paper is an effort to continue that enterprise, particularly through suggesting refinements in terminology and theoretical conceptualization. Some suggestions for cross-national research and counter-attrition strategies are also included.

## BASIC ATTRITION TERMINOLOGY

Some of the basic terms associated with attrition research need to be standardized and used less loosely. Relevant dictionary (Webster's New Collegiate) definitions of attrition are "3: the act of weakening or exhausting by constant harassment or abuse 4: a reduction (as in personnel) chiefly as a result of resignation, retirement, or death." The DOD Directive #1415.7 definition of attrition is "separation prior to completion of the contractual active-duty obligation." These definitions as well as those used or implied by various researchers are different. They therefore reduce the conceptual equivalency of research and, hence, its comparability and cumulation. For research purposes, the following attrition terms and generic definitions are suggested to help clarify and standardize their usage:

1. ATTRITION--the reduction in the number of personnel of a specified category through separation. (the process)
2. AMOUNT OF ATTRITION--the number of personnel of a specified category lost through separation. (the count)
3. RATE OF ATTRITION--the number of personnel of a specified category lost through separation within a specified time period. (the count/time ratio)
4. PERCENTAGE OF ATTRITION--the number of personnel of a specified category lost through separation compared to the total population of which the specified category is a primary part. (the count/population ratio)

Note that, among other figures, one can compute relative amounts, rates, and percentages; cumulative amounts and percentages; and average rates. For the sake of brevity, these computations are not explicated.

Given the foregoing set of generic attrition terms and definitions, the scientist need only precisely specify the category of personnel with which he is working to effectively perform his computations or present his results. The scientist should be careful, however, to distinguish the use of time for computing rates (e.g., the number of first term enlisted personnel in a given cohort who have not completed their contractual active-duty obligation separated per year) from the use of time to characterize a specified category of personnel (e.g., the number of first term enlisted personnel in a given cohort who separated in their first year of service). The use of the above terminology will lessen the incidence of loose phrases such as "high attrition" (amount, rate, or percentage? personnel category?) and "greater than expected attrition"

(amount, rate, or percentage? personnel category?). There is a need for more rigor in attrition research terms in general and in the specification of the personnel categories.

The amounts, percentages, and rates of attrition are aggregate figures which result from a count of the occurrences of an event, the separation. As such, the figures are dependent upon the group within which the event can occur, the specified category. This fact cannot be overemphasized. The causal factors and policy implications for different specified categories of personnel are likely to be quite different. Therefore it is crucial that the attrition figures are computed for an appropriate and meaningful level of aggregation, i.e., the right specified category.

## CAUSALITY

It is a mistake to equate reasons for premature separation, measured either by marks on discharge documents or by verbal or written question responses, with the causes of separation. Each premature separation is an individual, an organizational, and a manpower event. Causation factors exist at all of these levels, and a given separation results from a particularized blend of factors from all of these levels. Presumably, there is a generalized multi-leveled blend of causal factors within each specified category of personnel. The reduction in the amount of premature separations in each category of personnel then should result from a blend of counter-attrition actions and programs addressing factors at all of these levels.

At the manpower level of analysis, one must note that attrition in the personnel category "first term enlisted personnel who have not completed their contractual active-duty obligation" (i.e., premature separation) occurs because of the policy decision to permit that attrition. There would be no serious peacetime attrition issue without that policy decision. This fact, too, cannot be overemphasized. However, if attrition in this category were not permitted or subject to low ceiling levels, the underlying problems would simply be shifted to other arenas such as training, discipline, and individual/organizational adjustment. Manpower issues are highly interrelated. The obvious point in terms of causation is that the policy set level, whether a ceiling, goal, or unrestrained, is a substantial influence on the amount and rate of attrition in this first term enlisted category of personnel.

One example of a manpower policy influencing the amount of attrition is the policy to permit easy separation of low performers or those with adjustment problems at an early stage in their tour (especially during training). This policy is an extension of pre-enlistment screening. The rationale is that it is better to separate problem SMs early while the training and organizational investments are still low. It is assumed those with substantial early difficulties are likely to require separation later. Thus it was thought that by permitting easy early separations, there would be a reduction in the number of separations in a cohort occurring later. Presumably the reduction in later attrition would enhance morale, readiness, and the personnel management function. However, preliminary data seem to indicate that this reduction in later attrition

may not occur, at least not proportionately. Instead, the total cumulative amount of attrition during the first tour of a given enlisted cohort may be increased. Apparently early tour attrition is to a degree independent of later first tour attrition. More research, of course, needs to be done to verify this pattern and assess the impact of the early separation and other policies on attrition.

At the organizational level of analysis, one finds many factors directly or indirectly causing premature separation. For example, the rates and percentages of first term enlisted attrition may be affected by the occupational assignment system (Thomason, 1980), thematic training models (Siebold, 1979), beliefs of unit commanders (Goodstadt and Romanczuk, 1980), and/or work and work group structuring (Cathcart et al., 1978). Other examples are given in the previously cited literature reviews. Yet much research remains to be done on causal factors at this level not only in terms of their influence on attrition but in terms of their impact on, among other things, productivity, unit integration, and readiness.

The term "organizational level" is used herein to encompass all those systems, structures, processes, and personnel that fall hierarchically between manpower policy and the individual SM. This large grouping, in fact, needs to be placed in a coherent conceptual framework in order to determine on which of these organizational parts or features attrition research has been conducted and on which no attrition information is available. Then an assessment must be made as to the needed direction of future research within the framework. It is time to discover and modify those organizational features and parts which cause high rates of attrition. Research must progress beyond the previous emphases on screening out high risk personnel and on teaching problem SMs to adjust to the military.

At the individual level of analysis, one must be careful to distinguish the investigation of individual level causal variables (i.e., characteristics of individuals and their relationships) from the study of why a given individual SM or set of individual SMs were separated. This distinction is crucial because a study of the latter is not attrition research but separation research and is not likely to produce useful information to reduce attrition but only, if possible, to explain given separations (cf., definitions on page 2).

To explicate this distinction, let us consider what a separation is. A separation is a formal legal disentanglement between a military service and a SM with concurrent changes in the legal rights and obligations of the service and the SM. Is is consummated by the authoritative signing and issuing of certain documents. The event has an impact at multiple levels. At the manpower level, it is the loss of a SM to the military for which a replacement must be obtained. At the organizational level, the separation means there has been a break in the relationship structure. Another SM must perform the separated SM's jobs. At the individual level, rights, obligations, and expectations have been irrevocably altered between the discharged SM and others in the organization. For the separated SM, there are substantial changes, for example, in social circumstances, lifestyle, identity structures, and personal outlook.

Let us consider what causes a separation. There is an official "cause" or reason used to describe the basis for the separation. The official reason serves to explain and legitimate the separation but is not necessarily the only, the main, or the real reason(s) in the minds of the parties involved. But a reason is not a cause. The acts surrounding the creation and issuance of discharge documents effecting the separation are caused by a prior act, the authoritative order to do so. This order was given by some commander resulting from the fact that either the commander, the about to be discharged SM, or both wanted the separation to take place (cf., Goldman, 1970). In short, the separation was caused by the fact that someone with the power to effectuate it wanted the separation to occur. Separation research then must of necessity focus on why someone with the power to accomplish the separation wanted it to take place.

To establish what caused a "want" or to explain it is quite difficult. Wants not only are the result of a multitude of influences but are also capable of being explained within numerous inconsistent theoretical frameworks. The latter can range from theories positing a conscious internal locus of control (e.g., rational decision-making) to those positing either an unconscious (e.g., Freudian psychoanalytic theory) or external (e.g., "the devil made me do it") locus of control (cf., Mayhew, 1981). Unfortunately the length constraints on this paper prohibit further discussion of the problems associated with establishing causality or a satisfactory explanation for wants. Suffice it to say that the investigation of the causes or reasons for an individual separation is not likely to produce a solid parsimonious means to reduce the amount of attrition.

At the individual level of analysis, one is better off to conduct research through variables and approaches which are NOT mentalistic (i.e., not wants, attitudes, or reasons). And there is no shortage of promising non-mentalistic variables. Further, many of these individual level variables relate directly to organizational level variables. For example, at the organizational level one can investigate the amount and rate of unit attrition associated with the amount and rate of unit personnel turbulence, the amount of unit overtime, the degree to which a unit is over or understrength, the number of small fill MOS in a unit, the task variety and complexity in a unit, and a unit's physical isolation. At the individual level of analysis, one can investigate the amount and rates of attrition associated with individuals experiencing varying amounts of turbulence, pulling differing amounts of overtime, working at job sites with a shortage or surplus of personnel, occupying small or large fill MOS within a unit, performing typical tasks of less or greater variety and complexity, and operating alone or in groups of various size and integration.

The number of individual level variables causing or associated with attrition is extensive. Again, like organizational level variables, they need to be placed in a coherent conceptual framework. Accurate measures need to be developed (an ounce of measurement is worth a pound of analysis). Meaningful, testable hypotheses need to be deduced. Further, causality needs to be assessed, and that is not an easy job. But without a determination of causality, effective counter-attrition programs will be hard to create.

The amount, rate, and percentage of attrition are variables of a peculiar class for which the determination of causality is elusive. Variables in this class relate to numerous social phenomena (e.g., crime, births and deaths, church attendance). These variables have two important characteristics: 1) they pertain to the occurrence of a specified event within a specified group and 2) the precise causal linkages between the independent variables and the amounts, rates, and percentages of the occurring event are often indirect, probabilistic, and/or unknown. Causality for the class is non-determinative (Bunge, 1963). Consequently, it is difficult to understand exactly why there are changes in the rates and amounts. Similarly it may be difficult to predict whether the event will or will not occur to a specific individual. Yet simultaneously one may be able, with a given set of independent variables, to predict the amount, rate, or percentage with great accuracy or to explain a large percentage of the variance. From the independent variables, one must infer the underlying causal structure. From the inferred causal structure, one can develop actions and programs designed to reduce the number of occurrences of the event, i.e., the separation in the case of attrition. Note that these independent variables will not be status (demographic) or mentalistic variables.

## CROSS-NATIONAL ATTRITION RESEARCH

Cross-national (and cross-service) attrition research is becoming increasingly common. Presently it suffers not only from the general looseness in conceptualization and terminology mentioned above but from the limited comparability of different national (or service) terms, personnel categories, policies, and organizational structures. A major task facing the international research community is the establishment of ways to transform these items so that cross-national comparisons can be made. Similarly, standardized data bases need to be established so that the appropriate data will exist to make the comparisons necessary for meaningful cross-national research.

Cross-national (and cross-service) research can be extremely valuable. The country specific descriptions of the amounts and rates of attrition for common personnel categories can provide a better idea of the potential range of variation. The different national analyses can provide new independent information on the apparent and underlying causal structures. Universal patterns may be found. The investigation of attrition in just one country, in other words, is a limited endeavor simply because of the shortage of comparable observations.

One of the biggest potential benefits of cross-national research is in the design and evaluation of counter-attrition research. With different countries trying different sets of counter-attrition actions and programs, a wider variety of them can be tested than would be possible in one country alone. For the most promising actions and programs, equivalent experiments can be set up in multiple countries simultaneously. The results of these experiments will be more generalizable, and the effects of the actions and programs can be more thoroughly assessed. Stated briefly, cross-national (and cross-service) research has the potential to increase substantially the depth and breadth of the assessment of counter-attrition actions and programs within a given time frame.

## COUNTER-ATTRITION PROGRAMS

To drastically reduce first term enlisted attrition, a blend of actions and programs will be needed at the manpower policy, organizational, and individual SM levels. The optimal blend is not yet known and is not going to be easy to arrive at. In all likelihood, it will only come over time through a series of successive approximations. Also, the optimal blend may well change as time passes. In any case, the attrition problem exists now, and researchers must address it without waiting for all the answers to come in.

Comment on possible policy modifications is not within the scope of this paper. Obviously, military pay increases, the imposition of attrition ceilings, and the intensive recruiting of less attrition prone personnel would have an impact on the number of enlisted premature separations. But these actions and others would have coincident disadvantages, and all must be weighed within the total policy context (e.g., see Comptroller General, 1980). The research community, however, can and has begun to develop and assess counter-attrition actions and programs at the organizational and individual levels. It is to these levels that the rest of the paper is addressed.

Several candidate research topics were cited as worthy of support at the (Leesburg, Virginia) Conference on First Term Enlisted Attrition (Sinaiko, 1977). Research on some of these topics has been conducted. New useful information has been acquired, and some promising counter-attrition projects have been carried out and evaluated. Yet several conclusions about attrition research are apparent. First, there has been no development of a coherent conceptual framework/plan for the research. Second, not much research involving experimental policies has been undertaken. Third, the majority of the research is concentrated on the front end of the service tour (e.g., on applicant screening and initial SM expectations). Fourth, research has most often focused on ways for the new SM to adjust to military life (e.g., on coping skills). And fifth, research has been conducted usually on a specific narrow topic on a small SM group at a specific stage in the tour with the short term measurement of results. In other words, past and much current research present only small pieces of the puzzle. One of the main set of tasks to be accomplished is to put these pieces of the puzzle together, determine which pieces are still missing, and discard pieces which do not belong.

For every incorrect assumption about SMs and attrition, there is probably an equal and opposite incorrect assumption. For example, underlying a given suggested counter-attrition program may be the Father Flanagan assumption (there is no such thing as a bad boy), the silk purse assumption (you can't make one out of a sow's ear), or the bad apple assumption (a bad one spoils the bunch). Unit commanders make separation decisions based on certain assumptions and self-imposed rules (Goodstadt and Romanczuk, 1980). In developing an array of counter-attrition actions and programs, assumptions and decision rules need to be made explicit. Further, actions and programs based solely on one of these assumptions (e.g., on blame the "victim", blame the military, or blame society rationales) should be avoided.

What counter-attrition actions and programs ought to be developed? The question, of course, is best answered through hindsight. One should note, however, that during training SMs are, for the most part, simply confined aggregates of individuals. Except in long duration formal schooling, there are not many strong bonds between SMs in training. Thus during-training and post-training actions and programs to reduce attrition may have to be quite different although integrated. For post-training attrition, two general approaches are suggested.

The first general approach to reduce attrition is to build strong, positive primary group relationships between each SM, his co-workers, supervisors, and family (local or back home). This approach would go beyond the building of peer groups or the implementation of a buddy system (see Sinaiko, 1977). In essence, it is a re-emphasis on the communal or institutional nature of military life at the small group level. Research has shown that the isolated, alienated SM is more likely to have difficulties leading to premature separation (e.g., Shils, 1977; Georgoulakis and Bank, 1979). The full integration of a SM with the personal significant others around him can provide a number of benefits. It can help overcome a multitude of major and minor difficulties faced by the first term enlisted SM. In addition to being a long term, pervasive inhibitor of premature separation, strong primary group relationships may well enable the SM to withstand stress on the battlefield or during periods with heavy workloads.

The second general approach to reduce attrition is to keep SMs more informed about what is going on with their units, the service, their mission, and the international political situation. The purpose of this approach is to prevent a sense of anomie or meaninglessness from permeating the lives of the SMs. The typical SM needs to feel he is in on things. Further, before a SM can develop a commitment to something, he needs to know to what he is making a commitment. For example, research has consistently shown that those who do not fully expect to finish their full tour (i.e., low commitment) are less likely to finish it. Some scientists may wish to consider this approach as a matter of building goal congruence. In essence, though, it is a re-emphasis on military service as a service or calling rather than just an occupation or job (cf. Siebold, 1979). Like the first approach, this second approach may act as a long term, pervasive inhibitor of premature separation.

## REFERENCES


Albert, Walter G.
1980 Predicting Involuntary Separation of Enlisted Personnel. Technical Report 79-58, Air Force Human Resources Laboratory, Brooks AFB, TX, January.

Bunge, Mario
1963 Causality. Cleveland: World Publishing.

Cathcart, John S., Robert D. Goddard, and Stuart A. Youngblood
1978 Perceived Job Design Constructs: Reliability and Validity. Technical Report 7, Center for Management and Organizational Research, University of South Carolina, Columbia, SC, September.

Comptroller General of the United States
1980 Attrition in the Military--An Issue Needing Management Attention. Report FPCD-80-10, United States General Accounting Office, Washington, DC, February 20.

Dodd, Mary I., Ronald G. Bauer, Thomas J. Miller, and David R. Segal
1975 The Post-Service Adjustment of U.S. Army Trainee Discharge Program Dischargees. Technical Report BEETO-75-044, The Bendix Corporation, Ann Arbor, MI, November.

Fox, Alexander J.
1979 A Comprehensive Investigation of First-Term Enlisted Army Attrition. Draft Report, Military Strength Programs Division, ODCSPER, HQ, Department of the Army, Washington, DC, May.

Georgoulakis, James M. and Robert L. Bank
1979 "Social Factors and Perceived Problems as Indicators of Success in Basic Combat Training: Part Two." Military Medicine October:685-6.

Goldman, Alvin I.
1970 A Theory of Human Action. Englewood Cliffs, NJ: Prentice-Hall.

Goodstadt, Barry E. and Alan P. Romanczuk
1980 Research on Determinants of Unit-Level Attrition Decision Making: I. Unit Commander Survey Findings. Technical Report, Westat, Inc., Rockville, MD, April 17.

Goodstadt, Barry E. and Nancy C. Yedlin
1979 A Review of State-of-the-Art Research on Military Attrition: Implications for Policy and for Future Research and Development. Technical Report, Advanced Research Resources Organization, Washington, DC, June.

Hand, Herbert H., Rodger W. Griffeth, and William H. Mobley
1977 Military Enlistment, Reenlistment and Withdrawal Research: A Critical Review of the Literature. Technical Report 3, Center for Management and Organizational Research, University of South Carolina, Columbia, SC, December.

Lockman, Robert F. and Philip M. Lurie
1980 A New Look at Success Chances of Recruits Entering the Navy (SCREEN). Report CRC 425, Center for Naval Analyses, Alexandria, VA, February.

Mayhew, Bruce H.
1981 "Structuralism Versus Individualism: Part II, Ideological and Other Obfuscations." Social Forces 59 (March):627-48.

Randolph, W. Alan, Stuart A. Youngblood, Bruce M. Meglino, James Laughlin, and Angelo S. DeNisi
1980 Sources of Concern and Coping Skills Employed by U.S. Army Trainees During Basic Training. Technical Report 1, Center for Management and Organizational Research, University of South Carolina, Columbia, SC, May.

Sands, William A.
1977 Screening Male Applicants for Navy Enlistment. Technical Report 77-34, Navy Personnel Research and Development Center, San Diego, CA, June.

Seeley, Leonard C., Theodore Rosen, and Kenneth Stroad
1978 Early Development of the Military Aptitude Predictor (MAP). Technical Paper 288, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA, AD A052953, March.

Shils, Edward
1977 "Profiles of a Military Deserter." Armed Forces and Society 3 (Number 3):427-31.

Siebold, Guy L.
1979 "Trends in Army Training: Are They Consistent with the Industrial Model of the Army?" Paper presented at the 74th Annual Meeting of the American Sociological Association, Boston, MA, August 27-31.

Sinaiko, H. Wallace (ed.)
1977 First Term Enlisted Attrition--Volume I: Papers. Technical Report 3, Smithsonian Institution, Washington, DC, June.
First Term Enlisted Attrition--Volume II: Summary. Technical Report 4, Smithsonian Institution, Washington, DC, August.

Thomason, James S.
1980 Rating Assingments to Enhance Retention. Report CRC 426, Center for Naval Analyses, Alexandria, VA, February.

AD P001377

# An Evaluation of the Air Force Airman Retraining Program

Mary J. Skinner

Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

An evaluation of the progress and adjustment of retrained airmen in their second Air Force occupational specialties was conducted. The specific study objectives were to assess (a) retrainees' accommodation to the specialty change as reflected by their attitudes, motivation, and interpersonal relationships; (b) the impact of retraining on career progress indicators such as promotion and reenlistment; and (c) retrainees' performance, skills, and ability in the new job. Data were obtained from self-report questionnaires completed by over 13,000 retrained enlistees and 11,500 military supervisors. Preliminary analysis of selected survey items were in the form of item descriptive statistics and t-tests of mean differences in ratings assigned to retrainees versus non-retrainees and to pre- versus post-retraining specialties. Results provided strong support for the conclusion that retrained enlistees had made a smooth and successful transition between military occupations. In addition to the interim findings, research plans for a comprehensive and in-depth examination of the survey data were reported.

## Introduction

The Air Force and other military services, unlike employers in the private sector, must maintain balanced manning in enlisted career fields within the constraints imposed by a closed personnel system. Apprentice-level positions requiring enlisted personnel with minimal or basic skills are filled from the same manpower pool as available to private industry. However, the Air Force must utilize personnel who are already integrated into the military force to staff jobs above the entry level. To realign career field manpower overages and shortages, particularly in journeyman positions, military managers rely heavily on the capability to retrain enlisted personnel. Guided by the policies and procedures of the Airman Retraining Program (AFR 39-4, 1979), managers process retraining actions which change enlistees from one occupational specialty to another either within the same career field or into a different career field. This retraining capability permits the Air Force to staff positions and balance career field/manning more efficiently within the limitations of a closed personnel system.

During the past ten years between 10,000 and 15,000 enlistees annually have changed occupational specialties under the auspices of the Airman Retraining Program. Air Force data files indicate that among enlistees who have served 20 years approximately 80% have changed specialties at some time during their military careers. The volume of retraining witnessed in recent years is not expected to decrease. Factors such as attrition, changes in mission and weapon systems, and fluctuations in the recruiting pool will continue to influence military manpower needs and, in turn, will make personnel retraining necessary.

Until late 1977, no systematic research on retraining had been undertaken despite the integral role played by retraining in the total Air Force personnel management system. Recognizing the need for a comprehensive and objective evaluation of the Airman Retraining Program, Air Force retraining managers submitted a Request for Personnel Research (RPR 77-12) to the Manpower and Personnel Division, Air Force Human Resources Laboratory (AFHRL). An overview of the research stream supporting this RPR is presented below. Research undertaken to address certain issues using available historical personnel records will be briefly described while the paper will focus primarily on the study methodology, planned analysis, and preliminary results of a second major data source -- a survey of retrained enlisted personnel and their supervisors.

The evaluation of the Airman Retraining Program is a multiphase effort aimed at tracking the progress of retrainees in their new military specialties. Specific objectives are to study (a) performance in technical training; (b) job and career progress in terms of skill level upgrading, on-the-job performance, promotion, and separation/retention rates; (c) retraining success for different sub-groups of retrainees, including voluntary and involuntary retrainees, and as a function of personal and background characteristics; and (d) the accommodation of retrainees to their new occupations.

Data needed to address much of the first two research objectives are contained in historical personnel files maintained on a routine basis and periodically updated by AFHRL. The technical school performance of retrainees is being evaluated through comparisons with new recruits (non-retrainees).

Analyses of academic performance and school attrition rates for approximately 20,000 retrainees and 230,000 non-retrainees attending over 250 schools are in progress. Statistical analysis will examine the relationship between school performance criteria and retraining status and aptitude, as well as experience factors such as amount of military service and type of background experience acquired prior to retraining. A second major effort examines several job and military career advancement criteria for retrainees relative to Air Force averages. The time required to upgrade job skill levels in the new specialty and to achieve grade promotions by about 35,000 retrainees are being retrieved for a three-year period for comparison with the remainder of the force (approximately 480,000 enlistees). Similar comparisons of separation/ retention criteria are being conducted. This series of analyses also provides a cursory examination of retrainee on-the-job performance based on an annual official appraisal by supervisors. To supplement these data and to address the latter two research objectives, an analysis of data collected by survey is underway.

The survey was designed to provide information needed for a thorough evaluation of the adjustment and progress of retrainees in their second miltary specialties. The preliminary results from this survey will be reported. The main issues addressed are the retrainees' accommodation to the new specialty, the impact of retraining, and retrainees' job skill and performance. In addition to the interim results, research plans for an in-depth examination of the survey data will be described.

## Method

### Questionnaire Development

Survey topics and items were identified from an extensive review of pertinent literature on prior and ongoing research and regulatory guidance on the retraining system. Inputs were solicited from knowledgeable military personnel, including former retrainees and managers tasked with retraining policy formulation or program administration. Two preliminary questionnaires, one for retrainees and one for supervisors, were developed for pretesting. The questionnaires were administered in face-to-face, individual interviews conducted at eight bases in five major Air Force commands. A total of 213 retrainees and 158 supervisors of retrainees participated in the pretest. A descriptive summary of the pretest information was prepared as a reference document to guide revisions and modifications to the instruments.

The final instruments consisted of two standardized questionnaires with 85 items on the retrainee form and 66 items on the form for each retrainee's respective supervisor. Multiple response choice and forced choice forms of closed questions were used. Response options were presented in category rating scale form. Most ratings scales had five options. Appropriate scale adjectives (e.g., poor, fair, average, good, excellent) were selected for each question or set of questions. Both questionnaires called for the retrainee and supervisor to provide self-ratings on a variety of personal, attitudinal, and opinion questions. The instruments contained common sets of questions to permit, for example, comparisons of the retrainee's self-rating and his/her supervisor's appraisal. The supervisors provided overall assessments of retrained and non-retrained personnel as well as ratings on a specific retrainee. In addition, the supervisor selected and identified one of his/her non-retrained enlisted subordinates to rate on the same set of questions.

Thus, supervisors' appraisals of individual retrainees and non-retrainees were also available for comparative analysis. Survey materials for each respondent consisted of a retrainee or supervisor questionnaire, a separate answer sheet for recording responses, and a letter of instructions. Answer sheets were precoded with an identification number to link retrainee-supervisor pairs for later analysis.

In the questionnaires, items were organized by specific topic or issue with descriptive headings for each section. These major topic areas, with a description of the types of items in each, have been combined into the following general categories for ease of discussion:

Background Information. Personal and demographic information which were either not available or needed to verify the historical data file contents were collected. Items on the retrainee questionnaire included identification of the previous (before retraining) and current (after retraining) Air Force Specialty (AFS) and amount of experience in each specialty. Background items on the supervisor questionnaires included total amount of experience in the current AFS and length of time as a supervisor.

Impact of Retraining/Adjustment to Retraining Specialty. Ratings of the impact of retraining on chances for promotion and propensity to reenlist were collected from both supervisors and retrainees. Indicators of accommodation and adjustment to the new occupation were also collected. Retrainees rated their motivation, interpersonal relationships, and other attitudinal measures in their previous and current AFS. The corresponding questions for supervisors required their appraisals of the retrainee, a non-retrainee, and retrainees and non-retrainees in general.

Job Skill, Ability, and Performance Assessment. Measures of quality, quantity, difficulty level of work performance, job knowledge and supervisory skills, and amount of time and help required to upgrade skill level were collected. Retrainees were asked to rate themselves on these measures and supervisors to rate the retrainee, a non-retrainee, and retrained and non-retrained enlistees as a group.

Contributors to Successful and Unsuccessful Retraining. Retrainees' and supervisors' perceptions of transferability of technical, administrative, and supervisory job skill and general military knowledge between AFSs or AFS groupings were obtained. Retrainees were asked to rate how well technical school and on-the-job training prepared them to perform the job in their new specialty. Supervisors provided judgments of the best time to retrain enlisted personnel during their military career. Measures provided by retrainees on other factors which may influence the difficulty or ease of job changes were the type of retraining action (voluntary or involuntary), reason for retraining (to reenlist, to collect a monetary incentive, to learn skills helpful in finding a future civilian job) and attitudes toward the experience (desire to leave previous or enter current AFS).

Retraining Policy and Program Administration. Items on the retrainee questionnaire solicited their perceptions of retraining procedures including the expeditiousness of application processing, accuracy of counselor's advice, currency of job vacancy information, and amount and quality of publicity on

retraining opportunities. Opinions about involuntary retraining and associated policies and their impact on enlistees' morale, motivation, and productivity were also collected.

## Sampling Strategy and Subjects

Three major sampling objectives were identified during design of the sample selection plan. First, the sample size had to be sufficiently large to allow later study of different retraining movements between Air Force career fields. A career field is a grouping of related AFSs involving similar skills and knowledge. Characteristics of career fields and their similiarities and dissimilarities to other career fields were expected to influence the ease of transition during retraining. Second, it was important that the sampling strategy consider the voluntary or involuntary (selective) nature of the retraining action. Retraining managers identified voluntary versus involuntary retrainee comparisons as a priority interest. Since the rules to qualify an action as involuntary were very stringent, few selective retrainees were identified in the retraining history files. Consequently, oversampling on the factor was deemed necessary. Finally, in order to compare retraining ease and patterns of job changes between traditional and nontraditional career fields for males and females, consideration had to be given to gender in the sampling plan. Few female retrainees were available. Thus, it was necessary to select them in numbers disproportionate to their representation in the retrainee population.

Subjects were enlisted personnel who had retrained between July 1973 and August 1977 and between April 1978 and August 1979. Records to identify airmen who retrained during the seven-month interval from September 1977 through March 1978 were not available. About 20,000 subjects were targeted for selection from 1733 different between-career-field movements using a stratified, quasi-random sampling plan. For each type of career field change, a maximum of 66 cases was randomly selected such that up to 50% were female and/or involuntary retrainees and the remainder were male volunteers. A total of 20,968 retrainees were selected using these procedures. Later, administrative constraints were encountered which limited data collection to 200 retrainees maximum per base of assignment. Questionnaires were administered to a final sample of 18,065 retrainees and to their first-line supervisors.

## Survey Administration

Survey materials were batched for bulk mailing to 123 stateside and overseas Air Force bases. Survey Control Officers at these bases distributed survey materials to individual respondents and performed survey monitoring including follow-up contacts with nonrespondents.

## Analysis

After data were reduced and edited, several items from the retrainee and supervisor questionnaires were selected for preliminary data analysis. The rating scale options of these items were coded from 1 through 5. Descriptive statistics for the items in the form of frequencies, percentages, means and standard deviations were obtained on cases with valid data entries. Tests of significance were conducted with Student $t$ statistics using the Bonferroni technique to control Type I error ( $\alpha$ ) per family of comparisons among survey items (Miller, 1966) at the .01 probability level.

## Results

### Data Processing and Editing

Both completed and uncompleted answer sheets were processed by optical scanning into computer data files. Answer sheets which were returned from nonrespondents with all response entries missing were discarded from further processing. An additional 68 supervisors and 25 retrainee cases were deleted for failure to respond to certain critical items. Critical items were those for which skips were unpermissible or data unretrievable by reference to historical data files on supervisors and retrainees. After data reduction and clean-up, less than three percent of the cases had missing/invalid data on any one critical item. A sample of 13,070 retrainees and of 11,549 supervisors remained for preliminary analysis.

### Background Information

Distributions of background information items described occupational and retraining experience for retrainees and supervisors. Over 75% of the retrainees reported that the present specialty change was their first retraining experience since entering the Air Force. An additional 16% had retrained twice and about 6% had retrained three or more times. The retrainees most commonly reported that they had been assigned to their previous specialty three to four years before retraining (34%). In general, there was an almost equal split on amount of experience in the pre-retraining specialty at the four-year point with about half having four or fewer years and the remaining 50% having five or more years. At the time the retraining survey was administered, about 44% of the retrainees had acquired up to two years of experience in their new specialty. Another 50% had been in their current occupation for three to six years since retraining.

Supervisors who completed the questionnaire had more experience in their current specialty than their retrained subordinates. About 73% had been assigned to the occupation for five or more years, and nearly 30% had 15 or more years of experience. Most had functioned in a supervisory capacity in the specialty for over three years (70%). About 55% of the supervisors had been assigned to their present specialty since entering the Air Force. The others (45%) had experienced a change in Air Force specialty and had, like their subordinates, retrained into their current occupation.

### Adjustment to Retraining Specialty

Attitudinal, motivational, and interpersonal relationship measures were selected for the preliminary assessment of retrainees' adaptation to the retraining experience and accommodation to their new occupations. On these items retrainees used a rating scale with poor to excellent options to describe their experiences and perceptions in their prior and current specialties. Item means and standard deviations for the previous and current specialties are shown at Table 1. The results of t-tests (for correlated samples) to identify items with mean ratings which differed significantly for the two specialties are also tabled. Mean response values for the pre- and post-retraining occupations indicated that the retrainees described their experiences as average or good. The measures of motivation and interpersonal relations were assigned somewhat higher ratings than job attitude measures in both specialties. All statistical contrasts between perception of the

previous and current specialty were found to be significant ($p \leq .01$). That is, in the retraining specialty the retrainees (a) had more positive attitudes in terms of job satisfaction, utilization of talents and training, and general morale; (b) were more highly motivated to do a good job and to learn new skills; and (c) were more satisfied with their relationships with their supervisor, peers, and subordinates than in the previous specialty.

A similar set of items for supervisors provided further information for evaluating retrainee adjustment. Supervisors compared retrained and non-retrained enlistees in general on additional attitudinal, motivational, and interpersonal relationship items. The descriptions provided for the two end points of the rating scale read "retrainees much better" and "non-retrainees much better." T-tests were accomplished to identify item means significantly greater or less than the scale midpoint (3 = "both groups about the same") which represented a neutral opinion. Means, standard deviations, and t-ratios for these items are shown in Table 2. Without exception, the observed item means were between scale point 2 (retrainees better) and 3 (both about the same) and, furthermore, were significantly ($p \leq .01$) lower than the scale midpoint. These results indicated that supervisors perceived retrainees collectively as having better attitudes toward work and military life, as being more motivated, and as interacting with other people better than their non-retrained counterparts.

In summary, preliminary findings indicated that retrainees were making the transition to the new specialty successfully. Attitude and motivation difficulties which might have suggested that changing jobs promotes adjustment problems were not observed. Rather, the retrainees consistently assigned higher than average ratings to the factors of interest and were found to be more positive toward their new occupational environment than toward the former, before-retraining specialty. Supervisors' appraisals were highly supportive of the trend reflected in the retrainees' judgments. In appraising retrained airmen in a more positive light than non-retrained airmen, the supervisors depicted the typical retrainee as adjusting capably to a change in military occupation.

## Retraining Impact on Career and Job Progress

The impact of retraining on military career and job progression was assessed through the supervisors' and retrainees' perceptions of the effect of retraining on promotion, reenlistment, type of work assigned, and technical ability to perform the job. On the promotion and reenlistment measures, the respondents were asked to evaluate whether retraining had increased, had no effect on, or had decreased chances for promotion and likelihood of reenlistment. The retrainees provided a self-report of their own experiences, while supervisors assessed the situations for retrainees in general. Since supervisors who had and who had not retrained may hold disparate views on these issues, they were separated for analysis purposes. Item descriptive statistics are shown for the retrainees, for supervisors by retraining status, and for the combined supervisor group in Table 3. T-tests were conducted to determine if the item mean ratings assigned by supervisors (retrained versus non-retrained) and retrainees on the promotion and reenlistment measures differed. The resultant t-ratios and statistical significance decisions are shown in Table 4. First, items means on the promotion measure indicated that all of the analysis groups perceived that retraining increased slightly or had

no effect on promotability. The retrainees felt that their chances for promotion were less positively influenced than the supervisors, regardless of their retraining status. Mean ratings assigned by the supervisors who had retrained and who had not retrained did not differ. However, supervisors as a group perceived that retraining improved chances for military rank progression to a significantly greater extent than did the retrainees. On the reenlistment measure, all groups felt that retraining slightly increased the likelihood of the enlistees signing-on for another military tour. Retrained supervisors were found to have the most positive assessment of the influence of retraining on propensity to reenlist. Non-retrained supervisors were less positive, and retrainees were the least positive. The differences in mean ratings for each of the three comparisons -- retrained versus non-retrained supervisors, retrained supervisors versus retrainees, and non-retrained supervisors versus retrainees -- were found to be significant ($p \leq .01$).

To help evaluate the retraining effects on job progression, the retrainees' perceptions of the type of work assigned and their technical ability to perform were compared for their pre- and post-retraining specialties. Table 5 provides a summary of the item statistics and t-tests. Item means indicated that retrainees felt they had an average to good opportunity for challenging work assignments in terms of work difficulty and responsibility level in both the previous and current specialties. The ratings assigned by retrainees to describe their technical ability to do their job and to supervise others were also positive and were somewhat higher than ratings on the work assignment items. Item mean ratings for the previous and current specialties significantly differed, except for the item on technical ability to perform their job. Retrainees perceived their ability to perform the technical aspects of their work before and after retraining to be comparable. Otherwise, the retrainees' experiences in the current specialty received more favorable ratings. The retrainees reported that the opportunity for more difficult work and responsible positions improved in the current specialty, as did their ability to supervise others on the job.

The findings may be summarized by the statement that no repercussions of a negative nature ensuing from a change in the military job were evidenced in the supervisors' and retrainees' reports. The retrainees and supervisors agreed that neither an enlistee's chances for achieving higher military ranks nor his/her intentions concerning reenlistment were adversely impacted by the retraining experience. In fact, the job changes were generally perceived as having a stimulating effect; retrainees were, however, somewhat more conservative in their assessments than supervisors. Retrained and non-retrained supervisors shared a common viewpoint on the subject of promotions but not of reenlistment intentions. On the latter subject supervisors who had retrained took a more positive stand. In addition, the retrainees reported that they could perform their current and previous jobs equally well, but were more favorably disposed toward the current specialty on work assignment opportunities and their ability to supervise. Collectively, these findings reflected favorably on consequences of retraining.

Job Skill, Ability and Performance Assessment

Retrainees' ability and performance on the job in the new specialty were examined using supervisors' comparative ratings of retrained and non-retrained enlistees in general on six skill and performance items. Table 6 presents

means, standard deviations, and t-ratios by item. As described earlier, t-tests of differences between item means and the scale midpoint (neutral opinion) were conducted. The supervisors consistently rated retrainees better than non-retrainees. On all items the observed means were significantly ($p \leq .01$) different from the neutral opinion reference point (retrainees and non-retrainees about the same). The results indicated that retrainees as a group were, in the supervisors' opinions, superior to their non-retrained cohorts with respect to job skills and knowledge, supervisory ability and potential for promotion, as well as the quality, difficulty level, and amount of work performance. These preliminary findings lent support to a conclusion that retrainees generally have acquired the requisite job skill and knowledge and have performed to their supervisors' satisfaction in the new occupation.

## Conclusions and Recommendations

Overall, the current findings may be viewed as demonstrating that the Airman Retraining Program objective of interchanging enlisted personnel between military occupations is being accomplished with minimal impact on the individual participants and their supervisors. The change in their specialty does not appear to be viewed by the retrainees as a stumbling block to the attainment of their job-related or military career goals. Supervisors do not seem to be of the opinion that retraining interferes with or disrupts job accomplishment or personnel morale in their occupations. On the surface, the retraining program is apparently operating smoothly and promoting the successful assimilation of enlistees into second military specialties.

The study results presented are preliminary in nature. Inasmuch as the current findings are based on global measures of attitude, adjustment and performance, some caution in interpretation and restraint in generalization should be exercised. Additional work is needed to refine the data and further analyses are required before definitive statements about retraining program efficiency would be fully supportable. Analyses on a specialty-by-specialty basis are needed as retraining actions may be operating smoothly for some job changes but not for others. The influence of personal characteristics and reasons for and circumstances of retraining on adjustment and performance need to be assessed. The voluntary or involuntary nature of an enlistee's retraining is of particular interest. Likewise, an evaluation of the transferability of administrative and technical skills between specialties and of the effects on retraining success would be helpful. In conclusion, while preliminary findings are encouraging, further research to comprehensively evaluate the Airman Retraining Program is recommended.

## References

Air Force Regulation 39-4. Airman retraining program. Washington, D.C.: Department of the Air Force, 28 November 1979.

Miller, R.G., Jr. Simultaneous statistical inference. New York: McGraw-Hill Book Company, 1966, 67-70.

Table 1. Retrainees' Appraisal of Attitudes, Motivation and Interpersonal Relations in Previous and Current Specialty

| Item | Valid N | Previous Specialty (Pre-Retraining AFS) Mean | SD | Current Specialty (Retraining AFS) Mean | SD | t-ratio |
|---|---|---|---|---|---|---|
| Attitudes | | | | | | |
| Job satisfaction | 12802 | 2.92 | 1.43 | 3.58 | 1.24 | -36.58[a] |
| Use of talents & training | 12784 | 3.17 | 1.34 | 3.62 | 1.17 | -27.32[a] |
| General morale | 12730 | 2.89 | 1.39 | 3.41 | 1.28 | -29.15[a] |
| Motivation | | | | | | |
| To do a good job | 12797 | 3.34 | 1.36 | 3.85 | 1.16 | -32.35[a] |
| To learn new skills | 12720 | 3.28 | 1.35 | 3.91 | 1.17 | -39.25[a] |
| Interpersonal Relations | | | | | | |
| With supervisor | 12825 | 3.69 | 1.26 | 4.05 | 1.05 | -26.22[a] |
| With subordinates | 12765 | 4.00 | .96 | 4.13 | .89 | -13.18[a] |
| With co-workers | 12756 | 4.09 | .92 | 4.20 | .86 | -12.52[a] |

Note. Rating Scale Points: 1=Poor, 2=Fair, 3=Average, 4=Good, 5=Excellent.

[a]Indicates significance (Bonferroni $\alpha$ = .01).

Table 2. Supervisor Appraisal of Retrainees Versus Non-Retrainees on Attitude and Adjustment Items

| Item | Valid N | Mean | SD | t-ratio |
|---|---|---|---|---|
| Attitude | | | | |
| Toward work | 11449 | 2.64 | .80 | -47.64[a] |
| Toward military life | 11434 | 2.60 | .77 | -54.78[a] |
| General morale | 11448 | 2.69 | .75 | -43.55[a] |
| Motivation | | | | |
| To learn new skills | 11432 | 2.45 | .87 | -66.94[a] |
| To do a good job | 11420 | 2.52 | .84 | -61.17[a] |
| Interpersonal Relations | | | | |
| With co-workers | 11448 | 2.80 | .68 | -31.52[a] |
| With supervisors | 11450 | 2.73 | .73 | -39.07[a] |
| With subordinates | 11423 | 2.87 | .73 | -19.65[a] |

Note. Rating Scale Points: 1=Retrainees much better, 2=Retrainees better, 3=Both about the same, 4=Non-retrainees better, 5=Non-retrainees much better.

[a]Indicates observed mean is significantly less than (Bonferroni $\alpha$ = .01) scale midpoint value (neutral opinion).

Table 3. Effect of Retraining on Retrainees' Chances for Promotion and Likelihood of Reenlistment

| Rater | Chances for Promotion | | | Likelihood of Reenlistment | | |
|---|---|---|---|---|---|---|
| | Valid N | Mean | SD | Valid N | Mean | SD |
| Retrainee | 12832 | 3.18 | 1.25 | 12839 | 2.67 | 1.17 |
| Supervisor | | | | | | |
| Retrained | 5185 | 2.95 | 1.01 | 5167 | 2.31 | .85 |
| Not-retrained | 6208 | 2.92 | .95 | 6179 | 2.42 | .85 |
| Total group | 11393 | 2.94 | .98 | 11346 | 2.37 | .85 |

Note. Rating Scale Points: 1=Increases (chances/likelihood) a lot, 2=Increases somewhat, 3=Has no effect, 4=Decreases somewhat, 5=Decreases a lot.

Table 4. T-tests Between Supervisor (Retrained Versus Non-Retrained) and Retrainee Ratings on Promotion and Reenlistment Items

| Test | Chances for Promotion | Likelihood of Reenlistment |
|---|---|---|
| 1. Supervisor-Retrained vs Not Retrained | 1.78 | -7.12[a] |
| a. Supervisor (Retrained) vs Retrainee | n/a | -20.35[a] |
| b. Supervisor (Not Retrained) vs Retrainee | n/a | -15.06[a] |
| 2. Supervisor (All) vs Retrainee | -16.55[a] | n/a |

Note. n/a - designates comparisons which were not appropriate based on prior tests.

[a]Indicates significance (Bonferroni $\alpha$ = .01).

Table 5. Retrainees' Appraisal of Work Assignments and Technical Ability in Previous and Current Specialty

| Item | Valid N | Previous Specialty (Pre-Retraining AFS) Mean | SD | Current Specialty (Retraining AFS) Mean | SD | t-ratio |
|---|---|---|---|---|---|---|
| Work Assignments | | | | | | |
| Difficult work assignments | 12791 | 3.12 | 1.40 | 3.75 | 1.20 | -37.62[a] |
| Responsible positions | 12723 | 3.14 | 1.38 | 3.71 | 1.25 | -34.06[a] |
| Technical Ability | | | | | | |
| To do job | 12810 | 3.84 | 1.13 | 3.88 | 1.01 | -3.20 |
| To supervise others | 12749 | 3.51 | 1.24 | 3.66 | 1.15 | -11.15[a] |

Note. Rating Scale Points: 1=Poor, 2=Fair, 3=Average, 4=Good, 5=Excellent.

[a]Indicates significance (Bonferroni $\alpha$ = .01).

Table 6. Supervisor Appraisal of Retrainees Versus Non-Retrainees on Skill, Ability and Performance Items

| Item | Valid N | Mean | SD | t-ratio |
|---|---|---|---|---|
| Skills and Ability | | | | |
| Job skills and knowledge | 11449 | 2.85 | .86 | -18.92[a] |
| Supervisory skills | 11393 | 2.69 | .82 | -39.93[a] |
| Potential for promotion | 11442 | 2.81 | .83 | -24.40[a] |
| Work Performance | | | | |
| Quality of work performance | 11441 | 2.75 | .81 | -32.59[a] |
| Amount of work performance | 11438 | 2.78 | .78 | -30.48[a] |
| Difficulty of work performance | 11424 | 2.83 | .81 | -22.20[a] |

Note. Rating Scale Points: 1=Retrainees much better, 2=Retrainees better, 3=Both about the same, 4=Non-retrainees better, 5=Non-retrainees much better.

[a]Indicates observed mean is significantly less than (Bonferroni $\alpha$ = .01) scale midpoint value (neutral opinion).

AD P001378

# A Rationale for Designing and Managing Technical Training Programs:

Brandon B. Smith, Associate Director
Minnesota Research and Development Center
for Vocational Education
University of Minnesota
Minneapolis, Minnesota

## Rationale

The effective design and management of technical training programs is in a function of four factors (1) the learner, (2) the type and form of the stimulus content, (3) the type of behaviors to be developed and (4) the instructional variables to be controlled in presenting the stimulus content to learners.

The model discusses these four variables as major factors for designing and managing any technical training program regardless of the form or mode of the instructional delivery.

Implications of the model suggests that content must be "chunked" and that the structure of content can effectively be taught from whole to part rather than from part to whole as is currently practiced in most education and training programs. This does not negate the importance of the task analysis, but rather places the tasks within a larger, more functionally relevant context, to facilitate initial learning, retention and transfer of skills and knowledge.

The current instructional practices will be discussed and compared to the proposed model together with a general discussion of the variables affecting the learning environment.

## Introduction

Both the military (Zeidner, 1981) and vocational education are concerned about the quality and quantity of the pool of applicants in terms of increasing the effectiveness and efficiency of instruction for various technical traning programs. The training model generally assumes that the purpose of technical training programs is to modify the behaviors of trainees in such a way that they are capable of learning the technical content or skills necessary to perform in an occupational speciality.

The purpose of this presentation is to present and discuss a rationale for designing and managing technical training programs in the military and vocational education. This paper is based on a comprehensive review of research (Smith and Currey, 1981) conducted in the fields of educational psychology, the military, vocational education and industry. The review identified the variables which are common to all types of instructional delivery systems (e.g., group instruction, individualized instruction, self-paced instruction, computer assisted instruction).

This paper is divided into the following three parts: (1) review of the rational of existing instructional systems in the military and vocational education, (2) discussion of the proposed instructional rationale and variables, and (3) a proposed instructional strategy designed to increase the effectiveness and efficiency of the initial learning retention and transfer of skills and knowledge in military and vocational training programs.

## Part 1

## Current Instructional Practices

Current instructional practices in the military and Vocational Education are for the most part derived from the concept of task analysis (Prosser, 1938) (Fryklund, 1944) (Gagné, 1964) (Christal, 1970) supplimented with specification of behavioral objectives (Ammerman and Melching, 1966; Mager, 1961, 1968). In more recent years, new technological innovation have been proposed such as individualized instruction (Pucel and Knaak, 1975; Howes, 1971) and still more recently the implementation of computer assisted instruction for the Air Force (King, 1975), (Brown, et al, 1976) (Judd, et al, 1974) (Brown, et al, 1976) and (Dallman and Deleo, 1977). In a recent study of the cost-effectiveness of individualized and computer assisted instruction as used in the military (Orlanski, 1979), the finding suggest that attrition rates are higher than other forms of instructional delivery, but that it is more effective for those learners who do complete the training program. While these new technologies have made it possible to present information or tasks to learners in a more efficient manner and with a greater degree of learner control over the rate at which stimulus content is presented to the learner, the fact remains that the content is presented to learners in small pieces or learning tasks.

More recently, Gagné (1963, 1964, and 1959) discussed the psychological conditions of learning as they relate to learning stimulus content. His theory suggests (as do Prosser and Fryklund's) that the learning process progresses through eight stages of hierarchical development: (1) signal learning, (2) stimulus response learning, (3) chaining, (4) Verbal associations, (5) mulitple discriminations, (6) concept formation, (7) principle learning and (8) problem solving. It is believed that mastery of concrete signal learning tasks are prerequisite to learning higher order tasks or concepts. In general, mastery of lower order concrete tasks or behaviors are believed to be prerequisite to learning the next higher order tasks or behaviors. Content stimuli are therefore presented to learners in the form of a series of concrete learning tasks.

While the historical perspective of instruction in the military, vocational education and education in general, has primarily focused on task analysis and behavioral objectives, other theorists have proposed what might be referred to as chunking or whole to part learning (Miller, 1956) (Bruner, 1960, 1963) (Ausubel, 1963, 1968) (Shoemacher, 1967) and (Brown, et al, 1959). While there may be theoretical differences among these theorists, their commonalities focus on (1) readiness of the learner, (2) the structure of knowledge, (3) restructing or chunking stimulus content, and (4) focus on deductive/whole to part organization and sequence of stimulus content.

## Part 2

### Proposed Instructional Variables

Figure 1 shows the factors or variables which are believed to be generalizable to any instructional delivery system. Each of these variables together with their component elements will be discussed separately. Figure 1 suggests that modifications of individuals necessarily begins with a learner who interacts with some form of content stimulus to produce some desired behavioral changes in an individual through some planned instructional process. The process is implemented in an manner which accounts for individual learner differences.

### The Learning Process

There seems to be no general disagreement about the definition of learning (Hilgard, 1966) that learning is a process that brings about changes in an individual resulting from some planned or unplanned situation which cannot be explained by native response tendencies, maturation, or drugs. The definition of learning posed by the rationale of this paper suggests that all learning is the result of an individual who receives some stimulus through psycho/sensory perceptions (the five senses) and who encodes, decodes, and assimilates the stimulus with other stimuli already familiar to the learner. In general, learning is the process of receiving, encoding, decoding, and assimilating some stimulus content. Formal instruction is therefore the controlled process by which stimulus are presented

to, received by, and subsequently encoded, decoded, and assimilated with other stimulus content already familiar to the learner.

Stimulus Content

According to Bruner (1966) all instructional content (stimulus content) must necessarily be drawn from some combination of three types of stimuli: (1) enactive, (2) iconic, and (3) symbolic.

Inactive stimuli are those stimuli which require hands-on trial and error experiences. Examples of inactive stimuli may include driving a car/bicycle, flying a plane, water skiing, etc. In general, inactive stimuli are those sets of tasks which must be demonstrated to a learner, and are hands-on psychomotor tasks which usually cannot be acquired or developed by reading about them.

Iconic stimuli are those which are presented to a learner in the form of some graphic, pictorial representation. An iconic stimuli can display information or concepts in a manner which allows the learner to form a mental picture of some stimulus content.

Symbolic stimuli are most typically represented by spoken or the written word such as might be found in textbooks, or technical training manuals. It is probably the most frequently used and accepted form presenting stimulus content to learners but, requires learners to be able to receive, encode, decode, and assimilate these stimuli in some meaningful way in order for it to be learned, retained and used in subsequent learning tasks.

Behaviors

There appears to be general agreement among learning and instructional theorists (Gagné, 1963) (Bloom, 1968) (Bruner, 1966) (Mager, 1968) (Ammerman and Melching, 1966) that learning is designed to produce observable behavioral changes in individuals. Most learning and instructional theorists would also agree that there are three categories of behaviors: (1) Cognitive, (2) Psychomotor and (3) Affective. Each of these behaviors are developed as a result of presenting some stimulus content (enactive, iconic or symbolic) to learners through some formal instructional process.

Cognitive behaviors are perceived to be those related to the knowledge possessed by a learner and usually measured by paper and pencil achievement tests. These behaviors/knowledges are developed by the learners by presenting them with iconic or symbolic stimuli (e.g., textbooks, transparencies, technical manuals, lecture, or some form of printed material). Psychomotor behaviors are most readily developed by presenting enactive stimulus content to a learner (usually be demonstration) which requires some hands-on trial and error practice sessions. Psychomotor behavior are measured most frequently by performance tests or performance ratings.

Affective behaviors generally are perceived as attitudes, opinions or beliefs which are related to the work performance behaviors of learners

(Dawis and Weitzel, 1975). They may be perceived as a latent psychomotor behavior. That is, attitudes may not be readily observable, but will generally show up at some point in time in either a positive or negative psychomotor act or behavior (e.g. use/abuse of tools and equipment, timeliness, attendance, safe/unsafe work practices, etc).

In summary, learners must be able to receive, encode, decode and assimilate a combination of three types of content stimulus (enactive, iconic, or sumbolic) in order to develop each of three types of training related behaviors (cognitive, psychomotor or affective). The process by which this is accomplished is through the designing, planning, and managing a set of instructional variables.

Instructional Variables

Figure 1 also shows two categories of instructional variables which must necessarily be a part of any planned instructional delivery system: (1) planning and (2) implementing. Each of these and their respective elements will be discussed.

Planning is the a priori organization of the stimulus content of a training program which is to be presented to a learner. It includes the following four elements: (1) readiness of the learner, (2) organization/structure of stimulus content, (3) purpose/goals of instructional and (4) stimulus content sequence.

Readiness is perceived to be one of the most important yet probably the least understood and implemented part of any instructional plan. It is simply defined (Ausubel, 1963, 1968), (Bruner, 1960, 1966) and (Bloom, 1980) as presenting information or concepts (stimulus content) to learners with which the learner is already familiar and at a level of abstraction and inclusiveness such that it is most readily received, encoded, decoded and assimilated with existing information or concepts. Failure to do this requires learners to resort to rote learning strategies.

Organization and structure of the stimulus content is the second major element or instructional planning. It includes a prior organization, and structure of the stimulus content to be presented to a learner at a level of abstraction and inclusiveness already familiar to students. The structure and organization of stimulus content is a most important variable (Ausubel, 1963, 1968) (Bruner, 1960, 1966) (Shoemacher, 1960) and (Miller, 1956). It is best summarized by Miller's article "Seven Plus or Minus Two", which suggests that by structuring or chunking stimulus content for learners, initial learning, retention and transfer may be facilitated.

Goals and objectives are the third element in planning and have the effect of serving as an advanced organizer (Ausubel, 1963) for the learner by alerting the learner to the focus, on the standards and outcomes of instruction (Ammerman and Melching, 1963) (Mager, 1965, 1968).

Instructional sequence is the fourth component of instruction planning generally agreed to be important for facilitating effective learning

(Gagné, 1963 and Briggs, 1967). It is generally accepted that mastery of certain stimulus content is prerequisite to the learning of other related stimulus content. Gagne suggests that learning progresses in a hierarchical manner (1963).

In general, instructional planning is perceived to be prerequisite to the formal design and management of any technical instruction program. The specifies of an instructional system may differ (group vs. individualized computer assess instruction), but the elements or readiness, organization and sequence of stimulus content to achieve prespecified goals or objectives, are elements which must be considered in designing any instructional program.

Implementation

The second component part of an instructional system may be referred to as elements which attempt to account for individual differences in learner capabilities. The following are the four component parts followed by a brief discussion of each: (1) rate of stimulus content presentation, (2) frequency/contiguity and practice of stimulus content, (3) reinforcement/reward and (4) knowledge of results or feedback.

Rate of presentation of stimulus content represents a factor which compensates for different rates of learning. Some learners are able to encode, decode and assimulate stimulus content more rapidly than others. Some learners may learn from iconic stimuli (pictures) more readily than enactive or symbolic stimuli.

Frequency, contiguity/practice is another factor related to principles of instruction. In general, the more frequently and contiguous stimuli are presented to and practiced by a learner, the more important and proficient the learner becomes with respect to the use of that information or skill.

Reinforcement/reward is a well accepted fact for theories of learning and instruction (Bruner, 1966) (Skinner, 1969) (Ausubel, 1963). Reinforcement takes the form of intrinsic and extrinsic rewards. Extrinsic are those provided by the instructional environment (e.g., money, praise, tokens, etc.), where intrinsic rewards are those which the learner derives from an instructional experience. Regardless of the form of reinforcement or reward, reinforcement is a necessary and powerful factor to facilitate or control learner behavior.

Knowledge of results is closely related to the concept of intrinsic reinforcement since based on feedback of correct or incorrect performance, learner behavior is modified in acordance with correct or expected standards/answers. Failure of an instructional system to provide immediate feedback to learners reduces instructional effectiveness.

## Part 3

## A Proposed Rationale for Instruction

The proposed rationale for military and vocational education technical instruction is admittedly eclectic since it is based on many of the ideas represented in a review of research (Smith and Currey, 1981) and does not draw upon one school psychology but rather represents a synthesis and application of the best of thinking of many theorists about instructional effectiveness. The rationale is based on the validity of the following five assumptions:

Assumption 1: A learner is a stimulus receiving, seeking, and responding organism without which psychological frustration is experienced.

Assumption 2: Learners differ in terms of the rate at which they can receive encode, decode, and assimilate stimulus content.

Assumption 3: There is some finite number of concepts individuals are able to remember at a given time. By chunking content in more inclusive concepts, retention will be facilitated.

Assumption 4: All stimulus content for instruction must be drawn from one of three classes of stimuli presented to a learner at some level of specificity or abstraction: (1) enactive stimuli, (2) iconic stimuli or (3) symbolic stimuli.

Assumption 5: Learners are able to most readily encode, decode and assimilate material or concepts with which they are already familiar.

Based upon these five assumptions, it is now possible to suggest a rationale for the design and management of technical instructional programs which is believed to facilitate intitial learning, retention, and transfer of skills and knowledges.

### Planning Instruction

The following five steps are suggested as a basis for designing and managing technical instructional programs.

1. Conduct a task analysis to identify all the technical, (enactive stimulus) tasks to be presented to learners. These technical skills will most probably be taught separately through repeated demonstrations or practice, but will be functionally related to a more inclusive structure of knowledge and take on greater meaning and importance within this structure.

2. Conceptually and graphically restructure the stimulus content (tasks and knowledges) into no more than seven plus or minus two mutually exclusive and exhausitve clusters such that all or most skills and knowledges are subsumed by or can be included within the clusters.

3. Present the graphic reconceptualized or restructured stimulus clusters to a learner in a way that is already familiar to the learner.

4. Specify the goals and objectives of instruction to establish a state of readiness for the learner and instructor as well as making explicit the instructional procedures and standards of performance.

5. Sequence instruction deductively beginning with the most inclusive technical clusters of instructional stimuli using combinations of enactive, iconic and symbolic stimuli, while presenting specific skills knowledges or facts in a manner that relates facts or skills to the broader, more inclusive technical concepts.

## Implementation

The factors or variables that relate to the concept of implementation have already been discussed in a previous section of this paper and include the following four elements: (1) rate of stimulus content presentation, (2) frequency and contiguity with which content stimuli are presented to and practiced by the learner, (3) rewards or reinforcement schedule, and (4) knowledge of results or feedback.

It is believed that in the proposed rationale, the factors related to planning are essentially the same for all learners regardless of the mode of instructional delivery since the content stimuli and performance standard are set by the expectations of the program designers. The factors related to implementation are intended to account for individual differences in terms of differential abilities of the learner to receive encode, decode, and assimilate information or skills. From a management/control perspective, only the four elements of implementation mentioned above are variables which are intended to compensate for individual differences.

## Summary and Conclusion

The purpose of this paper was to present, discuss and propose a rationale for designing and managing technical instructional programs in the military and vocational education. The ideas contained in the paper were based on a comprehensive review of research learning and instructional theories, but the ideas were eclectic and do not necessarily reflect any one particular school of thought.

## References


Ammerman, J.K., and Melching, William H. The Derivation, Analysis and Classification of Instructional Objectives. Technical Report 66-4., HumRRO, George Washington University, Human Resources Research Office, May 1966.

Ausubel, David P. "Some Psychological Aspects of the Structure of knowledge," In Education and the Structure of Knowledge, Rand McNalley and Company, 1964.

Ausubel,David P. The Psychology of Meaningful Verbal Learning, Grune and Stratton, New York 1963.

Bloom, Benjamin. "Mastery for Learning, Evaluation Comment, Vol. 1, No. 2., Center for the Study of Evaluation of Instructional Programs, UCLA, Los Angeles, CA.

Briggs, Leslie, V. Sequencing of Instruction Relation to Hierarchies of Competency. American Institute for Research, Palto Alto, CA 1967.

Brown, George H. et al. Development and Evaluation of Improvided Field Radio Repair Course. Technical Report 58, Human Resources Research Office, George Washington Univesity, 1958.

Brown, John Seely, et al. Intelligent Computer Assisted Instruction (CIA) Applications. DHEL-TR-76-67, Technical Training Division, Lowry Air Force Base, Colorado 1976.

Brown, John Seely, et al. Reactive Learning Environment for Computer Assisted Electronics Instruction, AFHRL-TR-76-88 Technical Training Division Lowry Air Force Base, Colorado, 1976.

Bruner, Jerome S. et al. A Study of Thinking, Science Editions, Inc. New York, 1965.

Bruner, Jerome S. The Process of Education. Harvard University Press, 1968, Cambridge, MS 1963.

Bruner, Jerome S. Toward a Theory of Instruction. Harvard University Press, Cambridge, MS 1966.

Christal, Ray E. "Implications of Air Force Occupational Research for Curriculum Design", In Process and Techniques of Vocational Curriculum Development. Minnsota Research Coordinating Unit for Vocational Education, Minneapolis, Minnesota 1970.

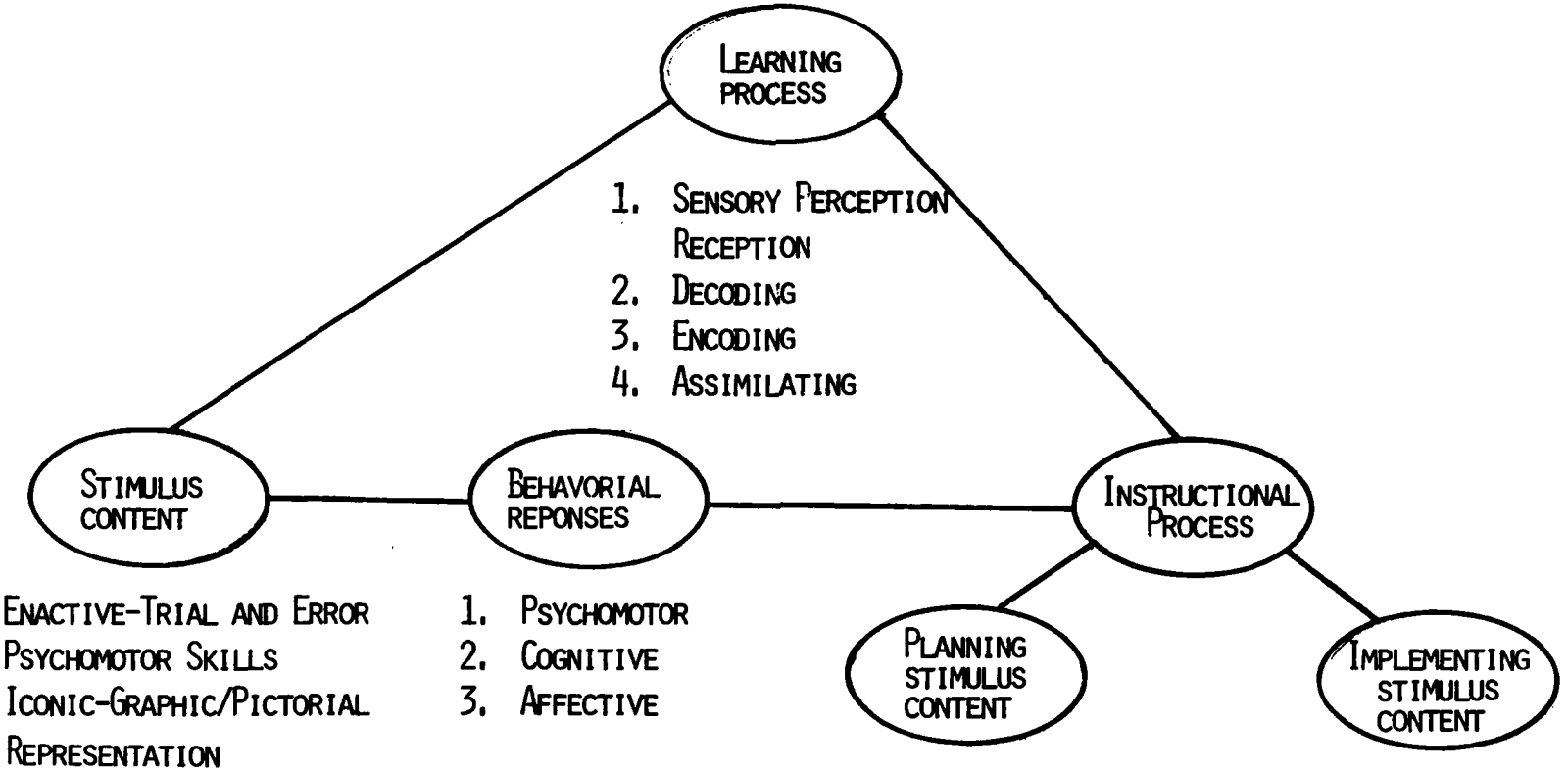
Dallman, Brian E. and Delea, Philip J. Evaluations of Plato IV in Vehicle Maintenance Training, AFHRL-TR-59. Technical Training Division, Lowry Air Force Base, Colorado, 1976.

Fryklund, Verne C. Trade and Job Analysis. Bruce Publishing Company, 1944.

Gagné, Robert M. Conditions of Learning. Halt Renehart and Winslow, Inc. New York, 1965.

# APPENDIX A

## ELEMENTS FOR DESIGNING AND MANAGING TECHNICAL INSTRUCTION PROGRAMS




1130

FIGURE 1

AD P001379

# Computer Assisted Task Selection in US Army SQT Development

Robert M. Smith, Ph.D.
USATSC Test Psychologist, SQT Management Directorate
Fort Eustis, Va 23604

The SQT is the US Army's major diagnostic instrument for evaluation of individual training and is an annual assessment of performance on 25 to 35 selected tasks. These tasks are selected by the service school which is proponent for each Military Occupational Specialty (MOS). However, there is little formal input from field units within this task selection process. This omission is compounded by an increasing rate of attrition on among subject matter experts at the schools and the diversion of experienced military personnel away from the training development area. Although the service school is formally responsible for the content of the SQT test plan, the knowledge and experience base at the operational level may be far removed from the field; thus, perceptions of the appropriate task emphasis may be inaccurate.

The Computer Assisted Task Selection (CATS) system is a data driven technology intensive alternative to the current SQT development system. CATS actively involves the field in the task selection process early in the SQT development cycle and continually solicits and monitors input from the field units regarding the importance of evaluating various tasks.

The CATS system operates on an automated data base of military tasks which is updated frequently with results of field surveys and on-going SQT testing and adjusts the content of the developing test plan in accordance with the subjective and objective requirements of the field. Through the use of self-adjusting automated systems like CATS the Army may produce more relevant, timely, and higher quality training products at lower cost.

The Skill Qualification Test (SQT) is the US Army's major individual training diagnostic instrument for enlisted personnel. It is developed and administered on an annual basis and evaluates performance on selected critical tasks in the soldier's Military Occupational Specialty (MOS). These critical tasks, the conditions under which they are performed, the standards of acceptable performance, and the specific elements which comprise the task are detailed in the Soldier's Manual (SM) for each MOS. The SQT is the primary determinant of individual training emphasis throughout the Army.

Although the Army Training Support Center (ATSC) is responsible for the production and distribution of SQT material, and scoring and reporting of SQT performance results, each of the 22 separate service schools and agencies, through their Directorates of Training Development (DTD), is contractually responsible to ATSC for the construction of their specific SQTs each year. The actual number of SQT produced, along with all the supporting material for each test, ranges from 12 to 170 separate test series annually for each school.

The ATSC provides guidance and test construction assistance to the schools through the SQT Management Directorate (SMD) and the publication of circulars, pamphlets, and directives. During the past 2 years, ATSC has also placed one or more test development specialists at each school for the purpose of monitoring the construction of, and assuring a high level of quality in, the SQT.

Within each school, the SQTs for which it is responsible are typically in different phases of a 15 month test development cycle, since each SQT has its own publication and administration dates. Additionally, the testing orientation and performance emphasis is different for combat arms, combat support, and combat service support branches of the force due to their differing technical requirements.

As can be seen in this brief description, the SQT is a large, complex, and resource intensive production and evaluation effort. To be effective and efficient, the technical quality and field relevance of SQT products must be as high as possible.

An attempt to assure the quality of SQT products has been made through provision of technical guidance and the establishment of developmental procedures by the ATSC test psychologist and specialists. A attempt to provide field relevant input to the SQT development process has been made through provision of SQT test plans to Major Army Command (MACOM) and Army Readiness and Mobilization Region headquarters and the provision of on-going test results to the schools. Unfortunately, neither of these attempts has been sufficient to accomplish the goal of high quality, field-relevant SQT. First, there is no mandatory documentation and audit of the process by which tasks to be evaluated by the SQT are selected. Second, there is no mandatory consideration of field input in the selection

of tasks for the SQT. An additional negative influence is the rate of attrition among test development staff at the schools and the resulting loss of institutional memory and the rotation or diversion of recently experienced military personnel away from the SQT development area. The knowledge and experience base at the schools thus may be far removed from current field reality and the resulting perception of appropriate training emphasis, which is driven by the SQT, may be inaccurate.

Clearly, new mechanisms must be developed and implemented in order to enhance the quality and relevance of the SQT evaluation system.

A brief review of the SQT development process at this point may be helpful in identifying the areas in which an alternative approach is required. The first crucial step in SQT construction is the selection of tasks from the SM to be included in the SQT test plan. The SQT test plan is comprised of 25 to 35 tasks selected from approximately 200 to 500 tasks which are appropriate for an MOS. The SQT test plan for an MOS is usually developed by one subject matter expert (SME), military or civilian, working alone, supplied with published guidance from ATSC. The SME may have recent field experience or he may have been assigned to the school for several years. The guidance with which the SME is supplied specifies, in detail, the selection process for each task, but does not require

documentation of the process. The SME is required to evaluate each task which is appropriate for the MOS (the task pool) on approximately 20 different characteristics. The characteristics include performance deficit, frequency of use, performance time, doctrinal currency, etc. If the task obtains high ratings on these characteristics, then its testing payoff is high and it is selected for inclusion in the test plan. If the task obtains low ratings on these characteristics, then its testing payoff is low and it is not selected for inclusion in the test plan. Ideally, all tasks in the MOS task pool will be subjected to this process, and only the highest rated (highest testing payoff) tasks will be selected for the SQT test plan. A common concern among those who observe this process at the schools is that these procedures are not followed consistently due to personnel shortages, lack of training, or lack of incentive to accomplish this time-consuming, labor-intensive task. Monitoring this undocumented, largely mental and unobservable process adequately consumes an exhorbitant amount of time and duplication of effort on the part of the evaluator.

Additionally, beyond the requirements for task selection on their individual characteristics, there are additional requirements for the composition of the final test plan as a whole. The number of tasks which fall into certain categories must not exceed quotas which are determinded by ATSC and by the relative number of tasks in certain other categories in the SQT test plan. These quotas are often unrelated to the importance of the task for the MOS and, in many instances, do

not accurately reflect the characteristics of the MOS task pool as a whole. If these SQT test plan quotas are exceeded, then the test plan must be reconstructed or a request for exception to policy must be initiated and forwarded to ATSC through the commandant of the school. Since the final approval of the SQT test plan rests with ATSC, and the produce test plan is easier to evaluate and verify in terms of task quotas than the process by which it is developed, it is far easier to construct an SQT test plan be selecting tasks which simply fit the guideline quotas than it is to carefully construct the test plan with tasks which form an integrated, meaningful and relevant whole which is seen as important and useful. The current system puts great pressure on the test developer to produce a test plan which satisfies policy but may ignore the actual training needs of the Army.

In regard to field input, under the current SQT development system, SQT test plans are sent to field commands (for informational purposes only) approximately 1 year prior to the test administration date. The importance of task groups and individual tasks to the field commander may or may not be communicated and considered by the schools for future SQT development. The comments are often not received in time to effect changes in composition of the current SQT test plan. The current system effectively ignores the input of the ultimate user of SQT results, the field unit commander.

The lack of documentation of task characteristics, objectivity in task selection, the mechanistic quality of task selection and the loss of intitutional memory through military personnel rotation and civilian personnel attrition, has converged in the need to automate the SQT test plan construction system. The use of computers to assist in the development of tests, test administration, and instructional management generally is extensively documented in the literature and will not be reviewed here. (Baker, 1980, Lippey, 1974, Smith, 1977, 1980.) It would be appropriate for the Army to investigate the effectiveness of a prototype computer assisted SQT development system which is now under development at the US Army Infantry School and which operates on the existing TRADOC computer network.

The Computer Assisted Task Selection (CATS) System establishes an automated file containing documentation of individual task characteristics and the selection criteria for the SQT test plans. This file is processed by a computer program which mimics the behavior of the human test developer in selecting tasks for the SQT test plan. The computer program makes an initial pass through the task file and marks as rejected those tasks which are doctrinally inappropriate for the MOS. During this initial pass the program also sums the testing payoff ratings for each task and adds or subtracts testing payoff points for increased or decreased emphasis of certain subject areas which are specified by the test developer. At the end of this initial

pass, the system reports the number of tasks in the various categories, thereby summarizing the composition of the task pool. The task file is also rearranged at this point by skill level, track, and testing payoff from high to low values. This assures that the final test plans will contain tasks which are specific to the MOS and track, and will be high in payoff. The second phase of CATS continues with several passes through the task file, one for each skill level and track in the MOS. Each task record is evaluated and accepted or rejected for inclusion in the test plan. Running totals of task characteristics and other program control parameters are maintained as the run progresses until the test plan is completed. The system then is reset and processing begins for the next test plan. Each test plan is evaluated for guideline compliance as it is developed and the status of success or failure is reported to the user.

The CATS system thus relieves the schools of the labor intensive and time consuming task of test plan construction. ATSC guideline compliance is automatically verified as the test plan is developed. CATS provides increased field input through incorporation in individual task records of the results of frequent surveys of field unit commanders regarding the emphasis they put on different aspects of the total task pool.

The CATS system is a data driven, technology intensive alternative to the current SQT development system. It operates on an automated data base of military task characteristics which is updated frequently with results of field surveys and on-going SQT testing. The system adjusts the content of the developing test plan in accordance with the objective requirements of the field. The utilization of this technology may provide the necessary mechanism to enhance the SQT construction effort.

## REFERENCES


Baker, F. B. Computer Managed Instruction: Theory and Practice Englewood Cliffs, New Jersey, Educational Technology Publications, Inc., 1980.

Lippey, G. (Ed.) Computer-Assisted Test Construction. Englewood Cliffs, New Jersey: Educational Technology Publications, Inc. 1974.

Smith, R. M., An Operational Computer Assisted Test Production System, Mid-South Educational Research Association Conference Proceedings, Birmingham, Alabama, November 9-11, 1977.

Smith, R. M., A Cybernetic Model of Instructional Management: Design, Development, and Implementation (Doctoral Dissertation, University of Alabama, 1980).

AD P001380

Staley, Michael R. & Weissmuller, Johnny J., Air Force Human Resources Laboratory, Brooks Air Force Base, Texas. (Wed. A.M.)

## Interrater Reliability: The Development of an Automated Analysis Tool

The Comprehensive Occupational Data Analysis Programs (CODAP) system is a basic tool in both operational job analysis and in occupational research. This system of programs is augmented and enhanced to meet changing requirements from the research community. In particular, this paper discusses the development and evolution of interrater reliability procedures. The history of these procedures is traced from techniques in use prior to its inclusion in the CODAP system through to the present time. This paper explains the program capabilities in terms of the research requirements which necessitated the enhancements and may be used as an analyst's guide in the future. The paper closes by examining current research streams, potential applications, and the operational use of the newest version of this program aimed at fulfilling these anticipated needs.

# INTERRATER RELIABILITY: The Development of An Automated Analysis Tool

Michael R. Staley
Johnny J. Weissmuller

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

## INTRODUCTION

Within the United States Air Force, as well as many other military and civilian agencies, there is a fundamental technology which supports occupational analysis for both operational and research programs. The core of this technology is called CODAP, an acronym for the Comprehensive Occupational Data Analysis Programs. The CODAP "system" is a set of analysis tools and procedures which use, as raw material, information provided by members of the occupational field being studied (Thew & Weissmuller, 1979). This system may be used to improve classification structures, assess job-related skills, verify the relevance of training courses and support a host of other applications for which an accurate knowledge of job content at the task level is desirable (Weissmuller, Moore, & Thew, 1980).

Prior to the 1970's, the CODAP system primarily analyzed relative time spent ratings from job incumbents. About 1970, it was decided that the value of CODAP reports would be significantly improved if supervisors' ratings of task importance could be reported alongside of incumbent performance data. Because the concept of "importance" is multifaceted, a technology was required to handle any number of factors which might be considered "important" for tasks in any specific application. One of the more promising "task factors" was Task Learning Difficulty, which allowed supervisors to rate the relative difficulty of training new personnel to perform various tasks within their specialties. A program called TSKDIF was added to the CODAP system to compute the mean task difficulty ratings of supervisors across all tasks in a job inventory. These ratings were used to resequence the tasks in standard job descriptions into descending order based on average task learning difficulty. Hence, within any specified job, Air Force managers could easily see the tasks hardest to learn and the degree to which journeyman level personnel were performing and spending time on those tasks.

About the same time, an interrater reliability program was being developed within the Laboratory. Although reliability theory was developed to measure the trustworthiness of test instruments, the applications within the Laboratory were concerned with assessing the level of agreement within a group of raters. This interrater reliability program allowed researchers to make some statements about the stability of the ratings they were collecting and reporting. In order to increase the credibility of the CODAP reports, it was decided that this procedure should be incorporated into the system and routinely applied to the task learning difficulty ratings. This paper traces the development and evolution of the interrater reliability technique within the CODAP system.

**PRE-CODAP VERSIONS: EARLY 1972**

One of the earliest interrater reliability programs within the Laboratory was developed by Mr. Manuel Pina. The program was later modified and applied to aptitude rankings of 207 task statements from eight career ladders. This work was accomplished by Mr. Charles A. Greenway in a non-CODAP applications section. Although the use of ratings from 10 research scientists was a modest beginning, the results were promising enough to warrant further development. Subsequent uses in "real-world" settings uncovered some operational shortcomings. The program required that all raters rate ALL tasks. Some analyses had to drop as many as 80 percent of the raters because they failed to meet this condition. As this was deemed unacceptable, it was clear that further changes were required before inclusion in CODAP.

**THE FIRST CODAP VERSION: AUGUST - NOVEMBER 1972**

Dr. Raymond E. Christal, then Chief of the Occupation and Career Development Branch of the Personnel Research Division of the Air Force Human Resources Laboratory, assigned to Mr. William J. Phalen the responsibility of integrating an interrater reliability procedure in the CODAP system. Mr. Phalen worked with Master Sergeant (MSgt) William D. Stacey, head of the CODAP Programming Unit, to coordinate the development of the required program. MSgt Stacey assigned the programming task to Airman First Class (A1C) Johnny J. Weissmuller. Many of the details of the first CODAP version were worked out between Mr. Phalen and A1C Weissmuller (Stacey, Weissmuller, Barton & Rogers, 1974).

The new CODAP program (RXXNDX) followed the example of previous versions and reported the interrater reliability coefficients ($R_{11}$ and $R_{kk}$), computed by Lindquist's intraclass correlation technique (Lindquist, 1953, p.361) and, optionally, the mean rating for each task. In addition, this program provided some new capabilities. It compensated for raters who did not rate every task (Haggard, 1958, p.14) and, optionally, adjusted the ratings of raters who tended to rate either high or low as compared to all other raters. The program also reported the number of tasks rated, the rater's correlation with the grand mean vector, the mean of the rater's ratings and the mean of task means for the tasks rated by the rater.

As the program was nearing completion and several tests were being run, it was noted that the mean of the task means was not 4.0 as would be expected from a seven-point relative scale. The standard deviations also varied widely from one data set to another. It was decided that if two different factors were ever evaluated within a single analysis, it would be important to standardize the task means to some common value with a specific standard deviation. This would allow one to compare different ratings on a given task and determine which factor was "more critical." The program was modified so that if any adjustments were made to individual ratings, the task means would be standardized to a mean of 5.0 and 1.0 standard deviation. This version of the RXXNDX program was released in November 1972.

**THE FIRST YEAR REVISIONS: NOVEMBER 1972 - NOVEMBER 1973**

During the first year of use, several revisions were required to handle ongoing research projects. One of the most prominent research streams was the investigation of multiple rating scales conducted by Lt. Col. James B. Carpenter (Carpenter, Giorgia, & McFarland, 1975). Interrater reliability measures were desired for ratings on 7-, 9-, 25-, 99- and 9999-point scales. It should be pointed out that the 9999-point scale was actually a direct percent time spent estimate in which two decimals of accuracy (e.g. 45.67) were permitted. The original version of RXXNDX only permitted single digit ratings. These ratings were assumed to be in the range of 1 to 7, with the SETCHK program doing all error-checking and resetting out-of-range values to zero. The RXXNDX program was modified to handle ratings of up to six digits.

Later in the first year, it became apparent that the report of rater correlations with the grand mean vector was effectively identifying deviant raters. As the reliability of a test instrument can be improved by the removal or replacement of poor items, so can the reliability of the mean ratings be improved by the removal of deviant raters. Since the primary measure of stability is the $R_{kk'}$, removing too many raters may actually reduce stability (Haggard, 1958, p.89), and hence the removals should not be done indiscriminately. For this reason, a capability was added to the RXXNDX program to bypass any analyst-specified raters. A t-value column was added to the rater correlation report in order to help analysts evaluate the significance of correlations based on the differing number of tasks rated by each rater. Although the rules-of-thumb varied from analyst to analyst, deviant raters were usually defined to be raters showing a very low (or negative) correlation with the task mean vector across all raters, or those having a t-value of 1.65 or less.

During this period the program name was changed because "RXXNDX" was deemed unpronounceable. As the program was more commonly called "REXALL," the name change was made official to avoid confusion and was justified as "RXXNDX with ALL options." As time went on, however, it became clear that ALL options had not yet been added.

## THE EXPLOITATION OF TASK FACTORS: NOVEMBER 1972 - OCTOBER 1976

During this time frame, REXALL became a standard part of the Air Force's operational occupational analysis procedures. By mid-1975 other task factors were being considered and programs were developed to exploit this newly acquired methodology (Christal & Weissmuller, 1976). Some of the motivation for pursuing task factors at this time was the high-level discussion in the Australian Department of Defence aimed at adopting a single occupational analysis system as their DoD standard. Because there were two competing systems being used in the Australian armed forces, their basic choice was between the USAF version of CODAP adopted by the Australian Air Force and an independently developed Australian Army system. A major advantage of the Army system was its ability to effectively handle various task factors. As these capabilities seemed highly desirable, the USAF developed and integrated a generalized task factor capability into the CODAP system. This action may have contributed to the Australians' final decision to adopt the CODAP approach as their standard. The set of programs which comprise this capability has become known as the Task Factor extension or the Task Factor Applications Package.

With the increased emphasis on new factors and the intention of making more use of these data, the practice of removing so-called "deviant" raters came into question. Some researchers felt that the raters who were removed might represent a consistent, valid, albeit minority viewpoint. Squadron Leader (SQN LDR) Kenneth Goody, an Australian exchange officer working at the Laboratory, investigated this hypothesis. He concluded that, in general, raters removed using the REXALL process were simply uncooperative (Goody, 1976).

During the course of this analysis, however, he wondered if the deviant raters were members of a smaller, easily identified subgroup with a different policy. He hoped that these groups might be identified by some background characteristic such as grade level, major command, etc. SQN LDR Goody obtained a copy of the REXALL program and modified it to permit the screening of raters based on predetermined categories. He had trouble in testing these ideas because background data were not routinely collected from the supervisory raters. Even when he obtained the necessary information, his technique required the use of special programming to code each rater according to the category of interest prior to running REXALL. Although this was inconvenient, the potential use of that option was considered important enough to warrant making his version of the program the CODAP standard. Also based on his recommendation, data collection procedures were modified to solicit more background information from raters.

## THE IMPACT OF TRAINING EMPHASIS: 1978 - 1979

Training Emphasis is a task factor designed to capture a field supervisor's recommendations about which tasks ought to be included in entry-level training and how much relative emphasis should be placed on each task (Ruck, Thompson & Thomson, 1978). This task factor departed from the

CODAP standard 1 to 9 rating scale. For this factor, raters were asked to use a scale of 0 to 9, in which zero meant "Do not train." Although this is quite logical, there were some problems in using these data in the REXALL program. REXALL considered ratings of zero to be nonresponses and substantial program changes had to be made. In addition, the SETCHK program which range checks the data was also modified to permit zero as a valid rating.

Training emphasis continued to impact on REXALL. Although many career fields produced stable vectors of task means, some Air Force specialties failed to achieve a minimally acceptable value for interrater agreement. These specialties were labelled "complex specialties" and further analyses were conducted. It was thought that there might be different rater policies for various subsets of tasks. Various policy-capturing techniques were attempted without much success. One subsequent attempt involved using REXALL to identify not only a reduced set of raters, but also a reduced set of tasks on which there would be general agreement. The REXALL program was modified to make multiple passes of the data and compute interrater agreement on sets of tasks that were rated by at least a specified percentage of raters (e.g. 10%, 20%, 40%). Of course, for these purposes, a zero had to be considered a nonresponse.

Unfortunately, this modification also failed to achieve the desired results. It was next hypothesized that the 0 to 9 scale was inadequate to fully represent the range of emphasis intended by the raters. Because zeroes were averaged in with the typical 1 to 9 ratings, the overall means for tasks only reached a maximum value of about 3 for the most important tasks. The proposed solution was to reweight the responses so that 9's became 512, 8's became 126, and so on. This line of thought was explored by another Australian exchange officer, SQN LDR Michael J. Cassidy, and several such weight substitutions were suggested. The REXALL program was modified to permit this capability, but none of the substitution schemes seemed to solve the problem.

The next approach to complex specialties was to try clustering raters using the standard CODAP procedures. Using these procedures, the research analyst was provided with a cluster-merger diagram from which to identify rater groups of interest. It was hoped that these empirically defined groups would represent the various rating policies. Using this approach, a problem was identified in that the REXALL program was designed to use all raters on the input file with the exception of those it was specifically told to ignore. The clustering programs, on the other hand, produced short lists of raters to be used instead of the longer lists of raters to be ignored. For example, in this analysis groups typically consisted of 20 to 30 raters, while the rest of sample usually contained 300 raters. Rather than require researchers to specify all raters NOT included in each group, a new version of REXALL called "REXNON" was made which used only the raters requested. The new program caused "REXALL" to be reinterpreted as "interrater reliability on ALL raters" and "REXNON" to mean "interrater reliability on NONE but the specified raters."

**THE IMPACT OF STRENGTH & STAMINA: 1979 - 1980**

The Strength and Stamina research project was designed to help identify career ladders and tasks within career ladders that warranted a further in-depth study of various physical demands (Gott & Alley, 1980). As in the Training Emphasis area, a 0 to 9 scale was used. Because of the lessons learned, however, raters were required to enter an actual zero digit to signify "no lifting requirement" and an "X" to signify "unknown requirement." Again, REXALL had to be modified -- this time to interpret an "X" as a nonresponse. The SETCHK program edits the data file that is later used by the REXALL program. This program also had to be changed and the required changes were not straightforward. For this reason a new version of SETCHK called "SETCKR" was developed.

As this project had short suspenses and spanned nearly all career fields within the Air Force, a large level of effort was required within a short time frame. For this reason a major portion of the logistics was completed under contract. In order to minimize training time, many of the computer runs were preprogrammed to require only clerical support. In the REXALL program, an option was added to record which raters were deemed "deviant" and on subsequent passes, automatically remove those raters.

**TRAINING EMPHASIS, CONTINUED: 1980**

In the intervening time period, several approaches outside of the REXALL methodology were tried with the training emphasis ratings in the "complex specialties". In 1980 another REXALL-based approach was suggested by William J. Phalen in which only those tasks with low standard deviations would be considered for inclusion in a stable, majority viewpoint. Instead of adding another option like the "percentage of raters" mentioned above, a more generalized approach was taken that reduced the task set based on the values of any specified task factor. For example, with this new option one could investigate the training emphasis for only those tasks that are performed by a high percentage of members. This and other approaches are still being considered but to date, no operational method has been adopted.

**CURRENT RESEARCH & FUTURE REQUIREMENTS -- REXSPC: 1981 - 1982?**

There are two projects currently underway which would be greatly facilitated by program changes. The first is a continuation of the Training Emphasis research project and is being done by the most recent Australian exchange officer, SQN LDR Hans P. Jansen. This analysis is exploring factor analysis techniques and addresses both the question of sample sizes and the reliability of task means. A large number of raters is being randomly subdivided into various groups and the stability of the interrater agreement cofficient is being monitored. Under the present system, approximately six computer runs are required to prepare the data, identify samples, select subsets and compute interrater reliability for each sample. A program called "REXSPC" is currently being developed which will reduce this to two runs. This may not seem significant, but there are over 300 random subsamples to be processed and this could save over 1200 computer runs.

The second project, being done by Captain James H. Gilbert at the USAF Occupational Measurement Center, is designed to evaluate the difference in training emphasis policies between supervisory personnel for maintainance of different aircraft systems. In this analysis, over 20 different aircraft systems are represented, and the intent is to produce a task mean vector for each system. Again, under the current system, this requires three computer runs, while the new REXSPC program would handle this in one run. If Capt. Gilbert's research finds significant differences between these groups of supervisors, this approach may become the operational standard for all future analyses. If this happens, REXSPC will result in a greater savings impact than is expected for the 20 systems currently being studied.

Not all questions have been answered -- indeed, not all questions have yet been asked. New questions usually require new tools; hence, all development has not yet been completed. The interrater reliability capability began simply as a method to produce a single number representing the confidence in mean ratings. It has expanded and is helping to form new guidelines for the types of raters and ratings that are necessary to provide Air Force managers with sound information on which to base their policy decisions. As long as Air Force managers face new challenges, the development and evolution of automated analysis tools will be required to meet their changing needs.

## References


Carpenter, J. B., Giorgia, M. J., & McFarland, B. P. **Comparative analysis of the Relative Validity for Subjective Time Rating Scales.** AFHRL-TR-75-63, AD-A017-842. Lackland AFB, TX: Occupational and Manpower Research Division, December 1975.

Christal, R. E., & Weissmuller, J. J. **New CODAP programs for Analyzing Task Factor Information.** AFHRL-TR-76-3, AD-A026-121. Lackland AFB, TX: Occupational and Manpower Research Division, May 1976.

Goody, K. **Comprehensive Occupational Data Analysis Programs (CODAP): Use of REXALL to Identify Divergent Raters.** AFHRL-TR-76-82, AD-A034-327. Lackland AFB, TX: Occupation and Manpower Research Division, October 1976.

Gott, S. P., & Alley, W. E. "Physical Demands of Air Force Occupations: A Task Analysis Approach." Proceedings of the Twenty-Second Annual Conference of the Military Testing Association. Toronto, Canada: Canadian Forces Personnel Applied Research Unit, October 1980.

Haggard, E. A. Intraclass Correlation and Analysis of Variance. New York: Dryden Press, Inc, 1958.

Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin, 1953.

Ruck, H. W., Thompson, N. A., and Thomson, D. C. "The Collection and Prediction of Training Emphasis Ratings for Curriculum Development." Proceedings of the Twentieth Annual Conference of the Military Testing Association. Oklahoma City, OK: United States Coast Guard Institute, October-November 1978.

Stacey, W. D., Weissmuller J. J., Barton, B. B., & Rogers, C. R. **CODAP: Control Card Specifications for the Univac 1108.** AFHRL-TR-74-84, AD-A004-085. Lackland AFB, TX: Computational Sciences Division, October 1974.

Thew, M. C., & Weissmuller, J. J. "CODAP: A Current Overview." Proceedings of the Twenty-First Annual Conference of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center, October 1979.

Weissmuller, J. J., Moore, B. E., & Thew, M. C. "CODAP: Applications and Their Implications for Higher Level Design." Proceedings of the Twenty-Second Annual Conference of the Military Testing Association. Toronto, Canada: Canadian Forces Personnel Applied Research Unit, October 1980.

AD P001381

# AN INNOVATIVE APPROACH TO DATA CAPTURE IN AUTOMATED ASSESSMENT.

Author - Major John F Taylor MSc, Training Consultant, Army School of Training Support, England.

Summary

The microcomputer has had a significant impact on the development of computer assisted learning. It has not resulted in a parallel development in computer managed learning especially in the area of computer based testing. One of the reasons for this is that testing, especially in the conventional group testing situation, requires efficient data capture. The usual medium for this has been by marked cards. Equipment to read cards is expensive compared with the processing power required. The development of a low cost input system is discussed in terms of the appropriateness of imposing program controlled answering strategies on testees attempting a conventional multiple choice objective question achievement test.

"Chips with everything" has taken on a new meaning in the British idiom. Articles in the popular press continue to debate the effect that microprocessors will have on everyday life. The British and other West European governments have initiated programs to support the spread of microcomputers throughout their educational systems. But what have been the consequences of computer technology for training and education? First; courses of instruction have been developed to improve computer appreciation and literacy; second; courses of instruction have been developed to produce trained personnel for the computer industry and to operate related technological equipment; third; increasing use has been made of computer technology as an aid to the training and learning processes. This paper is not concerned with the sociological or technological aspects but with the development of the microcomputer as an aid to training development.

Computational facilities have been available to trainers for a considerable period, usually through access to batch or time sharing systems. Initially this kind of facility was used for the provision of courses in computer programming, computer science and computer appreciation. Recent trends consequent to the increased availability of processing power have been to expand usage to support the learning of other subjects. In many subjects programs have been implemented which embody a variety of educational roles, including drill and practice, applications, tutorial, simulation and modelling. Applications programs and drill and practice programs are relatively easy to implement and use in conjunction with syllabuses taught primarily through an instrumental (learn by being told) educational philosophy. However, it is much harder to build simulation programs and integrate them into existing syllabuses, and even harder to accommodate modelling activity despite its enormous potential at the relational (learn by discovery) end of the learning dimension . However, the plethora of programs available does indicate that a great deal of resource has been committed in recent years to the development of quite sophisticated computer programs as an aid to learning. This effort, accelerated by the introduction of the microcomputer, has gone on in spite of the fact that the effectiveness of the computer program in each of these educational roles has not been rigorously evaluated.

Concurrent with the initial development of large systems as an aid to learning was the development of systems used for the management of learning. This included such roles as the monitoring of student progress, resource scheduling, routing of students through individual paths towards mastery of objectives, test item banking, test production, marking and analysis. Unfortunately these applications, which are fundamental to efficient and effective control of training are not nearly as transferrable to the microcomputer as the CAL applications.



PREVIOUS PAGE IS BLANK

There has not therefore been a corresponding wide spread development in the quality control aspects of training as a consequence of the microcomputer revolution. The primary reasons for this are that these applications demand a lot of processing power and/or the collection of large amounts of data. The management of learning by computer is therefore likely to be based on large systems which, in order to be cost justified, will be located in large training organisations. The problems of processing power are being overcome by advances in technology, as are the data collection problems through the development of networking techniques. However, the immediate need is to develop the stand alone microcomputer to be a more effective aid especially in the quality control aspects of training.

In the British army the lack of evaluative evidence for the efficiency of CAL and the organisational implications of a move towards self paced individual learning suggest that the group based instructional and testing system will be norm. Although limited progress is being made with a fixed mastery variable time system similar to that implemented by the US Marine Corps there is little likelihood of developments in the field of computer based testing or the introduction of such techniques as adaptive testing. However,the introduction of the systematic development of training programs and the increasing demands for greater efficiency have highlighted the need for accurately and efficiently making classification decisions in the existing criterion referenced achievement testing system. To assist in the quality control aspects of the army training system two levels of computer support have been developed. These are the Student Performance and Evaluation by Computer system (SPEC) and the Partially automated test marking and analysis system (PATMAS).

Each system is appropriate to a particular situation. PATMAS is a system which marks multiple choice objective question (MCOQ) tests, standardises scores and carries out item analysis. It is designed to be used in situations where there is a high throughput of data from testing during and/or on completion of short courses. SPEC does all that PATMAS does and much more, its raison d'être is to produce student progress profiles. SPEC is therefore appropriate in units which have students on long courses. Both these systems rely on marked cards for their data input. The Optical Mark Recognition (OMR) hardware used to read these cards is expensive (£6,000 - £10,000). This is not a problem in SPEC type systems costing £40,000 but the OMR hardware comprises 85 - 90% of the cost of a PATMAS system. OMR equipment was justifiable as the only means of relatively simple (for the student) but rapid (for immediate feedback) data capture. Unfortunately there is no indication that the price of OMR equipment is falling in line with the general trend in computer hardware. This means that the proportion of the cost of a PATMAS system devoted to data capture is rising above the already high level. Therefore, the only way to bring the PATMAS facility into the price range which will make it a justifiable system for many more units is to reduce the cost of data capture. The problem is to ensure that data collection continues to be simple for the student but sufficiently quick for the benefits of immediate feedback to be realised.

The use of cards as an input medium is generally being replaced by program controlled keying (PCK) where data is punched directly to magnetic medium, thus avoiding the intermediate cards. This is not only quicker but it overcomes the reliability problems which are a feature of OMR devices. However, the development of such a system is dependent upon the production of a low cost input device. Unfortunately the cost of such a device is inversely proportional to the degree of program control. Consequently the cheapest input device will result when the way in which the testee answers the test is controlled by the system.

This is of course at variance with current trends in the development of man machine interfaces and "user friendly" systems. The provision of the facility for each testee within the group to use his own strategy not only increases the cost of each individual imput device but will also make the control software more complex. Both of these factors will reduce the cost benefits of this approach to data capture and therefore mitigate against the more widespread use of quality control techniques. It was therefore necessary to establish what facilities need to be provided for the testee so that his performance on a test is not altered by the data capture device.

The basic hypothesis used was that there are three different phases to the answering of pencil and paper type MCQ tests.

a. An INITIAL phase - when an attempt is made to select one of the responses to as many questions as possible.
b. A REVIEW phase - when each question unanswered during the previous phase is reconsidered.
c. A REVISE phase - when all responses are reconsidered and changed if necessary.

The strategies which could be used to answer tests can therefore include some or all of these phases. They could range from the strictly sequential strategy which has only an INITIAL stage to a reiterative strategy which has a cursory INITIAL stage where only the "easy" questions are answered, followed by repeated REVIEW phases - each iteration dealing with "harder" questions. Hence, while it was thought that the three phases were common to most testees, the relative importance that a testee will attach to each phase will differ. In terms of automation of the data collection it is obvious that the model of the predominently serialist approach would be much easier to implement than the more complex predominently reiterative strategy. Unfortunately while a great deal of research has been conducted on other areas of testee response patterns (guessing, minority groups etc) nothing appears to have been published on this particular area of testee response strategies. It was therefore decided to carry out experiments to answer the following questions.

- Do individuals employ different strategies when answering a multiple choice objective question test ?
- If such differences do exist are they important ? ie will the imposition of a particular strategy adversly effect the score of those testees who prefer an alternative strategy ?

An initial experiment was carried out using 13 male Officer cadet who were studying English as a second language. The object of this experiment was to provide some evidence to support the initial hypothesis that testees employ differing strategies when answering tests. Each testee was observed as he answered the 40 written questions on an English language ability test. Observers sat beside testees as they attempted the test and recorded the number of the questions answered sequentially and the time taken to answer each question. Graphs of the response patterns for each testee were drawn and the strategies classified into four categories, namely, reiterative, reiterative/serialist, serialist/reiterative and serialist. The breakdown of testees by category is in table 1.

Table 1

| STRATEGY | S | S/R | R/S | R |
|---|---|---|---|---|
| NUMBER OF TESTEES | 6 | 2 | 2 | 3 |

From this limited evidence using relatively naive test takers it was decided that the hypothesis was supported and that further evidence should be collected to confirm these findings. A second experiment was therefore set up to gather more evidence of testee strategies and to make an initial investigation of the second question stated above.

Eighteen male officer cadets, all English, who were attending an educational foundation course prior to attending the Royal Military Academy were the testees in this experiment. The instrument used for the test was chosen from a bank of intelligence tests, the test used having been designed and validated for a population which best matched the background of the testees. In this case the rather cumbersome observation method of collecting data from which strategies could be ascertained was replaced by a self reporting form. It was also decided to use a repeated measures design. On the first application of the test testees were allowed to use their own strategies to answer the test and concurrently completed a self reporting form which showed the order in which questions were answered. In the second part of the experiment testees were forced to answer in a strictly serialist way with no recourse to a review or revise phase. To maintain the experimental conditions the testees were asked to complete a self reporting form giving levels of confidence in their answers.

Unfortunately several candidates failed to complete the test within the time allowed. It was therefore necessary when categorising testees by answering strategy to subjectively allocate these testees on the basis of incomplete information. The results of this second experiment are summarised in table 2.

Table 2

| STRATEGY USED IN FREE TEST | S | S/R | R/S | R |
|---|---|---|---|---|
| NUMBER OF TESTEES | 1 | 5 | 6 | 6 |
| MEAN SCORE IN FREE TEST | 41 | 46 | 32 | 34.7 |
| MEAN SCORE IN FIXED TEST | 42 | 51.6 | 35.5 | 37.5 |

The information produced from this experiment was insufficient to draw any conclusions about the effect that the imposed strategy had on testees, especially on those who were forced to make the greatest adjustments. It appeared that serialists on the whole scored better than reiterators. This factor cast suspicion on the decision criteria used to distinguish reiterators. However, the experiment did serve to confirm the method used. It was therefore decided to repeat the experiment using a more typical sample and to base the experiment on an achievement test of the type used throughout the military technical training field. In this case then the testees were part of the target population for which the testing system was being developed and the test was one which the testees were expected to finish.

The third experiment used the repeated measure approach based on a free strategy test followed by a forced serialist strategy retest. The sample used for the experiment was 39 apprentice electronics technicians. The test used was a 40 item MCCQ test, the items having been selected, in accordance with the test specification for the topic area chosen (Direct Current theory), from a validated bank of questions. A self reporting technique was used to gather data from which strategies could be identified. Strategies were identified by plotting the sequence in which questions were answered. Details of the self reporting sheet and graphs of examples of each category of strategy are at Annex A along with the results of the experiment.

The analysis of the results was concentrated on the two extreme subgroups as identified by the first application of the test; those testees who used a wholly serialist strategy and those who preferred a highly reiterative approach.

The statistics calculated for the whole group performance during the experiment are summarised in table 3, the calculations being based on scores calculated from percentage correct.

Table 3 Whole Group statistics

| | TEST | RETEST |
|---|---|---|
| Number | 39 | 39 |
| Mean Score | 62.95 | 63.58 |
| Range | 43–85 | 40–85 |
| SD | 10.97 | 10.95 |
| Kuder Richardson 20 reliability | 0.62 | 0.63 |

The descriptive statistics give no indication that the imposition of a serialist strategy for the majority (69%) of testees who had preferred, to varying extents, a different strategy. The Kuder Richardson 20 reliability calculation for both test and retest produced an acceptable figure for this kind of test. There was therefore tentative evidence that the imposition of a fixed answering strategy had no significant effect on the performance of testees on this achievement test.

The analysis was continued for the two extreme subgroups, the serialist group effectively becoming a control group. The statistics calculated are given in table 4.

Table 4 Subgroup statistics

| Subgroup | Reiterators | | Serialists | |
|---|---|---|---|---|
| Number | 12 | | 12 | |
| | Test | Retest | Test | Retest |
| Mean Score | 66.08 | 65.0 | 59.92 | 61.42 |
| Range | 50 - 85 | 55 - 80 | 43 - 78 | 50 - 85 |
| Mean Rank | 16.25 | 16.67 | 21.83 | 20.67 |
| Range of Ranks | 1 - 32 | 4 - 30 | 3 - 38 | 1 - 51 |

Once again inspection of these descriptive statistics gives no indication of any differences. This initial impression was investigated to ascertain whether:

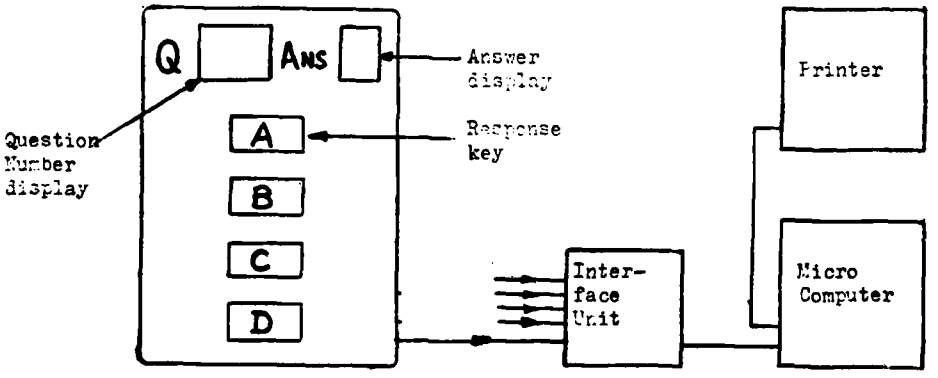
a. There was any significant difference between the performance of the two subgroups within the first and second application of the test.

b. There was any significant difference in the performance of each subgroup between the first and second applications of the test.

The details of the calculations are contained in Annex A. The outcome of the calculations was that there were no statistically significant differences.

This outcome was not unexpected considering the errors of measurement involved in achievement tests of this nature. However, it does provide sufficient justification for the development of a prototype system which can be used for further investigation of the concept. The prototype system to be developed will be similar to that illustrated in Figure 1.

Figure 1. Prototype data collection system.




The software system supporting the hardware illustrated in Figure 1 will assign the number 1 to the question number display of each testee and then poll each data pad in turn searching the Answer display buffer until an entry, made by the testee pressing the appropriate key, is found. The software will then store the response in the microcomputer, clear the display and assign the number 2 to the question display. This will continue for each testee until the final question number is reached.

The facilities considered and rejected were the provision of a "don't know" button and an "enter" button. The "don't know" button was rejected in the first instance because the literature concerning the development of multiple choice tests items tends not to support this technique. Although the purpose of the "don't know" key is in this instance different the evidence from the experiment shows that recourse to such a key would have been made for approximately 12% (192 of 1579 responses) of the key depressions. Combining these factors with the evidence from the experiment it was decided not to offer this facility. Similarly the provision of an "enter" key was considered. Again evidence from a slightly different context (CAL experiments) tends to favour a one key rather than two key entry system.

The evidence from these experiments shows that different strategies are used by testees when attempting MCQ achievement tests. The results of the experiment suggest that the use of these strategies does not contribute significantly to success on tests. Hence, a single serialist strategy can be imposed without significantly influencing the performance of those testees whose behaviour is altered. The experimentation will continue using a prototype system to investigate the degree of user acceptance of the data input device. Should the system prove acceptable to testees and trainers then the indications are that the interface and keypads can be made available for approximately £350. The major benefit of this will be that the processing power already available can be used to improve the validity and reliability of the tests used in training by making the means to monitor these factors much more widely available.

RESULTS OF AN EXPERIMENT TO INVESTIGATE THE EFFECT OF AN IMPOSED ANSWERING STRATEGY ON PERFORMANCE ON AN MCOQ ACHIEVEMENT TEST.

Design

1. The experiment used a repeated measures approach. The first administration of the test instrument was combined with a self reporting system. The data collected on the self reporting form was used to discover the preferred answering strategy for each testee. Those testees identified as preferring the serialist strategy which was to be imposed in the retest situation then became the control group. The second administration of the test instrument was used to measure the performance of testees under an imposed answering strategy. The imposed strategy was serialist; each testee was forced to respond positively to each question in turn, in his own time, with no recourse to a subsequent attempt on any question. In order to keep the experimental conditions as close as possible for each administration of the instrument a self reporting system was also used during the second administration. In this case testees were asked to attribute confidence levels, on a three point scale, to the answers they had selected.

Instrument

2. The testing instrument used for the experiment was a 40 multiple choice item test, each item having 4 responses. The questions were selected from a live item bank used for monitoring the progress of apprentices throughout a 3 year military and technical training course. All the questions had been validated on the population of which the testees were a sample. The test covered four topic areas within the subject of Direct Current theory and tested knowledge, application and comprehension.

Method

3. Test. The testees were briefed on the purpose of the self reporting sheet and instructed in its completion. They were then given the test paper and a response card and allowed to attempt the test in their own way within a time limit of 35 minutes. (It should be noted that a more typical time allowed would be 30 mins, extra time was allowed to compensate for the self reporting system).

4. Retest. Testees were once again briefed on the self reporting system. They were then briefed on the imposed answering strategy. Each testee was given the complete test paper and a response card and allowed 35 minutes to complete the test, compliance with the imposed strategy was monitored by close invigilation.

5. Scoring. The test was scored on the basis of percentage correct, no corrections were made for guessing. The results of the experiment are detailed below.

Results

6. Answering strategies. The data on the self reporting forms for the test situation was graphed in order to allocate a preferred strategy to each testee. An example of the self reporting sheet is in figure 1A below. It can be seen from Figure 1A that this testee included element of each phase; initial, review and revise. The initial phase included responses to all questions except 1, 2, 11 and 36 and also included a revision of 27. The review phase then consisted of responses to 1, 2, 11, and 36. Graphs were plotted for each testee. Examples of the graphs are at Appendix 1 to this Annex. From the graphs testees were allocated to categories as follows.

a. Serialist - A strategy which included only one initial phase in which each question was answered in turn.
b. Predominently Serialist - A strategy which included less than five entries in the review and/or revise phase.

c. Partially Reiterative - A strategy which included 5, 6 or 7 entries in the review and/or revise phase.

d. Highly Reiterative - A strategy which included at least 8 entries in the review and/or revise phase.

The breakdown of allocations is given in table 1A.

Figure 1A. Self Reporting Sheet - Answering strategies.

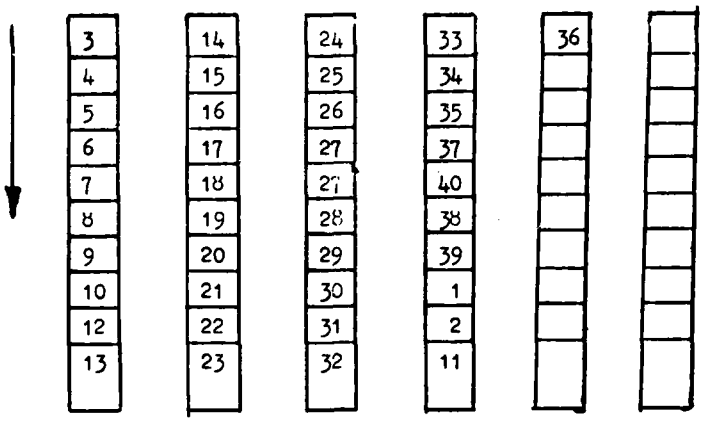



Table 1A Preferred answering strategies

| STRATEGY | S | PS | PR | R |
|---|---|---|---|---|
| NUMBER | 12 | 10 | 7 | 12 |

7. Performance. The results of the two administrations of the test are at Appendix 2 to this annex. The analysis of the results was concentrated on the two extreme subgroups S and R whose results are given in tables 2A and 3A respectively.

Table 2A Performance by subgroup S(Serialist).

| | PERCENTAGE SCORE | | | RANK | |
|---|---|---|---|---|---|
| TESTEE | TEST | RETEST | DIFFERENCE | TEST | RETEST |
| 1 | 43 | 50 | +7 | 38 | 37 |
| 4 | 63 | 60 | -3 | 20 | 19 |
| 16 | 78 | 85 | +8 | 3 | 1 |
| 18 | 68 | 63 | -5 | 14 | 19 |
| 19 | 48 | 60 | +12 | 34 | 19 |
| 21 | 70 | 60 | -10 | 10 | 19 |
| 23 | 73 | 55 | -18 | 6 | 30 |
| 24 | 58 | 60 | +2 | 26 | 19 |
| 29 | 65 | 75 | +10 | 17 | 8 |
| 31 | 43 | 53 | +10 | 38 | 33 |
| 33 | 60 | 63 | +3 | 24 | 15 |
| 35 | 50 | 53 | +3 | 32 | 33 |
| MEAN | 59.92 | 61.42 | | 21.83 | 20.67 |
| RANGE | 43 - 78 | 50 - 85 | | 3 - 38 | 1 - 37 |

Pearson's $Q_3$ estimate of test - retest reliability for this subgroup $r_{tet}$ = .64

Table 3A Performance by subgroup R (Reiterators).

| | PERCENTAGE SCORE | | | RANK | |
|---|---|---|---|---|---|
| TESTEE | TEST | RETEST | DIFFERENCE | TEST | RETEST |
| 5 | 70 | 75 | +5 | 10 | 8 |
| 7 | 58 | 55 | -3 | 26 | 30 |
| 9 | 78 | 68 | -10 | 3 | 12 |
| 12 | 50 | 58 | +8 | 32 | 27 |
| 14 | 70 | 68 | -2 | 10 | 12 |
| 15 | 65 | 60 | -5 | 17 | 19 |
| 17 | 63 | 55 | -8 | 20 | 30 |
| 22 | 73 | 73 | 0 | 6 | 10 |
| 27 | 60 | 60 | 0 | 24 | 19 |
| 30 | 85 | 80 | -5 | 1 | 4 |
| 34 | 58 | 65 | +7 | 26 | 14 |
| 38 | 63 | 63 | 0 | 20 | 15 |
| MEAN | 66.08 | 65.0 | | 16.25 | 16.66 |
| RANGE | 50-85 | 55-80 | | 1 - 32 | 4 - 30 |

Pearsons $Q_3$ estimate of test - retest reliability for this subgroup $r_{tet}$ = .94

8. Comparison of Groups R & S. The Mann-Whitney U Test was used to compare the performance of the two subgroups both in the test situation, where different strategies were being used, and the retest situation. The results are summarised in table 4A.

Table 4A Comparison of group performance.

| | TEST | RETEST |
|---|---|---|
| Mann-Whitney U | 53 | 50 |
| Standard Error Z | 1.09 | 1.27 |
| Significant (p 0.01) | NO | NO |

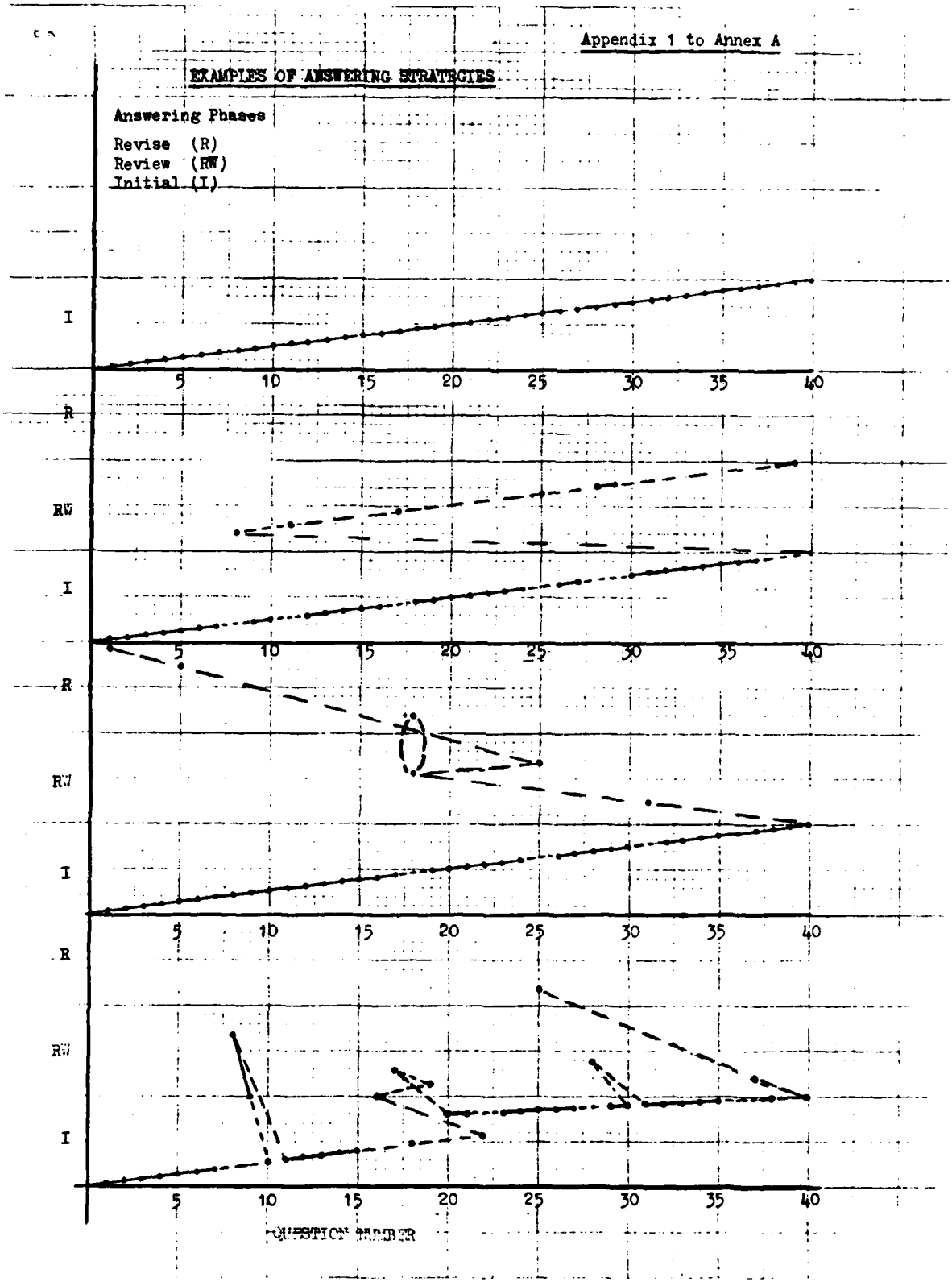
The performance of each group on test - retest was then compared using the Wilcoxon (related samples, signed difference) Test. The results are summarised in table 5A.

Table 5A Test - Retest comparison by groups.

| | SERIALIST | REITERATORS |
|---|---|---|
| NUMBER | 12 | 9 |
| WILCOXONS T | 36 | 20 |
| SIGNIFICANT (p 0.01) | NO | NO |

The difference in performance was not statistically significant for either group.

Appendix 1 to Annex A



Group Experimental Results

| TESTEE | TEST SCORE % CORRECT | RETEST SCORE % CORRECT | TEST RANK | RETEST RANK | ANSWERING STRATEGY |
|---|---|---|---|---|---|
| 1 | 43 | 50 | 38 | 37 | S |
| 2 | 53 | 63 | 30 | 15 | - |
| 3 | 70 | 75 | 10 | 8 | R |
| 4 | 63 | 60 | 20 | 19 | S |
| 5 | 80 | 83 | 2 | 2 | - |
| 6 | 73 | 80 | 6 | 4 | - |
| 7 | 58 | 55 | 26 | 30 | R |
| 8 | 48 | 53 | 34 | 33 | - |
| 9 | 78 | 68 | 3 | 12 | R |
| 10 | 58 | 43 | 26 | 38 | - |
| 11 | 45 | 40 | 37 | 39 | - |
| 12 | 50 | 58 | 32 | 27 | R |
| [illegible] | 65 | 58 | 17 | 27 | - |
| [illegible] | 70 | 68 | 10 | 12 | R |
| 15 | 65 | 60 | 17 | 19 | R |
| 16 | 78 | 85 | 3 | 1 | S |
| 17 | 63 | 55 | 20 | 30 | R |
| 18 | 68 | 63 | 14 | 15 | S |
| 19 | 48 | 60 | 34 | 19 | S |
| 20 | 78 | 78 | 3 | 6 | - |
| 21 | 70 | 60 | 10 | 19 | S |
| 22 | 73 | 73 | 6 | 10 | R |
| 23 | 73 | 55 | 6 | 30 | S |
| 24 | 58 | 60 | 26 | 19 | S |
| 25 | 48 | 53 | 34 | 33 | - |
| 26 | 68 | 60 | 14 | 19 | - |
| 27 | 60 | 60 | 24 | 19 | R |
| 28 | 63 | 73 | 20 | 10 | - |
| 29 | 65 | 75 | 17 | 8 | S |
| 30 | 85 | 80 | 1 | 4 | R |
| 31 | 43 | 53 | 38 | 33 | S |
| 32 | 68 | 60 | 14 | 19 | - |
| 33 | 60 | 63 | 24 | 15 | S |
| 34 | 58 | 65 | 26 | 14 | R |
| 35 | 50 | 53 | 32 | 33 | S |
| 36 | 70 | 78 | 10 | 6 | - |
| 37 | 73 | 83 | 6 | 2 | - |
| 38 | 63 | 63 | 20 | 19 | R |
| 39 | 53 | 58 | 30 | 27 | - |
| MEAN | 62.95 | 63.58 | 18.97 | 18.77 | |
| RANGE | 43 - 85 | 40 - 85 | | | |
| SD | 10.97 | 10.95 | | | |
| KR 20 | 0.62 | 0.63 | | | |

AD P001382

# SIMULATION OF COMMAND GROUP OPERATIONS : AN EVALUATIVE REPORT

Authors: Major John P Taylor MSc and Major Robert R Begland PhD.,
Training Consultants, Army School of Training Support, England.

## SUMMARY

Technological advances in computer hardware and software have greatly improved the ability to simulate the complex information systems on which commanders depend to control warfare. The ability to generate data in real time introduces the capability to exercise both the individuals and the group in decision making in situations which approximate to the stressful conditions of war. The Battle Group Trainer (BGT) was developed to provide battalion and regimental commanders and their staffs with the opportunity to train collectively for their role. The BGT design included a high degree of environmental realism. However, the changing nature of the battlefield and the lack of recent experience of mechanised warfare makes the definition of the tasks to be carried out one of conjecture. To evaluate a Command group trainer it is therefore necessary to establish how accurately the skills used during training and the tasks for which these skills are used approximate to the "real" situation. Three perspectives of the "truth" are considered in an attempt to produce the criteria against which the BGT can be evaluated.

## INTRODUCTION

The problem with having highly sophisticated technological capabilities is that you have to find uses for them. The old "Law of the Hammer" seems appropriate; "If you give a child a hammer he will eventually find something that requires pounding"

The utilization of computer simulation as a means of training, rehearsing, testing, assessing, and playing with command groups is a commonly accepted practice. The degree of fidelity (both physical and task based) can vary from low to high, with various effects upon training efficiency and effectiveness. Yet the one area that seems to elude quantification is what should be taught (experience) and how can computer simulation maximize that experience.

There is a substantial amount of subjective data to indicate that particpants who experience a command and control computer simulation trainer really enjoy it and believe it to be a worthwhile experience. However, from an empirical perspective there is scant research to quantify the nature and merit of this type of training.

The essence of this problem is found in both the lack of a strong analytically derived set of tasks for the command and staff of an operative Battle Group HQ (BGHQ), and in the commensurate training design for the acquisition and maintenance of these same tasks.

In order to conduct an evaluation of such a computer supported simulation trainer, it is essential to be able to separate fact from fiction, truth from hearsay, and objective from subjective. The basis of such an evaluation must reside in the definition of the performance (operational) requirements of the members of the command group to be trained. Both individually and collectively the task type behaviours must be established.



PREVIOUS PAGE IS BLANK

This is not a new problem. Xenophanes, many centuries ago, stated the problem thus:

> "Let us conjecture that this is like the truth, but as for certain truth no man knows it. ..... And if by chance he were to utter the final truth he himself would not know it for all is but a woven web of guesses".

Acknowledging the weaknesses and strengths of any one analytic approach and the complexity of the interactions within the BGHQ it was decided to approach the problem from three different perspectives. The analytic approaches developed and described are an attempt to capture the reality/truth of what a BGHQ is required to do in war.

## THE BATTLE GROUP TRAINER

The BGT was designed to subject the commanding officer and his staff to stressful situations, relating exercises to real ground and portraying battle conditions which are as realistic as possible within a realistic time frame. The training system can best be described as a combination of a tactical exercise without troops (TEWT), a command post exercise (CPX) and a war game. The trainer has three main components; the BGHQ, the control room and a tactical HQ.

The BGHQ is very much like a theatrical set. It uses actual vehicle bodies and realistic mock ups to portray the situation of a BGHQ which has taken up temporary accommodation in an old farm and outbuildings. The "set" produces, as much as is possible, the conditions that the commanding officer and his staff would expect to work under in the field, in terms of varying light levels and the background noise of battle. The information which flows into the HQ through simulated radio networks is generated in the control room.

During the TEWT phase of the training period the commanding officer makes his plan, issues his orders and co-ordinates compliance with his plan. All of this activity takes place over real ground near the trainer. The second phase of the exercise is to "fight" the battle using a computer supported simulation in the control room. The main tool used during the simulation is a master map board. The map (scale: 4 cm represents 100m) is very detailed, showing features down to individual hedgerows and ditches. Own and the enemy's men, individual vehicles and major equipments are represented by symbols on the map. Contouring is shown by coloured layering and, as an aid to intervisibility, valleys and ridges are highlighted by coloured lines. Around this map sit the commanders of all the sub units involved in the battle, (company commanders, reconnaissance troop commander etc). These people "fight" the battle in respect of their own troops and report to, and seek guidance and assistance from BGHQ staffs via the simulated communications system. The play of the battle is free to the extent that the actions and reactions in the battle depend, on the one hand, on the commanding officer's plan and the way he and his subordinate commanders carry out this plan and on the other hand, upon the result of each and every individual sighting, engagement or movement. The computer sub-system is used to generate the combat information by assessing the likelihood of sightings, of detection and the outcome of engagements between opposing units. Engagements are assessed using a data base of hit and kill values which take into account the range and circumstances of opponents as well as the weapon types. Engagements can be as simple as one on one, one on many or many on many. The computer sub system also accounts for ammunition, records battle casualties and assessed the effects of artillery, mortar, helarm, fighter ground attack and air defence weapon systems. The simulation is therefore used to generate

the information used by the BGHQ to control the battle and is, except in the circumstances described for the tactical HQ, closed to the BGHQ staff.

The tactical HQ is a simulation, again using vehicle bodies and theatrical set which allows the commanding officer to leave his main HQ during the battle. All the normal communications are provided and a combination of slides and close circuit television, focussed on the map board, is used to allow the CO to see the battlefield as he "moves".

Overall the environmental realism is achieved by theatrical means whereas the realism of the information generated during the simulation is a consequence of the attention paid to timings, movement, intervisibility, the chance of sightings, the application and assessment of direct and indirect fire, accounting for ammunition and battle casualties and the use of realistic enemy tactics. The emphasis during the design process was therefore on environmental and information realism. However, the training benefits which accrue from simulations are dependent upon task fidelity as much as realism. No matter how scientifically the information is generated during the simulation process, if the information is being used in inappropriate ways and causing the development of disfunctional behaviour then the expenditure of resources on the training system is inefficient.

## THE PROBLEM

Traditionally the Commanding Officer of a battalion or regiment has been given a great deal of freedom to run his unit in his own way. The emphasis in any evaluation of the units, and consequently the CO's, readiness for combat has been summative. Hence, although guidelines are provided in field manuals and personnel are established to man the HQ in accordance with the guidelines there are as many variations on the theme as there are commanding officers. The organisation of the BGHQ is further influenced by the operational role of the unit providing the HQ. Armoured regiments are established differently from mechanised battalions in terms of personnel and vehicles. A move from the British Army of the Rhine to a station in the United Kingdom will also cause changes in emphasis. Finally the organisation of the BGHQ reflects the CO's own experience and his perception of the level of training of the individuals who become part of his HQ on operations. The tasks to be carried out in a BGHQ, both individually and collectively, are therefore a function of the organisational structure of the HQ, the level of training of the individuals who constitute the HQ staff and the operational role of the battle group. In terms of any individual within the HQ, the tasks he will have to perform will be an amalgam of those things demanded of him by the CO through giving him in a pratical role within the HQ, the tasks forced upon him by his involvement in a particular type of operation, and his background and training. The analysis of the operational working of a BGHQ to establish the individual and collective task type behaviours must therefore take into account these factors. For this reason it was thought to be necessary to study the BGHQ from three perspectives, the combination of which would give a data base from which the "truth" would emerge.

## THE ANALYTICAL PERSPECTIVES

The perspectives chosen from which to view the BGHQ were those of:

a. the Commanding Officer

b. the Training system

c. the Operational Role.

The basis for this choice was a decision to treat the BGHQ as an information processing/decision making organisation made up of several subsystems, each manned by personnel who ideally will have been trained to occupy positions in the subsystem, the system working within the environment of an operational role. The model is illustrated in Figure 1.

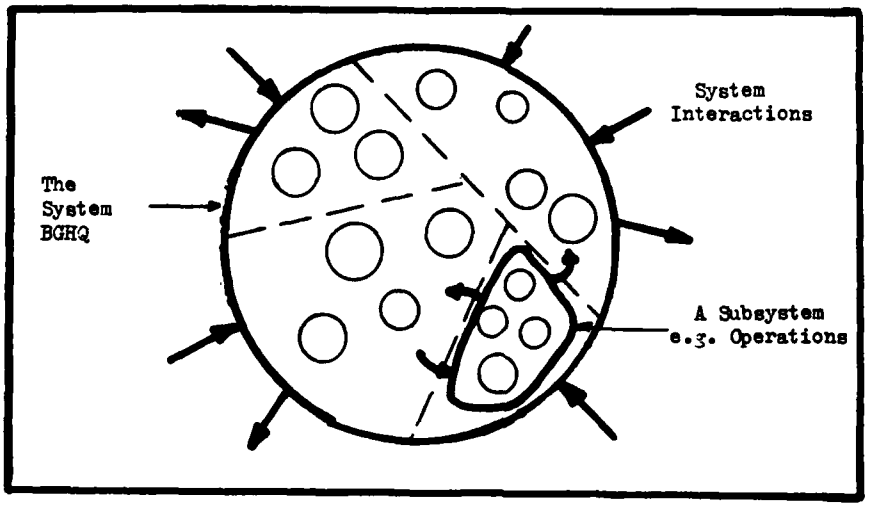



FIGURE 1. MODEL OF BGHQ

The environment provided by the operational functioning of the battle group is the source of all the information used by the BGHQ in its decision making processes. It contains the enemy, the forces under direct control of the BGHQ, flanking and higher formations. The BGHQ has only partial control over its environment and will therefore be driven, to a certain extent by the environment. The BGHQ consists of the Commanding Officer and his staff. The constituents of the system will be dictated partially by the environment in that it will influence the grouping of the battle group as a whole, hence the staff available to the CO. However, for a given operation the constituents of the HQ will be fixed and within the CO's influence. His appreciation of the capabilities of the personnel available and his concept of operating procedures will influence his organisation of the BGHQ. The subsystem boundaries (indicated by broken lines) will therefore change depending upon the CO. The subsystem functions and the personnel allocated to them will vary and will partially be a consequence of the skills and knowledge each individual brings to the system. The skills and knowledge possessed by each individual will be a consequence of his experience and training and will therefore frelect the current objectives within the training establishments. The trainers will therefore have a perception of the needs of individuals who are being trained to fill roles in an operational BGHQ. The multi-analytic perspective approach should therefore allow the truth of the operational role of the BGHQ to be approached by considering:

a. what is forced upon the system (Operational perspective)

b. what is expected of the system (CO's perspective)

c. what the components of the system are capable of doing (Training perspective).

## THE OPERATIONAL PERSPECTIVE

This analytic approach uses the Battle Group Trainer's Scenarios in an attempt to zero in on the actual operations that do or should occur in a Battle Group Headquarters both before and during battle. These battle group operations are to be defined and described in terms of the various functions (command and staff) that are performed by the members of a BGHQ organisation at Battalion level.

The procedure to be employed involves a series of analytic steps that when followed will progress from a specific activity/event in a given tactical scenario to the various functions and steps that would be performed by the staff of a BGHQ as a direct/indirect result of that event.

The series of analytic steps begins with a chronological analysis of each scenario. During this analysis each event or possible event is listed and the input stimulus is described. Additionally, a more precise description is produced for each event. Subsequent to the chronological analysis, a Functional Analysis is conducted. The functional analysis attempts to examine each event in terms of the various functions that relate to that event, who has responsibility for action for each of the components of the event, and proposes a desired response for each, along with an indication of how each component would be measured/critiqued.

Through this functional approach to establishing the operational roles performed by the members of the BGHQ, the individual and collective performance requirements can be produced. Once this level of specificity is achieved, the staff of the BGT will then attempt to prioritize all of the diverse events/functions according to their relevance/importance. Each event (function) is examined for its importance to the successof the mission of the Battle Group, and sorted into one of the following categories:

a. Critical Combat Behaviour

b. Relevant but less important Combat Behaviour

c. Unimportant Behaviour

Having sorted the events/functions into these 3 categories, the BGT staff will then review each of its scenarios in detail, examining each event/function that has been previously categorized, and make a determination as to the absence/presence of that function and the opportunity for realistic "play" given the present scenario. Through this Scenario Analysis, the degree of incorporation of Critical Combat Behaviours for each scenario can be established.

This multi-stage analytic approach has been developed as a way of examining each scenario presently used in the BGT, and to compile detailed information on the inner workings of a BGHQ.

The Battlegroup Trainer Staff has been selected as the most qualified available group of officers, who are able to provide insight into these staff and command innerworkings. The staff of the two BGTs (both in UK and Germany) represent a multi-arm perspective (Infantry, Armour, Artillery, etc) that over several years has seen a variety of BGHQ go through these Trainers. These officers, through their prior training and experience as BGT staff are uniquely qualified to, as a whole, speak to the operational requirements of a battlegroup under the variety of combat situations that can confront a unit. Thus by examining the events/functions that can occur in the various scenarios used in the two BGT's, it is possible to produce a listing by scenario of what the total play of the problem could entail. Similarly, by reformatting the data, it is possible to extract every operational requirement for a given staff position (e.g. CO, Operations, Intelligence, etc) and to start examining the

potential training requirements for that staff position.

The method of working for the analysis involves joint activity by the multi-arm team working chronologically through a battle group operation from receipt of mission through appreciation, planning, preparation and execution phases to completion of operation. The team uses the questions "What happens next?" and "What happens if?" to produce the stimuli for BGHQ activity. Each stimulus (input) will set in train a group of events to be actioned. The team will use the proforma illustrated in Figure 2 (not to scale) to record the activities produced.

| SCENARIO 1B | SITUATION POSITIONAL DEFENCE | | SHEET 13 |
|---|---|---|---|
| INPUT | ACTION | FUNCTIONAL AREA | EVENT DESCRIPTION |
| BG Comd Orders | Anti tank Pl. Comd | Fire Support (Anti tank) | Co-ordinate the employment of BG Anti tk resources |
| DESIRED RESPONSE | | AS MEASURED BY | REMARKS |
| 1. Interpret COs guidance<br>2. Conduct Recce<br>3. Coordinates with Ops<br>4. Develop plan<br>5. Coordinates/obtains approval<br>6. Issues plan<br>7. Monitors compliance | | 1. Anti tank defences are integrated with total plan of defence<br>2. ..... etc<br>3. | |

FIGURE 2. THE OPERATIONAL PERSPECTIVE ANALYSIS PROFORMA

The input "BG Comd Orders" will also be the stimulus for other events in the functional area of fire support and will involve action by the Battery Comd and others. The functional analysis will extract all event descriptions with a given label in the functional area column. Similarly the tasks of each individual can be identified by sorting on the action column. As this column will also record joint action the collective events will also be indentifiable. These proformae and the consequent analyses will then form the basis for the subsequent steps into critical combat behaviour and scenario analysis.

## THE COMMANDING OFFICER'S PERSPECTIVE

This investigation is particularly concerned with the BG commander's perception of how his BGHQ operates. It will be restricted to the activities within the BGHQ and the relationships within it. The model used will be to consider the BGHQ as an information processing/decision making organisation.

Previous investigations which have taken the information processing approach have identified five categories of activities which organisations must carry out in order to be successful. These categories are:

a. Acquiring Information

b. Disemminating Information

c. Decision making

d. Disseminating Instructions and Orders

e. Monitoring compliance with orders

The key category is, of course, decision making. The activities in the other categories are necessary to provide the information from which decision can be made and thus ensure the execution of decisions made.

For this investigation the categories Acquiring Information and Monitoring have been combined as activities resulting in information coming in to the BGHQ. Similarly the activities of Disseminating Information and Disseminating Orders have been combined as activities which result in information leaving the BGHQ. The study is therefore concerned with the categorisation of BGHQ activities into those which involve the following:

a. the receipt of Information

b. decision making/problem solving

c. the transmission of orders/instructions/information

In order to produce information from which the realism of activities at BGT can be assessed supplementary information will also be collected. For activities categorised as involving the Receipt of Information the supplementary information required concerns the channels through which the information is available to the CO and the means used to acquire it. For Decision making activities further details of the decision making process is required. Although all decision are ultimately the responsibility of the CO there are degrees of involvement and co-operation. The study is attempting to identify those decision making activities which are the sole concern of the CO, those involving co-operation between the CO and other members of the BGHQ and finally those delegated to a member of the BGHQ. For those activities concerned with the Transmission of Information/Orders the only supplementary information required is the means used for transmission.

To assist in the collection of the information required a set of activities has been identified. The sources used to compile this list were examples of Regt/Bn SOPs, field manuals and reports on the evaluation of other training systems.

The list of activities has been split into phases which cover initial appreciation, battle preparation, combat, BG Security and special situations. Some activities appear in more than one list in an attempt to reflect the cyclic nature of much of the decision making/problem solving carried out in BGHQ. The allocation to phases may be contentious, however, debate on this is not at issue; the reasons for the subdivision is to reduce the fairly extensive list to more easily handled sections.

The analysis will use a Delphi technique, the initial phase being to interview experienced battle group commanders who have previously completed the proforma illustrated in figure 3.

| ACTIVITIES | Information received | | Decision making | | | Information transmitted | Notes |
|---|---|---|---|---|---|---|---|
| | Appoint-ment | Means | CO only | CO plus | Delegated | Means | |
| 4. Control Deployment | | | | | | | |
| 4.1. Maintain Communications with BG elements | | | | | | | |
| 4.2 Inspect preparations | | | | | | | |
| 4.3 ..... ..... | | | | | | | |

FIGURE 3. EXTRACT FROM CO's PERSPECTIVE DATA COLLECTION PROFORMA

The proforma contained 81 activity statements. The interview will be used to investigate the completeness of the activity list and to even out the activity statements so that each activity will result in only one category. The responses from COs will then be consolidated into an overall picture before being offered back to the group. The process will continue until an acceptable activity list is produced. The analysis will then continue to identify the individual and collective activites which can be combined with the information from the operational perspective analysis and the training perspective analysis.

## THE TRAINING PERSPECTIVE

This investigation will concentrate on the individual appointments which are established within the BGHQ (Ops Officer, Adjutant, Battery Commander, Intelligence Officer etc) and on the training given to personnel who fill the appointments. The study will consist primarily of a literature search which will involve parallel studies of data from each of the training establishments involved in the training of BGHQ personnel. The literature used will include:

a. job analysis data
b. course entry criteria
c. course training objectives
d. course programs
e. on job training requirements etc

This data will be used to produce a "pen picture" of the personnel being trained to fill each appointment. The pen picture will identify the backgroun experience, including any relevant previous courses attended, the skills and knowledge trained during the particular course attended and the related experience that the individual is likely to gain between completion of the preparatory course and being involved as a member of a BGHQ.

The information collected from those parallel studies will be combined to produce a composite picture of the skills and knowledge available to the CO and will be used to cross reference the information gathered from the other two perspectives described above.

## EVALUATION

The combined data from the multi-analytic study should provide information which can be used to evaluate the existing BGT system. From the complete list of activities established it will be possible to categorise them into:

a. Activities practiced in all exercises at the BGT

b. Activities practiced in some exercises at the BGT

c. Activities not practiced at the BGT

For those activities practiced in all exercises there will be further information on the fidelity of the activity and its criticallity. This information can be used to assess the benefits of improving the way in which the system facilitates these activities.

For those activities practiced only in some of the exercises there will be information on whether this is simply a function of that particular mission or is an oversight. Again the criticallity of the activity will be known and recommendations can be made as to whether the activity should become common and if so how it should be implemented.

Similarly the activities not practiced at the BGT can be assess in terms of criticallity. Obviously those activities deemed to be critical but not practiced will have a large influence on the measure of the value of the BGT system. It is believed that the multi-perspective approach will therefore provide a sufficiently comprehensive model and set of decision criteria against which the BGT system can be compared in order to make an evaluation of the system as an effective and efficient training method. The study will continue to a comparative analysis range (vis a vis CPX and FTX) before making recommendations on the future utility of the system.

## BIBLIOGRAPHY


BARBER H.F., McGREEN J.F. and STEWART S.R. The Computer Assisted Map Manouvre System : A Preliminary Examination of its Training Effectiveness and its Use as a Research Vehicle (Research Memo 79-9) US Army etc. June 1979

HAYES-ROTH, B. Projecting the Future for Situation Assessment and Planning: A Cognitive Analysis (Interim Report N-1600-AF). Washington, D.C.: Requirements, Programs and Studies Group, HQ USAF, November 1980.

HAYS, R.T. Simulator Fidelity: A Concept Paper (ARI Tech Rep 490). Alexandria, VA.: US Army Research Institute for the Behavioural and Social Sciences, November 1980.

HULIN, C.L. and ROUSSEAU, D.M. Analysing Infrequent Events (Tech. Rep. 80-3). Arlington, VA.: Organisation Effectiveness Research Programs, Office of Naval Research, April 1980.

JONES, D.R., et al. Battle simulation board games : An analysis in terms of design characteristics and leader skills. (Research Note 80-2) Kinton Inc. Alexandria, VA.: US Army Research Institute for the Behavioural and Social Sciences, January 1980.

KAPLAN, I.T. and BARBER, H.F. Evaluation of a Computer - Assisted Battle Simulation: CAMMS versus a CPX (ARI Tech. Paper 355). Alexandria, VA.: US Army Research Institute for the Behavioural & Social Sciences, April 1979.

KAPLAN, I.T. and BARBER, H.F. Training Battalion Command Groups in Simulated Combat: Identification and Measurement of Critical Performances (ARI Tech. Rep. 376). Alexandria, VA.: US Army Research Institute for the Behavioural and Social Sciences, June 1979.

LACKEY LL., DELUCA A.J. and TREMBLE T.R. Decision Making and Training Techniques for Command and Control Systems Part III. Human Resources Research Organisation Alexandria VA. June 1975.

RAY, R.L., NELLIS, M.J., and EMURIAN, H.H. Event Time-Series Applications to the Analysis of Behavioural Events (Tech. Rep 1). Arlington, VA.: Organisational Effectiveness Research Programs, Office of Naval Research, January 1981.

SHRIVER, E.L., JONES, D.R., HANNAMAN, D.L., GRIFFIN, G.R., and SULZEN, R.H. Development of Small Combat Arms Unit Leader Tactical Training Techniques and a Model Training System (ARI Research Rep. 1219). Alexandria, VA.: US Army Research Institute for the Behavioural and Social Sciences, July 1979.

TIEDE, R.V., BURT, R.A., and BEAN, T.T. Design of an Integrated Division Level Battle Simulation for Research, Development, and Training: Vol 1 (ARI Tech. Rep. TR 420). Alexandria, VA.: US Army Research Institute for the Behavioural and Social Sciences, August 1979.

TREMBLE T.R. and COSTNER R.S. Information Flow in Training Exercises with the Combined Arms Tactical Training System. US Army Research Institute for the Behavioural and Social Sciences, November 1977.

WHEATON, G.R., ROSE, A.M., FINGERMAN, P.W., LEONARD, R.L. Jr., and BOYCAN, G.G. Evaluation of the Effectiveness of Training Devices: Validation of the Predictive Model (ARI Tech. Rep. TR76-A2). Alexandria, VA.: US Army Research Institute for the Behavioural and Social Sciences, October 1976.

AD P001383

# Shortening of Defense Language Aptitude Battery

John W. Thain
Education Specialist
Defense Language Institute
Foreign Language Center

The Defense Language Institute, Foreign Language Center (DLIFLC) is the proponent for the Defense Language Aptitude Battery (DLAB), used in screening recruits for aptitude to learn foreign languages. DLIFLC is attempting to shorten DLAB without decreasing its reliability or validity. Item analysis data was used to plan eleven possible shortening strategies with different number of items deleted from original test. One hundred sixty-four answer sheets were rescored and recorrelated with a foreign language course grade after items had been deleted according to each strategy. Tests from another sample will also be rescored and recorrelated with the criterion using each of the test shortening strategies in order to cross-validate the original results.

# SHORTENING OF THE DEFENSE LANGUAGE APTITUDE BATTERY

John W. Thain
Education Specialist
Defense Language Institute
Foreign Language Center

Background.

The Defense Language Aptitude Battery (DLAB) is administered to determine if enlisted personnel should be programmed for language training. Previous studies by the Defense Language Institute Foreign Language Center (DLIFLC) have shown the correlation between DLAB scores of students selected for language training and their course grades in DLIFLC language courses to be consistently around .40; if this correlation is corrected for restriction of range, it rises to around .60. In recent years, only about 20% of the recruits taking the test have achieved a passing score. The test takes 90 minutes to administer and consists of 119 multiple-choice test items.

The DLAB is administered at several locations; the most important of which are the local Armed Forces Entrance and Examination Station (AFEES) throughout the country, and Lackland Air Force Base in San Antonio, Texas (for Air Force enlisted men). The agencies responsible for administering DLAB at those locations are also responsible for screening recruits for most of the other occupational specialities in the Services. Largely on grounds of administrative convenience, these agencies favor any proposal that would shorten the process of programming enlistees to their subsequent assignments. In particular the AFEES have expressed interest in reducing the current administration time of 90 minutes required to administer DLAB. There are two specific reasons for this interest in shortening DLAB:

- A shorter test would contribute to the shortening of the overall screening process and thus help avert the possibility that the AFEES (Armed Forces Entrance and Examination Stations) would have to pay the overnight expenses of recruits if the processing of the recruits had to be extended an additional day.

- A minority of students taking DLAB pass the test. From the point of view of test examinees, every minute spent by failing examinees on DLAB is wasted.

Of course, shortening a test tends to decrease test reliability and validity. Using a less valid and reliable version of a test will result in less effective screening of potential students and lead to lower student performance and increased student attrition at DLIFLC.

Therefore, DLI designed a research study to determine whether DLAB could be shortened without substantially reducing test validity and reliability.

Procedure.

Two methods of shortening the test were considered - condensing redundant instructions or by deleting poorly functioning test items. At the very start of the project we needed to analyze in detail the factors contributing to length of test administration.

The following table breaks down the time required to administer DLAB. The total time required to administer each part of the test and the test as a whole is given in minutes and seconds. The average time required to administer each item in each part is computed by dividing the time required to administer each part by the number of items in that part. The total administration time and the average per item administration time is broken into two parts, that required by the test items themselves and that taken up by instructions. Although items are numbered 1 through 126, only 119 items are scored; seven practice items are not scored.

TABLE I

TIME REQUIRED TO ADMINISTER DLAB: TIME TAKEN BY ITEM TYPE AND CORRESPONDING SET OF INSTRUCTIONS

| PART OF TEST | TEST ITEMS | NO. ITEMS | ITEMS | | | TIME TAKEN BY INSTRUCTIONS | | | BOTH | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | TOTAL | | PER ITEM | TOTAL | | PER ITEM | TOTAL | | PER ITE |
| | | | MIN | SEC | SEC | MIN | SEC | SEC | MIN | SEC | SEC |
| I | 1-7 | 7 | 1 | 33 | 13.22 | 0 | 36 | 5.14 | 2 | 09 | 18.36 |
| II | 9-26 | 18 | 6 | 25 | 21.38 | 1 | 13 | 4.06 | 7 | 38 | 28.76 |
| III-1 | 28-40 | 13 | 5 | 31 | 25.46 | 2 | 46 | 12.76 | 8 | 17 | 38.22 |
| III-2 | 42-55 | 14 | 5 | 35 | 23.96 | 2 | 30 | 10.71 | 8 | 05 | 34.67 |
| III-3 | 57-73 | 17 | 8 | 46 | 30.94 | 3 | 37 | 12.76 | 12 | 24 | 43.70 |
| III-4 | 74.93 | 20 | 12 | 43 | 38.15 | 3 | 58 | 11.90 | 16 | 41 | 50.05 |
| IV | 97-126 | 30 | (22 | 00)* | 44.00 | (3 | 00)* | 6.00 | 25 | 00 | 50.00 |
| TOTAL | 1-126 | 119 | 62 | 29 | 31.50 | 17 | 44 | 8.94 | 90 | 13 | 40.44 |

*ESTIMATED AVERAGE. THIS PART NOT MACHINE PACED.

The Defense Language Aptitude Battery is a multiple-choice test with four parts. The first three parts of the test are paced by an audio tape. Part I is a self-report biographical inventory. In Part II and Part III the examinee learns an artificial language. In Part II the examinee discriminates stress patterns. In Part III the examinee selects the correct translation in the artificial language on the basis of grammatical rules provided in the instructions. Part III has four subparts. In Part IV, which the examinee is to complete in 25 minutes while working at his own pace, the examinee matches pictures to phrases in an artificial language according to rules given in the instructions.

When the test was reviewed, it was found that no meaningful savings would result if all four parts of the test were retained in their present form. Of course, if a whole part or subpart of the test were deleted, the instructions for that part could also be deleted.

Table I shows that every item deleted from Part III and IV will save more administration time than an item deleted from Part I or II; the table also shows that deleting all of Part IV or any subpart of Part III will save more administration time than deleting all of Part I or Part II.

An item analysis was conducted on DLAB in order to rank-order the items in terms of their overall contribution to test validity. 164 answer sheets were included in the sample. Final course grade at DLIFLC was used as a criterion. Item analysis data was used along with the analysis of the time required for test administration mentioned earlier to decide which items to delete. Eleven strategies for shortening the test were devised. The number of items deleted from the test in the various strategies ranged from four to sixty-seven items out of a total of 119 items on the original test. The time saved in the various strategies ranged from 3 to 49 minutes out of the 90 minutes required for the original test. The strategies that deleted more items either involved setting the minimum item-criterion correlation higher on the one hand, or deletion of whole parts of the test regardless of the intercorrelations between individual items and the criterion on the other hand. The tests were rescored according to each of the eleven strategies with the corresponding items deleted, resulting in a new distribution of test scores for each strategy. These new distributions were recorrelated with the criterion. By using the average time required to administer each type of item and the time required for instructions when whole sections of the test were deleted, the time required for test administration in each of the eleven strategies was computed. The results are at Table 2.

TABLE II

DLAB SHORTENING STRATEGIES -
ITEMS DELETED AND CORRESPONDINGLY
RECOMPUTED VALIDITY COEFFICIENTS
AND ADMINISTRATION TIMES

| ITEMS DELETED | RECOMPUTED CORRELATION WITH CRITERION | PROJECTED ADMINISTRATION TIME |
|---|---|---|
| Original Test | .366 | 90 min. 13 sec. |
| 70,78,98,113 | .390 | 87 min. 16 sec. |
| 59,70,76,78,89,96,98,113 | .396 | 84 min. 41 sec. |
| 2,14,26,30,42,51,55,70,77, 78,83,97,98,100,101,113, 114 | .452 | 81 min. 1 sec. |
| 70 - 89 | .383 | 73 min. 32 sec. |
| 2,14,26,30,40,42,51,55, 70,77,78,83,85,97,98,100, 101,103,106,113,114 | .468 | 78 min. 4 sec. |
| 90 - 119 | .347 | 65 min. 13 sec. |
| 2,14,15,23,26,28,29,30 36,40,41,42,50,51,55,59, 63,70-89,96,97,98,100,101, 103,104,106,113,114 | .480 | 59 min. 15 sec. |
| 70 - 119 | .353 | 48 min. 32 sec. |
| 2,14,15,23,26,28,29,30 36,40,41,42,50,51,55,59,63 70,76,77,78,80,83,85,89 90-119 | .4505 | 54 min. 11 sec. |
| 2,14,15,23,26,28,29,30,36, 40,41,42,50,51,55,59,63, 70-119 | .4225 | 41 min. 35 sec. |

The three most promising scenarios are compared at Table II.

TABLE III

THREE SCENARIOS FOR REDUCING DLAB LENGTH

| Scenario | Time | VALIDITY | |
|---|---|---|---|
| | | Uncorrected for Range | Corrected for Range |
| Original 119 Items | 90 min. 13 sec. | .366 | .555 |
| 72 Test Items | 59 min. 15 sec. | .480 | .684 |
| 64 Test Items | 53 min. 11 sec. | .451 | .650 |
| 52 Test Items | 41 min. 35 sec. | .423 | .620 |

Since item analysis data was used to decide which items to eliminate there is a high probability that some sampling error is involved. Our next step is to gather a large sample of approximately 350 answer sheets and crossvalidate our initial results. If the same shortening strategies yield the same results, we will feel justified in shortening the test.

Since DLIFLC receives the completed answer sheets of students from the AFEES and Lackland Air Force Base, and the final course grade of all past DLIFLC students are also available to us as criterion measures, there is no need for DLIFLC to conduct additional test administration or wait for a long period of time to collect criterion data when drawing a crossvalidation sample.

AD P001384

# MODELS OF HUMAN INFORMATION PROCESSING: IMPLICATIONS FOR TRAINER/SIMULATOR DESIGN

Spencer C. Thomason, Ph.D.
Essex Corporation
White Sands Missile Range, NM

Many studies have investigated the relationship between the fidelity of simulators and other training devices and the transfer of training they effect. Most of these have found that the greater the fidelity, the more positive transfer of training which occurs. There are, however, many problems with the measurement of transfer of training, and predicting it is well-nigh impossible. Therefore, the most common approach to the design of trainers and simulators has been to incorporate as much fidelity as is economically feasible into the device. This approach often results in low cost/training effectiveness, compared with other possible configurations. The reason for this is the diminishing returns on training effectiveness and, in some cases, even reductions in the transfer of training in the high-fidelity region of the transfer curve. Recent research in human information processing mechanisms and cognitive structures is discussed in terms of implications for designing training devices and maximizing their cost-effectiveness.

## INTRODUCTION

For those of you who may not be familiar with human information processing as a field of study, a brief description. As the name implies, it is concerned with how the human brain handles information, makes decisions based upon that information, and chooses or creates appropriate responses. As a basis for theorizing, the human brain is considered as a "black box" containing an unknown type of computer, and the object is to determine the architecture, processing capabilities, and other characteristics of that computer. This is done by observing the effects of inputing information of specific types and formats upon the output. This is a lot easier said than done.

This area of study actually began when Shannon and Weaver first described how to quantify information in 1949. The major impetus in the field, however, has been the development of the high-speed digital computer. The computer has given us many concepts in the area of information processing and handling which just did not exist before. It has also given us great flexibility in conducting experiements and in evaluating models of the human information processor through simulation. Considering the relative youth of the field and the complexity of the human brain, it is not surprising, however, that very little is actually known about that black box in people's heads.

At the present time, there are several major theories and models in existence, each of which fits the available empirical evidence pretty well and each of which is espoused by several well-known researchers in the field. In

earlier years, the wide disparity between different models of human information processing has discouraged training system designers from using them as a basis for training system design, due to the uncertain validity of any particular model. In more recent years, although the number of differing theories has not been significantly reduced (and in fact may actually have increased), the disparity between them has been reduced significantly. This reduction is primarily due to the existence now of a well-established body of empirical findings, which all of the models must account for.

The purpose of this paper is to present an overview of some of the more significant findings in the field and describe their implications for the design of training, specifically as applied to trainers and simulators.

## MODELS OF HUMAN INFORMATION PROCESSING

At this point, I should confess that, since I myself have done research in this field, I have certain prejudices toward particular models of human information processing. As a result, the orientation of this paper is based upon a certain class of human information processing models known as "dual-process models." The dual-process hypothesis postulates that human thinking is the result of two different types of information processing which are combined in various ways. Briefly, these are:

1. Controlled processing, which is under the conscious direction of the observer. This type of processing is generally assumed to be a sequential type of processing which is similar to that of a digital computer. It has certain limitations of capacity and, as the processing required becomes more complex, the longer it takes.

2. Automatic processing, which occurs automatically and is not necessarily under the control of the observer. This type of processing is generally thought to be parallel in nature (i.e., occurring "all at once") and is often equated with the type of processing characteristic of analog computers. This type of processing does not seem to have the same resource limitations as controlled processing and processing time is independent of complexity.

There are various models within this class, but the differences are irrelevant to the present discussion.

There is also another class, called "single-process models," which generally holds that only the first type of processing--controlled processing--exists. This class accounts for the phenomona attributed by the dual-process models to automatic, parallel processing as being the result of extremely fast serial controlled processing. Although the existence of these two classes is the source of a major dispute in the field, the preponderance of evidence presently seems to favor dual-process models (Thomason, 1981).

## LEARNING IN THE DUAL-PROCESS MODEL

The following discussion is based primarily upon a very detailed and well-supported dual-process model described by Schneider & Shiffrin (1977) and Shiffrin & Schneider (1977). The example used is the acquisition of a written alphabetic language.

The first step in acquiring a written alphabetic language is of course to learn to form and recognize the letters of the alphabet. The letters are at first perceived as being composed of various combinations of straight and curved lines. In the beginning, a child will examine a written letter, determine the line combination used, compare it mentally with those of the various letters until a match is obtained, and then name it. This involves the use of serial controlled processing. The child will eventually begin to recognize the letter without having to consciously analyze its structure. In the dual-process models, this is interpreted as being due to the establishment of an automatic process which makes all comparisons at once and outputs the name of the letter.

As learning progresses, the child will learn the sounds associated with each letter. When confronted with a word, the child will sound it out, letter-by-letter, then put the sounds together to obtain the word. Eventually, automatic processes are established which "recognize" familiar words without sounding them out. As the child becomes more and more familiar with the rules of spelling and pronunciation, more generalized automatic processes are gradually established which can apply these learned rules to establish a pronunciation even of unfamiliar words. These processes are applied without apparent conscious effort.

In due course higher and higher levels of automatic processes are established which can treat common phrases and even sentences as whole units. Practice in speed reading evidently takes advantage of and strengthens these processes.

In the end, these automatic processes may be visualized as a large network of filters. Each of the filters represents a well-learned concept or rule. The stimulus information is processed through this filter network and the output is used as the basis of response. To build up the network, the concepts must be learned and the appropriate links must be established between them.

Some of the linkages between nodes in the network may represent controlled processing, in which case the use of that link represents additional processing time. With extensive practice, that link may become an automatic process, indicating that the nodes connected by it have been combined into a higher-level concept or node. As more and more of the linkages become automatic, the overall network processing rate increases, leading to faster task performance and/or fewer errors in performance.

Several researchers have demonstrated the existence of such network-like structures in information processing. Briggs, Thomason, & Hagman (1978) found evidence for a decision tree, with parallel processing (pattern recognition) taking place at the branch points and serial processing occurring (reduction of uncertainty) along the branches, in a letter classification task. They also found that the structure of the tree changed as the probability of a target changed.

Smith (1980), found tree-like networks of technical concepts among radio and television repairpersons. He also found that those repairpersons who could perform a greater variety of repair tasks including novel, unusual, or complex ones had a much different concept structure than those who could not.

Schvanaveldt (1981), in a study particularly applicable to the present topic, found such concept structures among undergraduate fighter pilots, instructor pilots, Air National Guard pilots, and instructor weapon systems officers. Not only were there significant differences among the structures of these four groups, with the greatest differences being between those of the students and each of the other three groups, but he found that he could reliably classify the pilot population into the four groups based upon the structures alone.

In view of the above model of human information processing, the role of training of all types can be seen as a method of building a structure of knowledge pertinent to the skill being trained. In order to be maximally effective, the most efficient network of concepts and connecting links must be established by the training. If the network established by a course of training, including trainers and simulators, is identical to the network used by a skilled operator using operational equipment, then the training has accomplished its purpose.

In a training course, the classroom instruction will provide the concepts which establish the nodal points of the processing network, and it may present some minor connecting links. It is through practice of the skill itself, however, that most links are established and strengthened until they become automatic, thereby increasing the processing efficiency of the network. This practice is accomplished by using the operational equipment and in some cases, devices such as trainers and simulators.

## TRAINER/SIMULATOR FIDELITY AND TRANSFER OF TRAINING

One of the major difficulties with the design of trainers and simulators is determining how much fidelity is required in order to maximize transfer of training. This is because at present there are no reliable methods for predicting the transfer of training a given device will effect. Because of the general findings indicating that the greater the fidelity of the trainer or simulator, the greater the transfer of training to operational equipment which occurs, the approach generally adopted is to incorporate as much fidelity as is economically feasible into

the devices. It is quite possible, however, that this approach can lead to spending more money for increased fidelity, but with a resulting decrement in transfer of training. This is illustrated in Figure 1. As the fidelity of simulation increases, cost goes up at an accelerated rate. However transfer increases at a negatively accelerated rate and, at the high end of the fidelity scale, actually begins decreasing rapidly. This rapid decay in transfer as the cost of fidelity becomes extremely high reflects the empirical finding that increasingly complex military systems, because of their unreliability shown in Figure 2, become virtually impossible to maintain and use effectively.

Even when unreliability of the simulator equipment is not a problem, it is still possible to have greater transfer with less fidelity. A concrete example of this is found in a study (Caro, 1973) of the effectiveness of the Army's Synthetic Flight Training System (SFTS), also known as the 2-B-24 Link trainer. In this program, 43 hours in the SFTS replaced 53.5 out of 60 hours previously required on the operational equipment (TH-13T and UH-1-H helicopters) and 26 hours previously given in modified 1-CA-1 Link trainers. If one considers the operational equipment as a simulator with 100% fidelity, then the SFTS, with less than 100% fidelity, has greater transfer with less fidelity. The reason for this is that the complexity of the skills required to operate the operational equipment (and some complex trainers) may be so great as to actively interfere with learning. It is similar to asking a beginning piano student to learn by playing Mozart rather than by working up from simple scales and chords to more complex pieces.

In order to obtain the most transfer of training for the money spent, the objective of the designer should be to stay within the limits of the "honey region" shown in Figure 1, where there is a large amount of transfer of training for a relatively small investment. The human information processing model can aid the designer in achieving this goal.

## SOME IMPLICATIONS FOR TRAINER/SIMULATOR DESIGN

The most important implication the information processing model has for training systems in general is that the difference between the typical knowledge/processing structure of the incoming student and that of successful skilled operators is a key to designing an effective training system.

There are several techniques available for analyzing these structures, and each can be useful in different, often complementary ways. For example, a hierarchical clustering analysis of the structure of skilled personnel will give the designer a good idea of the way a course should be structured. The hierarchical structure yielded by such an analysis would correlate with a logical progression of concept formation which the training system should develop in the student.

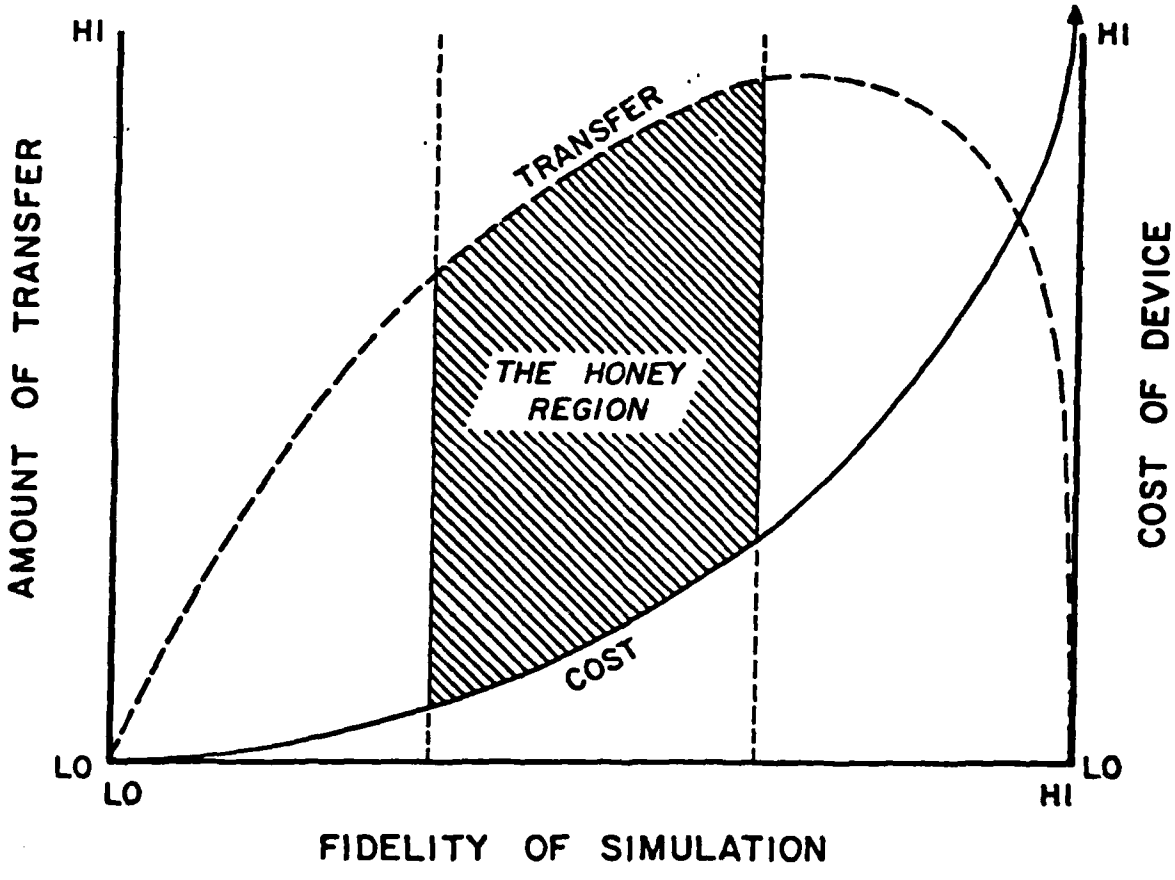



Figure 1. Positively accelerated cost and negatively accelerated transfer with increasing fidelity of simulation (From Roscoe, 1980).

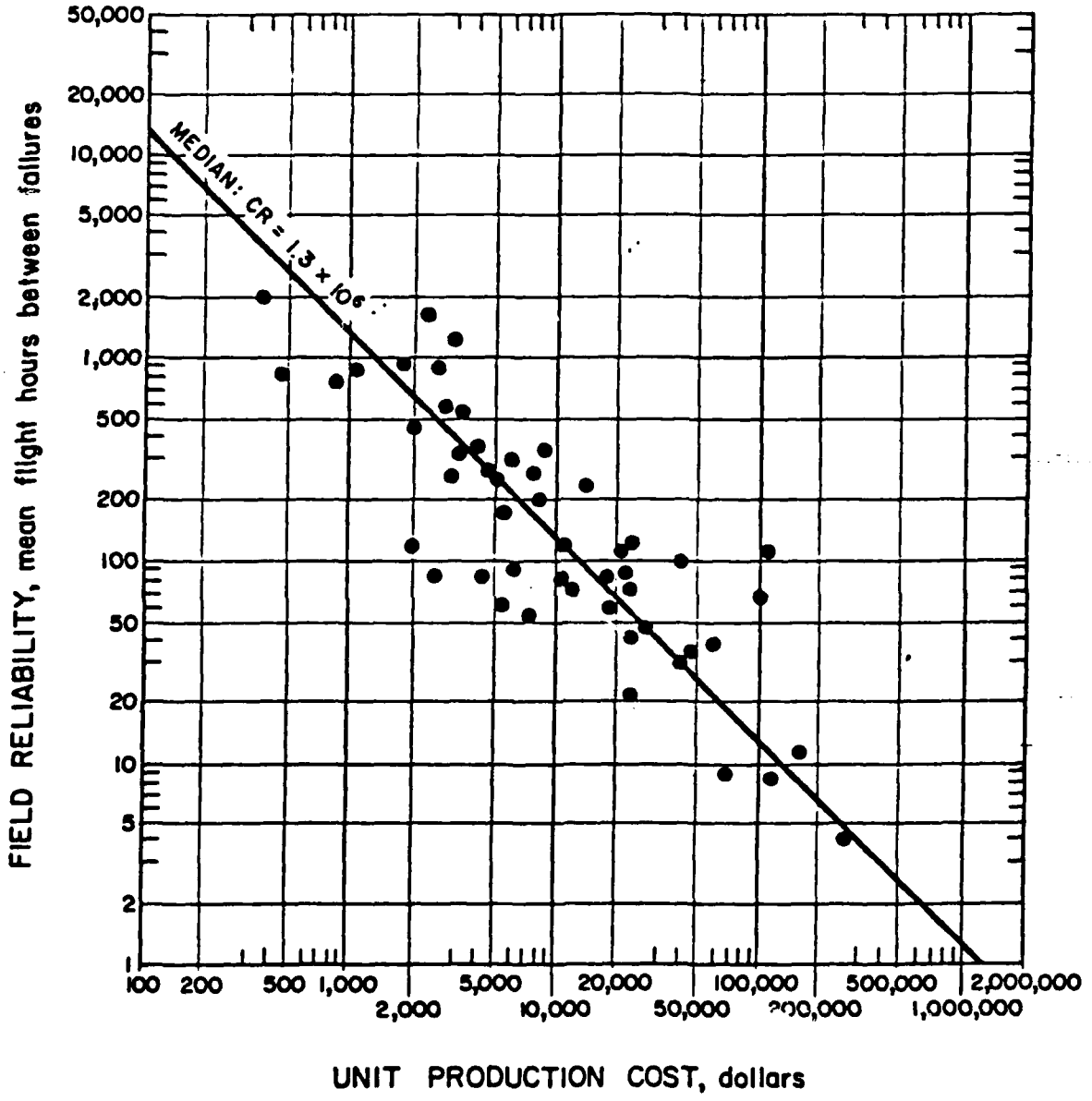



Figure 2. Avionics field reliability versus unit production cost of military systems (From Roscoe, 1980).

Another mathematical procedure known as Multi-Dimensional Scaling (MDS) can determine the different aspects of the concepts which skilled personnel use to differentiate between them. An effective training system should teach the students to perceive the skill being taught in terms of these conceptual "dimensions."

The network analysis which Schvanaveldt (1981) used in his study will give the overall knowledge structures, complete with linkages and relative link distances. By examining the structure of skilled personnel, certain "clusters" of tightly linked concepts will probably be detected. Such clusters might advantageously be trained as a unit; on a part-task trainer, for example. The length of the links connecting the concepts in the structure is also a clue to the importance of that link in efficient task performance. The shorter the link, the more important it is. This fact can be a help when making tradeoff decisions.

As an example of how such analyses could be used in flight training, the concept "descending turn" might be clustered with, and tightly linked to, the concept "landing approach," which is a higher-level concept, as shown by hierarchical cluster analysis. This would indicate that it would be advantageous to train the student to make descending turns within the general context of landing approach training, rather than as an exercise in itself. The advantage of the contextual approach is that it not only teaches the concept and skill of descending turns, but at the same time establishes and reinforces the link which relates it to landing approaches.

Another use of the network analysis is that, by analyzing the knowledge/processing structures of individual students, it may be possible to predict their success. Some indications of this were found by Schvanaveldt (1981) and are currently being investigated. Such individual analyses could also help establish a basis for individualized instruction.

The analysis of knowledge/processing networks can also be a useful diagnostic tool. For example, if a training system (or trainer/simulator) is less or more successful than expected, an analysis of the changes in the knowledge/processing structure it generates should give an indication where the problem lies. For example, an analysis may show that the knowledge/processing structure of the student after training is the same as that desired, except for the absence of a short (important) link. The training system can be modified to establish and reinforce that link.

It should be pointed out here that the existence of non-essential links in a knowledge/processing structure can be just as detrimental to the overall structure and as important to the training system design as a missing one. One characteristic of automatic processing in the dual-process model is that, once established, it is difficult to suppress (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). Practice must be expended to eliminate the link. Since irrelevant links essentially contribute false information to the network and represent a possible requirement for additional network processing time and capacity, their existence can contribute to increases

in time and errors in performance of the skill. It is therefore important that any training system adopted contribute a minimum of irrelevant links, since they may later have to be removed in order to attain peak skill levels and in the meantime may reduce the transfer of training effect.

In conclusion, approaching the design of trainers, simulators, and training systems in general from the standpoint of the current models of human information processing can provide the designer with several relatively powerful analytical tools. These tools can supplement, and may possibly even supplant, traditional approaches such as task analysis. In contrast to task analysis, which if properly done will give an accurate picture of the task structure as related to the equipment function, the information processing approach, through knowledge/processing network analysis, can delineate the structure of the skill required for successful task performance as it exists in experienced and qualified personnel.

## REFERENCES


Briggs, G.E., Thomason, S.C., & Hagman, J.H. Stimulus classification strategies in an information reduction task. Journal of Experimental Psychology: General, 1978, 107(2), 159-186.

Caro, P.W., Jr. Aircraft simulators and pilot training. Human Factors, 1973, 15, 503-510.

Roscoe, S.N. Aviation Psychology. Ames: The Iowa State University Press, 1980.

Schneider, W. & Shiffrin, R.M. Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review, 1977, 84, 1-66.

Schvanaveldt, R.W. Structure of Memory for Critical Flight Information. Presentation at meeting of the Rio Grande Chapter of the Human Factors Society, Sept 17, 1981.

Shannon, C.E. & Weaver, W. The Mathematical Theory of Communication. Urbana: The University of Illinois Press, 1949.

Shiffrin, R.M. & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 1977, 84, 127-190.

Smith, B.B. Cognitive structure of technical knowledge: a free association technique. Proceedings of the 22nd Annual Conference of the Military Testing Association, 1980, 903-914.

Thomason, S.C. An exploration, using varied and consistent mapping tasks, of some two-process models of human information processing (Doctoral dissertation, New Mexico State University, 1981). Dissertation Abstracts International, in press.

AD P001385

TUBBS, John D.; DEASON, Paul J.; EVERETT, James E.; and HANSEN, Alan D.;
USATRASANA, White Sands Missile Range, New Mexico

# CHAPARRAL TRAINING SUBSYSTEM EFFECTIVENESS ANALYSIS (TSEA)

This is the fourth in a succession of training studies conducted for the USA Air Defense School by USATRADOC Systems Analysis Activity, WSMR, NM. The REDEYE Man Portable Air Defense System (MANPADS) training was studied initially (1977) in a Weapon Systems Training Effectiveness Analysis (WSTEA). The REDEYE training was studied more extensively as a part of the Army Training Study (ARTS - 1978). The VULCAN Training Subsystem was evaluated in 1979, and the current CHAPARRAL study conducted in 1980. Other studies are currently being conducted or planned as a part of the on-going effort to evaluate training for all of the fielded air defense systems.

The CHAPARRAL system is a part of the Short Range Air Defense (SHORAD) group of weapons. It consists of the M48 guided missile system which is launched from a full-tracked carrier (M730). Both basic MIM-72A and improved MIM-72C missiles are used. The weapon system is normally manned by a 5-man crew, but more than half of the crew drills conducted for this study were performed by 4-man crews due to shortage of personnel.

This study evaluated 25 CHAPARRAL batteries from both CONUS and OCONUS locations. A total of 1076 individual crewmen were surveyed and tested. Over 150 crews were evaluated in selected crew drills. Significant differences were found between the performance level of CONUS and OCONUS units. The OCONUS units demonstrated higher proficiency in both individual and crew skills.

# CHAPARRAL TRAINING SUBSYSTEM EFFECTIVENESS ANALYSIS (TSEA)

JOHN D. TUBBS
USA TRADOC SYSTEMS ANALYSIS ACTIVITY
WSMR NM 88002

## 1.0 INTRODUCTION

### 1.1 PURPOSE

This report presents the results of the CHAPARRAL Training Subsystem Effectiveness Analysis. This TSEA was conducted to support the on-going program of training evaluation and improvement as directed by the US Army Air Defense School, Fort Bliss, Texas.

### 1.2 BACKGROUND

a. The CHAPARRAL system is a member of the family of Short Range Air Defense weapons whose purpose is primarily to provide close-in air defense to ground combat units, combat support units, and installations.

b. Several training studies have addressed the CHAPARRAL system in the recent past. These studies include the CHAPARRAL Weapons Systems Training Effectiveness Analysis (WSTEA), and the CHAPARRAL Initial Screening Training Effectiveness Analysis (ISTEA). These reports identified numerous problems with CHAPARRAL training in Advanced Individual Training (AIT) and in the units. These studies, while arriving at valid useful conclusions, did not address the overall proficiency of CHAPARRAL crewmen in AIT and in the units.

### 1.3 PROBLEM

The results of the CHAPARRAL WSTEA conducted by Litton-Mellonics, and the CHAPARRAL ISTEA conducted by TRASANA indicate that the typical CHAPARRAL system is not performing to its full design effectiveness as perceived by the Air Defense Artillery (ADA) community. The results of these studies indicated that the training subsystem contributes to this shortfall. In addition, personnel management problems and the reliability, availability, and maintainability characteristics of equipment currently in tactical units also contribute to the system effectiveness shortfall. A TSEA is needed to clearly document the shortfalls, as well as to attempt to identify the causes and suggest means to remedy the shortfalls.

### 1.4 IMPACT OF PROBLEM

The capability and proficiency of the soldiers associated with the CHAPARRAL directly affect the survival as well as the performance of protected elements. The current and the expected proficiency (performance) levels of CHAPARRAL crewmen must be assessed so that any shortfalls in actual performance can be corrected. Should identified problems not be amenable to solution through training subsystem or hardware modifications, current expectations of system performance should be re-evaluated.

### 1.5 OBJECTIVES

The objectives of the CHAPARRAL TSEA were:

a. To determine whether shortfalls exist in the training of Soldier's Manual (SM) tasks in the institution and unit and propose solutions for any shortfalls found.

b. To determine individual and crew proficiency and its relationship to design effectiveness.

c. To determine the effect of crew proficiency on combat effectiveness.

d. To determine the current CHAPARRAL soldier capability and its relationship to soldier proficiency.

e. To determine the impact of the Improved CHAPARRAL on training support requirements.

## 1.6 SCOPE

This study considered the current level of proficiency of Military Occupational Specialty (MOS) 16P Short Range Air Defense Artillery missile CHAPARRAL crewmen. The proficiency of CHAPARRAL crewmen is related to two main factors: institutional and unit training.

### 1.6.1 Institutional Training

a. The MOS 16P One Station Unit Training (OSUT) program is designed to qualify the individual in basic soldier skills and a few selected CHAPARRAL-related tasks. The soldier is also familiarized with all other of the skill level 1 CHAPARRAL system-related tasks. The final phase of OSUT corresponds to AIT under 2-Station Unit Training.

b. Graduates from the final phase of OSUT were tested to determine whether deficiencies existed in terms of meeting the minimum requirements for the MOS 16P. The analysis presented in this report assesses the proficiency of CHAPARRAL crewmen who completed OSUT during May - June 1980.

c. The parameters used to evaluate 16P OSUT graduates included visual aircraft recognition (VACR), hands-on proficiency, general aptitude, attitude, and system skills/knowledge. The survey and test instruments were constructed and administered to determine the capabilities of CHAPARRAL crewmen at the completion of OSUT.

### 1.6.2 Unit Training

a. Unit training is a combination of self-training and on-the-job training (OJT) with the CHAPARRAL squad, as well as scheduled collective training. At the unit, the 16P soldier is trained so that he can qualify in all SM tasks at his particular skill level.

b. This study utilized a broad base of survey and performance data from 12 ADA battalions and from those CHAPARRAL crewmen present-for-duty in each battalion. The test data was examined to determine if there were substantial differences between units. Although performance data are reported by unit, this report deals mainly with the overall proficiency of a typical CHAPARRAL crew.

c. The parameters used to evaluate 16P soldiers in the units included similar surveys and test instruments as those used for the OSUT evaluation. CHAPARRAL crew drill proficiency was also determined by additional hands-on testing.

d. Squad reaction times for target engagements were measured. These reaction times were compared with a simulation of engagements for various threats to demonstrate the effect of reaction time on combat effectiveness.

## 2.0 INSTITUTIONAL TRAINING

### 2.1 INTRODUCTION

The purpose of this section is to report the results of the evaluation of institutional training for MOS 16P CHAPARRAL missile crewmen students. This section presents the data collection methodology and the analysis of these data.

### 2.2 APPROACH

a. Training effectiveness was assessed by determining individual proficiency at the completion of OSUT. Specific task proficiencies were determined in relation to USAADS' requirements for task qualification and course completion.

b. The level of individual proficiency at completion of OSUT was determined by the administration of specially prepared or selected tests. A VACR test was constructed from GOAR kit slides and administered in the same way as the Skill Qualification Test (SQT). A skills and knowledge test was formed by selecting questions from several written tests used during past SQTs and other sources. Each question covered a specific task so the test results could be used as an individual measure of knowledge for tasks or task groups within skill level 1. Particular attention was given to the knowledge of the CHAPARRAL-related tasks in which the OSUT student was to be qualified at the time of graduation.

c. Personnel data for the soldiers tested were obtained from the MILPERCEN files. This information was used to determine if Armed Services Vocational Aptitude Battery (ASVAB) scores could be used as performance predictors.

### 2.3 ONE STATION UNIT TRAINING (OSUT) TESTING

#### 2.3.1 Discussion

OSUT at USAADS was a 13-week course for soldiers entering the Army. The emphasis at the first 6 weeks is basic combat training (BCT) and the emphasis of the final 7 weeks was on MOS training.

a. During the BCT phase, the new soldier is expected to learn the basic skills of soldiering. Examples of subject matter taught in BCT are: health and welfare, marksmanship, physical fitness, and military justice.

b. The students will normally receive approximately 180 hours of MOS instruction. Audio-visual aids (TEC tapes, synchronized slides, and VACR) account for 52% of a soldier's time in MOS training. Hands-on training accounts for only 22% of MOS instruction. The balance of the student's time consists of examinations, administration, in- and out-processing, and observation of CHAPARRAL range firings.

c. A short range ADA missile crewman must qualify in 55 critical tasks for CHAPARRAL and six critical tasks for REDEYE to be fully proficient in his MOS at skill level 1. This study examined 15 of the 55 tasks by either written or hands-on tests. A soldier is said to be qualified at a task when he has received sufficient training to perform the task in its entirety and his ability has been tested. He is said to be familiarized with a task when he has received some training on the task, but he may not be able to fully perform the task without additional training. The CHAPARRAL crewman is required to qualify in 20 of the 55 CHAPARRAL critical tasks at the completion of OSUT in order to be awarded the MOS 16P. The MOS 16P OSUT graduate needs to be familiarized with the remainder and then must work or study to qualify during his unit assignment.

2.3.2 Methodology

a. Students from three 16P classes, who were graduating in the late spring, 1980, were tested during their final week of OSUT. The test consisted of a VACR test, a written test, and hands-on proficiency tests.

(1) The VACR test consisted of two views of each aircraft projected one at a time on a screen in a classroom. There were 30 two-view sets presented so that, for each set, there was a 5-second showing of the first slide, 5 seconds for the second slide, then 10 seconds for the student to write his answer on a form provided. A correct answer consisted of the aircraft's alpha-numeric designator (MIG 25) or NATO name (FOXBAT). There were 25 different aircraft. Five aircraft were presented a second time using different slides.

(2) The written tests, administered in a classroom setting, required approximately 2 hours. These tests assessed: the student's attitudes and opinions regarding the instruction received; his perceived competence; his reading and mathematics ability level (using the screening pretest for the Adult Basic Learning Examination - called the SelectABLE); his knowledge of basic soldiering and CHAPARRAL specific items measured by a 61-question skills and knowledge test and a 6-question map test.

(3) Two hands-on tests were given. The first test was to measure how well the students could perform the following subtasks: preenergize, energize, and deenergize the CHAPARRAL launch station. The second test required the students to locate and indicate the proper procedures to perform six preventive maintenance (PM) checks of the CHAPARRAL carrier's fuel, coolant, and lubricant systems.

(4) One hundred and eight OSUT students were given the written tests and 92 took the hands-on tests. Total test duration was approximately 8 hours, spread over 2 days during the students' final week of training.

b. Background information was also collected from the OSUT students. The students' ASVAB scores were obtained from MILPERCEN to investigate the relationship of aptitude factors and performance.

c. Information pertaining to the CHAPARRAL mechanics course (MOS 24N) was collected by interviews with the instructors. Additional information was obtained from the Air Defense Accessions Training Effectiveness Analysis (ATEA) reported in TRASANA Technical Report No. TEA 7-81, Mar 81.

## 2.4 ANALYSIS

The analysis of data obtained from written and hands-on testing was used to determine the level of individual proficiency at completion of OSUT. Special attention was directed to proficiency in VACR and results of the skills and knowledge test. Hands-on proficiency in energizing and deenergizing the launch station and performing PM checks and services on the carrier were also analyzed. The results for all students are presented together since there were no significant differences in performance among the three classes.

### 2.4.1 Visual Aircraft Recognition Test

a. VACR is one of the CHAPARRAL-related tasks in which the student is required to be qualified at the time of graduation. A score of 90% correct is required for qualification.

b. The VACR test was graded for correct identification by either the alpha-numeric designator or NATO name. However, those aircraft that were incorrectly identified were categorized as to the degree or type of error as follows:

- Omission - could not identify
- Wrong - friendly identified as hostile
- Wrong - hostile identified as friendly
- Wrong - but no confusion as to whether hostile or friendly

c. All of the 25 aircraft included in the test had been taught as a part of the VACR instruction.

d. The VACR test results for each scoring category are shown in table 2-1.

TABLE 2-1. MEAN OSUT VACR RESULTS (N=106 STUDENTS)

| | ERROR CATEGORY | | | |
|---|---|---|---|---|
| CORRECT (%) | OMITTED (%) | FRIENDLY IDENTIFIED HOSTILE (%) | HOSTILE IDENTIFIED FRIENDLY (%) | NO CONFUSION HOSTILE OR FRIENDLY (%) |
| 54.0 | 14.4 | 5.0 | 6.0 | 20.6 |

e. The VACR proficiency of 16P students was well below the 90% required for qualification in all cases. The test was administered approximately 5 weeks following their block of instruction. The scores ranged from a high of 73% correct (three students) to zero correct (one student). If the mean score of 54% correct is combined with the 20.6% which were not confused as to the hostile or friendly category, satisfactory proficiency is still not met.

## 2.4.2 Skills and Knowledge Test

a. The skills and knowledge test presented 61 questions and was identical to the test given to the soldiers in the units. Since all skill levels were tested in the units, only 46 of the questions covered skill level 1 tasks. These were used to assess the knowledge of OSUT students. Thirteen of the 46 questions related to four tasks in which graduates were to be qualified at completion of OSUT. The remaining 33 questions related to 10 tasks with which students were only expected to be familiarized.

b. All OSUT students are required to perform at or above a 70% level of proficiency on qualified critical tasks prior to graduation. The results of the test are shown in table 2-2 for the 46 questions covering four "qualified" and 10 "familiarized" tasks. The mean correct score for the 13 "qualified" task questions was 55%. Eleven out of 108 students answered 70% or more of the "qualified" task questions correctly. The mean correct score for the 33 "familiarized" task questions was 35%. As expected, the students scored better on the tasks in which they were to be qualified, but scores were too low to be acceptable as demonstrating qualification.

TABLE 2-2. OSUT SKILLS AND KNOWLEDGE TEST RESULTS

| TASK TITLE | TASK NO. | TASK TYPE* | NO. OF QUESTIONS | NO. OF STUDENTS PASSING ALL QUESTIONS | NO. OF STUDENTS FAILING ALL QUESTIONS | STUDENT PERFORMANCE MEAN (%) | PHASE WHERE TASK IS TAUGHT |
|---|---|---|---|---|---|---|---|
| First Aid | 1001 | Q | 3 | 19 | 5 | 60 | BCT |
| M-16 Aircraft Engagement | 1020 | Q | 1 | 19 | 89 | 18 | BCT |
| Zeroing a M16A1 Rifle | 1030 | Q | 1 | 67 | 41 | 62 | BCT |
| Energize/ Deenergize | 1042 | Q | 8 | 0 | 0 | 58 | MOS |
| NBC Hazards | 1010 | F | 2 | 32 | 15 | 58 | BCT |
| Operate Telephones | 1022 | F | 1 | 61 | 47 | 57 | BCT/MOS |
| Proper RTO Comm | 1023 | F | 4 | 1 | 10 | 39 | BCT/MOS |
| Operate TADDS | 1037 | F | 3 | 0 | 32 | 16 | MOS |
| Observer Procedures | 1054 | F | 1 | 25 | 83 | 23 | MOS |
| PM Carrier Checks | 1044 | F | 3 | 4 | 5 | 41 | MOS |
| Misfire/ Hangfire | 1048 | F | 2 | 50 | 21 | 64 | MOS |
| Missile Upload | 1051 | F | 4 | 2 | 12 | 39 | MOS |
| Crew Drill | 1049 | F | 2 | 20 | 32 | 48 | MOS |
| Target Engagement | 1050 | F | 11 | 0 | 2 | 30 | MOS |

*Q - Qualified
F - Familiarized

### 2.4.3 Hands-On Tests

OSUT graduates were given two hands-on performance tests. The first hands-on performance test was to preenergize, energize, and deenergize the CHAPARRAL launch station, which is a "qualified" task. The second test was to perform the fluid level checks on the CHAPARRAL carrier, which is a "familiarized" task.

a. Of the 92 students who were administered the first test, 20 received a GO in all three subtasks (preenergizing, energizing, deenergizing) and 29 students failed all three subtasks. Table 2-3 shows the results for each of the three subtasks. Fifty-three percent of the students could properly preenergize the CHAPARRAL launch station, and only 37% could either energize or deenergize the system.

TABLE 2-3. OSUT HANDS-ON TEST RESULTS OF THE CHAPARRAL LAUNCH STATION (N=92)

| SUBTASK | NO. OF STUDENTS PASSING | NO. OF STUDENTS FAILING | % PASSING |
|---|---|---|---|
| Preenergize (12 steps) | 49 | 43 | 53 |
| Energize (20 steps) | 34 | 58 | 37 |
| Deenergize (15 steps) | 34 | 58 | 37 |
| All Subtasks | 20 | 29 | 22 |

b. Seventy-five students were asked to make five checks: three oil level checks, one fuel check, and one coolant level check. Table 2-4 shows the results for the 75 who were evaluated for actually performing the checks. Seven of the 75 students did not know how to perform any of the fluid checks. Eighteen students knew where to perform all of the five checks, but no one successfully completed all 13 how-to steps.

TABLE 2-4. KNEW HOW TO PERFORM FLUID CHECKS - OSUT (N=75)

| SUBTASK | NO. OF STUDENTS PASSING | NO. OF STUDENTS FAILING | % PASSING |
|---|---|---|---|
| Engine Oil (1 Step) | 49 | 26 | 65 |
| Engine Coolant (1 Step) | 52 | 23 | 69 |
| Differential Oil (2 Steps) | 27 | 48 | 36 |
| Transmission Oil (6 Steps) | 3 | 72 | 4 |
| Fuel (3 Steps) | 10 | 65 | 13 |
| All Subtasks | 0 | 7 | 0 |

# 3.0 UNIT TRAINING

## 3.1 INTRODUCTION

The purpose of this section is to report the results of the evaluation of unit training for MOS 16P CHAPARRAL missile crewmen. This section presents the data collection methodology and the analysis of this data for the units visited.

## 3.2 APPROACH

a. The training effectiveness addressed in this chapter was assessed in terms of both individual and crew proficiencies. The current levels of individual and crew proficiency in the unit were determined by the results of written and hands-on tests. The test standards or performance requirements were established by USAADS or developed jointly by USAADS and USATRASANA when standards did not exist.

b. The hands-on tests were selected crew drills described in FM 44-4. The grading criteria for the drills were also developed jointly by USAADS and USATRASANA. The criteria provided both a basis for pass/fail determination and a quantitative measure of performance. All the crew drills were timed, including squad reaction times for target engagement. These reaction times were compared with a simulation of engagements for various threats to demonstrate the effect of reaction time on combat effectiveness.

c. Information was obtained from both crewmen and trainers by a survey of the type of training aids used, their frequency of use, and their considered effectiveness. This information provided a measure of the type and level of training of each individual unit visited.

d. Personnel data for the tested soldiers were obtained from MILPERCEN files. This information was used to determine the relationship between the various performance data and ASVAB scores and derive predictors of performance, if applicable.

## 3.3 SOLDIER TESTING

### 3.3.1 Discussion

Unit training for CHAPARRAL is intended to build on the individual basic skills and knowledge acquired during institutional training. In addition, it is required to develop crew skills so individual responsibilities are carried out competently and efficiently in a team effort. The squad must be trained to function as would be required in combat either as a single defensive unit or as a part of a platoon supporting a maneuver unit. It is the responsibility of the unit to provide the on-the-job training so soldiers can qualify in all the SM tasks. Every soldier is then expected to reinforce his training with the self-help material available. There is also additional reinforcement of institutional training through the Non-Commissioned Officer Education System (NCOES). Unit training is structured toward meeting the individual requirements of the SQT and the collective requirements of the Army Training and Evaluation Program (ARTEP). The SQT is required for every soldier in grade E-4 or higher as a part of career development. The ARTEP is used to evaluate the mission capability of various sized units.

### 3.3.2 Methodology

Twelve C/V battalions were tested on a schedule which permitted one week to test each battalion. Six battalions were in CONUS, five in USAREUR, and one in Hawaii (the USAREUR and Hawaiian battalions are designated OCONUS in this report). It was requested that all MOS 16P CHAPARRAL personnel who were assigned to squads be made available during the testing period for both written and equipment hands-on testing. Four CHAPARRAL M-48 systems were to be provided for the hands-on tests at each site. A total of 1076 CHAPARRAL crewmen were tested.

a. Written tests were administered to obtain measures of individual proficiency in VACR, knowledge of critical tasks, and an indication of reading grade level (RGL) with the SelectABLE test. Information on soldier backgrounds, attitudes, and training methods and aids used in training were obtained by a written survey. ASVAB scores and other selected data were obtained from the MILPERCEN files.

b. Crew testing was accomplished to obtain measures of crew proficiency in three drills. The missile upload, prepare-for-action, and target engagement crew drills were conducted and graded by detailed standards.

## 3.4 ANALYSIS

a. The data obtained during the written and hands-on testing were used to assess the overall proficiency of the CHAPARRAL crews. Early in the analysis, it was noted that the proficiency levels of OCONUS units were consistently higher than CONUS units. Therefore, most of the results of the data are presented in the two major groups of OCONUS and CONUS.

b. Since a number of ADA battalions were surveyed and tested, this report also addresses how well each unit performed in comparison to all other units. To this end, most of the figures which include the results of the written and hands-on tests indicate the results by individual unit. Identification of each battalion included in this study is by a randomly assigned number (known to the battalion) preceded by the letter designator for OCONUS (O) or CONUS (C). The battery is identified by the corresponding battery letter in the third place. Thus, 05C would indicate an OCONUS unit, 5th battalion, and "C" battery, and C3D would indicate a CONUS unit, 3d battalion, and "D" battery. The purpose for this means of identification is to maintain anonymity of the units to the general reader, but to permit the units tested to identify their respective performance within the report.

### 3.4.1 Visual Aircraft Recognition (VACR) Test

a. The VACR test was graded for correct identification by either the alpha-numeric designator or the NATO name. Those aircraft that were incorrectly identified were categorized as to the degree or type of error as described in paragraph 2.4.1b. The relative consequence of each category of error was not weighted except to note that the last category is the least objectionable. Given an error in the last category, friendly aircraft would not be jeopardized and hostile aircraft would be engaged.

b. Twenty-six of the 30 sets of GOAR kit slides were of aircraft currently required for SQT qualification (a score of 90% is required for qualification). Therefore, the scores for the 26 sets of slides were used for

comparison of proficiency between the units because of the common SQT requirement. Table 3-1 shows the overall results of the 950 crewmen tested.

TABLE 3-1. MEAN UNIT RESULTS OF VACR - CONUS VS OCONUS

| UNIT LOCATION (NO. TESTED) | CORRECT (%) | ERROR CATEGORY | | | |
|---|---|---|---|---|---|
| | | OMITTED (%) | FRIENDLY IDENTIFIED HOSTILE (%) | HOSTILE IDENTIFIED FRIENDLY (%) | NO CONFUSION HOSTILE OR FRIENDLY (%) |
| CONUS (450) | 66.1 | 12.7 | 2.6 | 3.1 | 15.5 |
| OCONUS (500) | 84.7 | 4.1 | 1.4 | 1.3 | 8.4 |

c. Table 3-2 presents a summary of the range of VACR test results. In OCONUS, 53% of the soldiers tested qualified on the VACR test compared to 24% of the soldiers tested in CONUS. Over one-half of the soldiers tested in CONUS had a VACR score below 70% correct. It is of interest that 12% of all soldiers tested identified all aircraft correctly.

TABLE 3-2. RESULTS OF VACR TEST - PERFORMANCE INCREMENTS

| VACR SCORES | CONUS (N=443) | OCONUS (N=484) | TOTAL (N=927) |
|---|---|---|---|
| 90 - 100% (qualification) | 108 (24%) | 254 (53%) | 362 (39%) |
| 80 - 90% | 72 (16%) | 118 (24%) | 190 (20%) |
| 70 - 80% | 36 (8%) | 28 (6%) | 64 (7%) |
| Below 70% | 227 (51%) | 84 (17%) | 311 (34%) |

### 3.4.2 Hands-On Tests

Three different squad drills were conducted to evaluate hands-on task performance. The first was the missile upload in which four inert missiles were loaded on the launch rails in a timed drill. The second drill, which was also timed, included the prepare-for-action drill in which the squad moved into a position and prepared for action. The third drill was the target engagement drill which was conducted at the conclusion of prepare-for-action. This final drill, performed by only a portion of the CHAPARRAL squads tested, was not used in the evaluation of proficiency of individual squads but was used to obtain an indication of the reaction time required to engage targets. The results of the target engagement drill were compared with a simulation of engagements for various threats to demonstrate the effect of reaction time on combat effectiveness.

a. Missile Upload.

(1) The detailed procedure for removing four missiles stowed in the carrier compartments and installing them on the launch rails with wings and fins was evaluated. The procedure, delineated in FM 44-4, should be completed

in 8 minutes by a 5-man crew. The majority of the crews tested in this study had only four members, but no alternate time requirement existed for the 4-man crews. The performances of 4- and 5-man crews were examined separately to determine the impact of reduced crew size.

(2) This crew drill was usually conducted in the motor pool or an open field so the crew would be unhindered by other activity. One person from the test team observed the drill and recorded the "real time" commentary of the drill on audio tape. The information on the tapes was subsequently graded for compliance with specified procedures. Table 3-3 relates the results for both 4- and 5-man crews.

TABLE 3-3. MISSILE UPLOAD PERFORMANCE - CONUS VS OCONUS

| OCONUS | | | | CONUS | | | |
|---|---|---|---|---|---|---|---|
| CREW | FOUR CRITICAL PROCEDURES | | TIME | CREW | FOUR CRITICAL PROCEDURES | | TIME |
| 4-MAN/ 5-MAN | PASS | FAIL | 8 MIN OR LESS | 4-MAN/ 5-MAN | PASS | FAIL | 8 MIN OR LESS |
| 46/43 | 14/12 | 32/31 | 10/30 | 17/46 | 0/1 | 17/45 | 0/5 |
| Percent Passed | 30/28 | | | Percent Passed | 0/2 | | |

b. Prepare-for-Action.

(1) The detailed procedure for prepare-for-action is outlined in FM 44-4. This drill should be completed by a 5-man crew in 15 minutes, and again, no alternate criterion exists for 4-man crews. The drills were also conducted in the motor pool or open field area. Two members of the test team worked together to observe individual crew member performance. Each team member used portions of the FM 44-4 checklist to follow the progress of each crew member. The checklists were then used to grade performance following completion of the tests.

(2) Less than 46% of the 149 crews tested passed the "10 essential steps", and less than 12% passed the full 28-step criterion. Of the 17 crews who passed the 28 steps, the difference between 4-man and 5-man crew performance was not significant. Therefore, the results are presented in table 3-4 for all crews by OCONUS and CONUS groupings.

TABLE 3-4. PREPARE FOR ACTION RESULTS - CONUS VS OCONUS

| OCONUS | | | | CONUS | | | |
|---|---|---|---|---|---|---|---|
| BATTERY'S | NUMBER OF SQUADS | TEN ESSENTIAL STEPS PASS/FAIL | 28 STEPS PASS/FAIL | BATTERY'S | NUMBER OF SQUADS | TEN ESSENTIAL STEPS PASS/FAIL | 28 STEPS PASS/FAIL |
| All | 90 | 48/42 | 12/78 | All | 59 | 20/39 | 5/54 |
| Percent Passed | | 53 | 13 | Percent Passed | | 34 | 8 |

c. Target engagement performance data was obtained on a limited sample (48 crews). The communication between squad leader and senior gunner was recorded and the time measured from target "alert" to "missile away". Only six of the 48 crews were rated satisfactory.

## 4.0 RECOMMENDATIONS - INSTITUTIONAL TRAINING

It is recommended that:

a. USAADS pursue means to improve the proficiency level of 16P students and that minimum requirements be established for graduation from OSUT. The requirements should consider the unit's need for replacement crew members.

b. USAADS should consider improving the climate control (i.e., air conditioning) of the classrooms used in 16P OSUT.

c. USAADS should provide the number of instructors authorized for OSUT. There was a shortage of CHAPARRAL OSUT instructors causing a high student to instructor ratio.

## 4.1 RECOMMENDATIONS - UNIT TRAINING

It is recommended that:

a. The units emphasize crew drills, especially prepare-for-action and target engagement, as a part of regular monthly training, as a minimum.

b. Crew drill standards be developed that apply to procedure as well as time. Further, it is recommended that 4-man crew standards be developed.

c. The concept of a "master gunner" be studied to act as a unit operations sergeant/training NCO. The master gunner would be expected to be the expert at battalion on CHAPARRAL training.

d. VACR training be given additional emphasis until acceptable proficiency is achieved, and that effective training for maintaining proficiency be continued.

e. The T3 trainer and live targets be utilized for target engagement training and cross-training in order to meet the 15-second time standard.

f. A study be performed concerning the product improvement of the MPU and the communications subsystem.

g. USAADS direct resupply of training manuals, circulars, and information to assure replacement needs for training down to squad and individual level.

h. USAADS investigate means of simplifying the installation of the T3 trainers, to include avoidance of deadlining the readiness status of the CHAPARRAL system.

i. The unit trainers schedule regular training for squads wearing the MOPP-4 gear to develop confidence and alternative techniques for intra-squad communication.

AD P001386

Vandyke, G. A. CPT, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario, Canada. (Thurs. A.M.)

## The Personality Research Form (PRF) as a Prediction for Success in Pilot Training

This study examined the utility of Jackson's Personality Research Form (PRF) in the selection of aircrew in the Canadian Forces. A total of 1962 male candidates completed either the English or French version of Form E.

Major findings focused on the validity scales, anglophone versus francophone differences on the 22 scales and the predictive validity of the scales against performance in flying training.

Results show that while Infrequency scores are within the range reported by Jackson, the Desirability scores obtained were very high. There are also some significant correlations between Desirability scores and scores on other trait scales. There were a number of differences between performance of anglophone and francophone subjects and these will be discussed, together with other early stage psychometric evaluations of the Form.

Finally, the PRF may prove useful in the counselling of candidates who are interested in becoming military pilots.



PREVIOUS PAGE IS BLANK

# THE PERSONALITY RESEARCH FORM (PRF) AS A PREDICTOR OF SUCCESS IN PRIMARY FLYING TRAINING

Captain G.A. Vandyke
Canadian Forces Personnel Applied Research Unit
4900 Yonge Street
Willowdale, Ontario

## INTRODUCTION

1. The Canadian Forces Aircrew Selection Centre (CFASC) annually processes approximately 1000 candidates who wish to enter the Canadian Forces as aircrew. The process of selection is constantly monitored so that a valid test battery can be maintained and new innovations can be evaluated. It was felt that the addition of a personality inventory might improve the present selection system by adding information from the personality domain to those tests already used.

2. Since 1978, the Personality Research Form (PRF-E) incorporating 22 scales has been administered to applicants being processed at the CFASC. Over this period a total of 1960 subjects have completed the PRF, the scores being used for experimental purposes only.

### Detection of Invalid Records

3. The Infrequency scale on the PRF was designed to identify those individuals who tend to answer the questions in a non-purposeful or random manner. Such situations as failing to understand the questions, unco-operativeness or carelessness in answering will produce high scores in this scale. A raw score of four or greater indicates the possibility of errors in scoring or in responding.

4. The Desirability scale on the PRF was developed to permit detection of responses that were made in terms of what the individual saw as profile requirements rather than as true representations of his actual profile. High scores in this scale could indicate either conscious distortion or impression management or more subtly, influences of a typically high self-regard or of a high degree of conventional socialization. Low scores may indicate possible tendencies towards malingering or more likely, low self-regard.

### Background

5. As a preliminary step during this study, the Infrequency and Desirability scales of the PRF were examined. Based on this study the following conclusions were reached.

a. The Infrequency scores appear to be generally within the norms expected, that is 98.96% of all profiles contain scores where infrequency is three or less. It can therefore be said that based on these scores, the profiles can be considered valid;

b. The Desirability scores indicate possible problem areas with validation of the profiles. Jackson's expectations as to percentages of individuals falling into certain standard score areas are not in agreement with the percentages arrived at in this study. This would indicate either the subject group is somehow different from university norms or that the entire personality profile for the subject group is suspect. The first of these hypotheses has probable merit due to the fact that Canadian Forces Recruiting Centres (CFRCs) have already carried out a preliminary selection process and the subject group are all well motivated towards becoming pilots in the CF. Further study of this question is considered necessary. (FIGURES 1 and 2)

6. The purpose of the present study is to assess the predictive validity of the PRF against basic flying training.

## METHOD

### Subjects

7. The PRF was administered to 1960 subjects, male and female, from diverse ethnic backgrounds. All were less than 30 years of age and had applied for aircrew training in the Canadian Forces. From this initial group of 1960, a subgroup of 58 was selected. These comprised the male Anglophones who had been selected for, and subsequently attended, basic flying training. (The small number was due in part to the long wait between the ASC processing and the start of flight training.) A further 7 were dropped from the analyses because of their failure at 3CFFTS for other than flying reasons (e.g., medically unfit or failed to reach the required military or academic course standard). The sample size was, therefore, reduced to 51.

### Procedure

8. All candidates who were processed through the ASC during a two year period 1978 to 1979 were administered the PRF. The individual Social Insurance Numbers (SIN) were used for identification purposes. In April 1981 a search was made of data files held at the Canadian Forces Personnel and Applied Unit and personnel from the original group who had subsequently attempted basic flying training were identified. This group (N=58) were then matched with their PRF scores using SIN's. All cases of matched data on females, Francophones, infrequency scores of more than 3, or individuals who had failed flying training for other than ability reasons, were removed. The final analysis was then carried out using the remaining 51 cases as detailed above.

9. The Statistical Package for the Social Sciences (SPSS) was used to analyze the data. The 51 cases were classified in terms of performance during basic flying training (Pass/Fail) and of Previous Flying Experience (PFE/NPFE). Two analyses involving scores on the Desirability trait were undertaken, enabling t-test analyses to be applied to each of the 22 dimensions on the PRF. The number of cases involved is shown in Table 1.

10. Only comparisons involving groups with more than 10 subjects were deemed suitable for analysis, thus eliminating 2 of the original 6 subgroups from the analysis.

## RESULTS

11. No significant differences were found between Group 1 and Group 2 when their scores were compared on the 22 dimensions of the PRF. (TABLES 2-5)

## DISCUSSION

12. This study indicates that the PRF was unable to predict subsequent performance on flight training. It should be noted however that a small number of candidates was involved and for some of the groups the number of subjects is so small that no conclusions are possible. Perhaps a final verdict on the efficacy of the PRF should be suspended until further data accrue.

## REFERENCES

Jackson, D.N. Personality Research Form Manual. Research Psychologists Press, Inc., 1967.

Nie, N.H. Statistical Package for the Social Sciences. McGraw-Hill, 1975.

FIGURE 1

FREQUENCY DISTRIBUTION OF DESIRABILITY SCORES: ANGLOPHONE (N=1568)

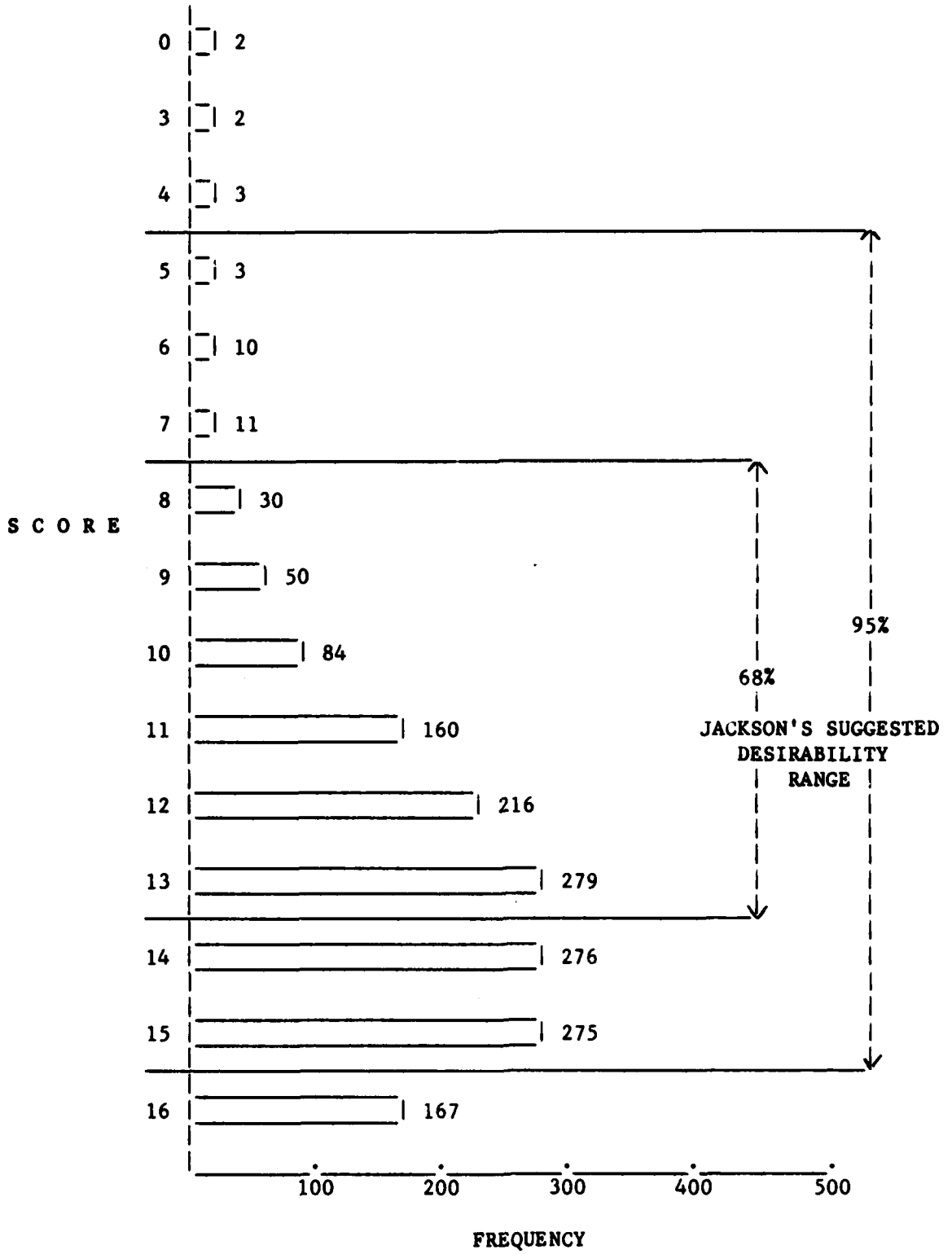


FIGURE 2

FREQUENCY DISTRIBUTION OF DESIRABILITY SCORES: FRANCOPHONE (N=394)

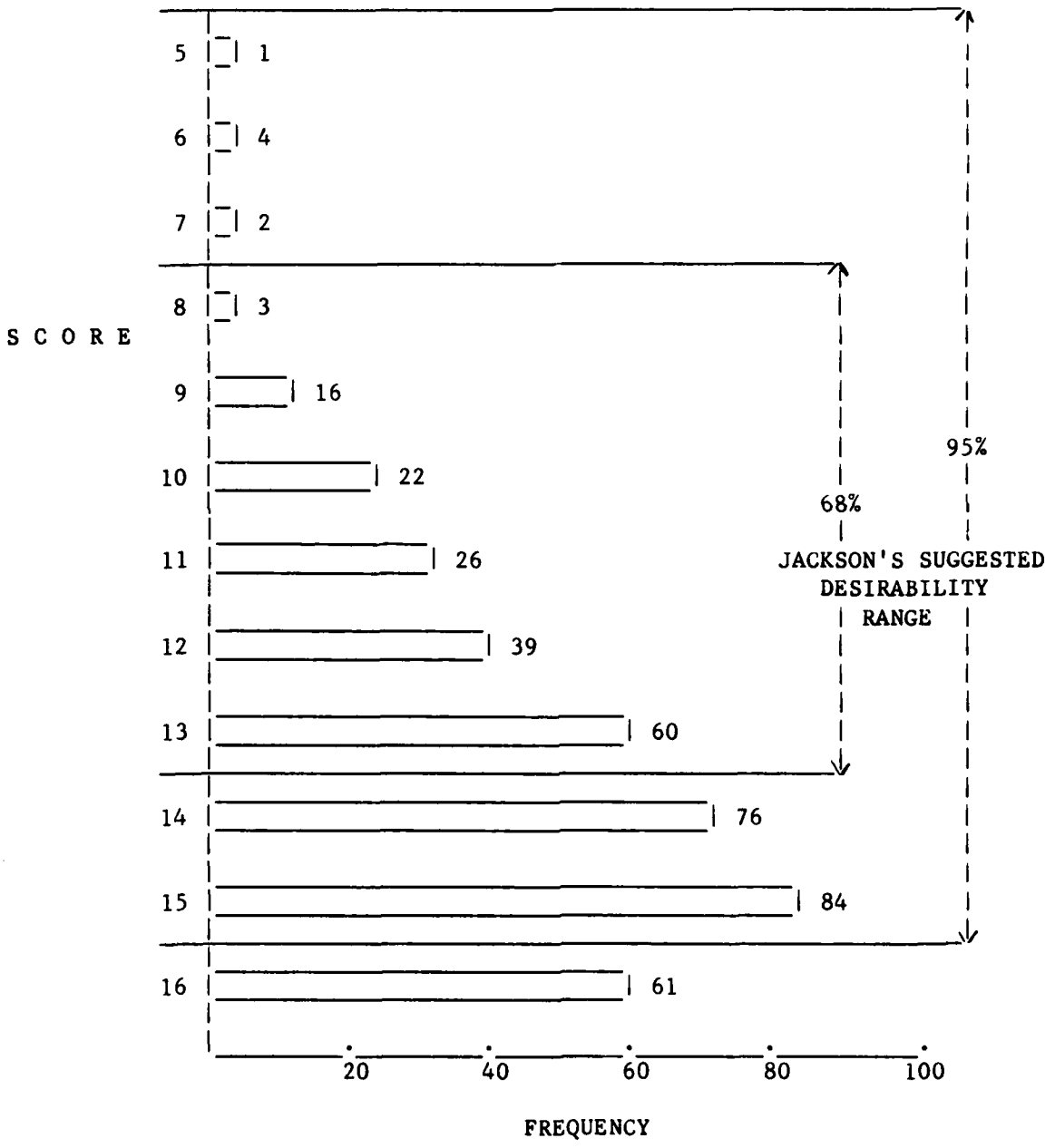


TABLE 1

THE NUMBER OF SUBJECTS IN EACH ANALYSIS GROUP
AFTER SUBDIVISION BASED ON
PREVIOUS FLYING EXPERIENCE AND FLYING TRAINING SUCCESS

| | | PREVIOUS FLYING EXPERIENCE | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | RANGE OF DESIRABILITY SCORES | YES | | NO | | TOTAL FLYING TRAINING | | GRAND TOTAL |
| | | FLYING TRAINING | | FLYING TRAINING | | | | |
| | | PASSED | FAILED | PASSED | FAILED | PASSED | FAILED | |
| ANALYSIS 1 | RESTRICTED TO 4-16(INCL) | 16 | 2 | 16 | 13 | 32 | 15 | 47 |
| ANALYSIS 2 | ALL | 18 | 2 | 18 | 13 | 36 | 15 | 51 |

TABLE 2

RESULTS OF T-TEST USING
GROUP 1 (PASSED FLYING TRAINING) AND GROUP 2 (FAILED FLYING TRAINING)
FOR EACH OF THE PRF TRAITS

| | GROUP 1 PASS N = 16 | | GROUP 2 FAIL N = 13 | | | |
|---|---|---|---|---|---|---|
| PRF TRAIT | $\bar{X}$ | SD | $\bar{X}$ | SD | t | P |
| ABASEMENT | 7.00 | 3.10 | 8.38 | 2.69 | -1.27 | .22 |
| ACHIEVEMENT | 12.15 | 3.26 | 13.38 | 1.76 | -1.13 | .27 |
| AFFILIATION | 10.56 | 3.08 | 11.54 | 3.23 | -0.83 | .41 |
| AGGRESSION | 7.19 | 3.41 | 6.38 | 2.90 | 0.67 | .51 |
| AUTONOMY | 7.56 | 2.22 | 6.69 | 2.78 | 0.94 | .36 |
| CHANGE | 10.63 | 1.86 | 10.92 | 2.43 | -0.37 | .71 |
| COGNITIVE STRUCTURE | 10.06 | 3.15 | 10.77 | 3.35 | -0.58 | .56 |
| DEPENDENCE | 4.56 | 2.66 | 5.15 | 4.24 | -0.46 | .65 |
| DOMINANCE | 12.88 | 2.58 | 12.92 | 2.66 | -0.05 | .96 |
| ENDURANCE | 12.31 | 3.03 | 12.77 | 2.35 | -0.45 | .66 |
| EXHIBITION | 9.75 | 3.42 | 7.54 | 3.18 | 1.79 | .09 |
| HARM AVOIDANCE | 4.44 | 3.10 | 4.69 | 1.75 | -0.26 | .79 |
| IMPULSIVITY | 4.44 | 3.81 | 4.62 | 2.79 | -0.14 | .89 |
| NURTURANCE | 9.44 | 3.10 | 11.08 | 2.33 | -1.58 | .13 |
| ORDER | 10.44 | 4.37 | 10.23 | 2.39 | 0.15 | .88 |
| PLAY | 8.13 | 2.58 | 7.70 | 3.03 | 0.41 | .68 |
| SENTIENCE | 7.81 | 3.19 | 8.00 | 2.00 | -0.18 | .86 |
| SOCIAL RECOGNITION | 7.25 | 3.28 | 7.31 | 3.52 | -0.05 | .96 |
| SUCCORANCE | 4.69 | 2.87 | 5.15 | 2.64 | -0.45 | .66 |
| UNDERSTANDING | 9.25 | 2.91 | 9.15 | 2.51 | 0.09 | .93 |
| INFREQUENCY | 0.38 | 0.72 | 0.07 | 0.28 | -1.41 | .17 |
| DESIRABILITY | 13.06 | 1.69 | 13.15 | 1.28 | -0.16 | .87 |

SUBJECTS ALL OBTAINED DESIRABILITY SCORES
GREATER THAN 4 AND LESS THAN 16 AND ALL HAD
NIL PREVIOUS FLYING EXPERIENCE.

TABLE 3

RESULTS OF T-TEST USING
GROUP 1 (PASSED FLYING TRAINING) AND GROUP 2 (FAILED FLYING TRAINING)
FOR EACH OF THE PRF TRAITS

| | GROUP 1 PASS N = 32 | | GROUP 2 FAIL N = 15 | | | |
|---|---|---|---|---|---|---|
| PRF TRAIT | $\bar{X}$ | SD | $\bar{X}$ | SD | t | P |
| ABASEMENT | 6.78 | 2.86 | 8.27 | 2.58 | -1.71 | .09 |
| ACHIEVEMENT | 13.00 | 2.75 | 13.67 | 1.80 | -0.86 | .40 |
| AFFILIATION | 10.16 | 2.64 | 11.33 | 3.40 | -1.30 | .20 |
| AGGRESSION | 7.63 | 2.92 | 6.60 | 2.85 | 1.13 | .26 |
| AUTONOMY | 7.41 | 2.12 | 7.13 | 2.85 | 0.37 | .72 |
| CHANGE | 10.38 | 1.91 | 10.93 | 2.25 | -0.88 | .38 |
| COGNITIVE STRUCTURE | 10.28 | 3.14 | 11.00 | 3.16 | -0.73 | .47 |
| DEPENDENCE | 4.53 | 2.46 | 4.93 | 4.01 | -0.42 | .67 |
| DOMINANCE | 13.38 | 2.37 | 13.13 | 2.59 | -0.32 | .75 |
| ENDURANCE | 12.34 | 2.66 | 12.87 | 2.20 | -0.66 | .51 |
| EXHIBITION | 9.19 | 3.05 | 7.67 | 3.11 | 1.58 | .12 |
| HARM AVOIDANCE | 4.84 | 3.73 | 4.67 | 2.09 | 0.17 | .87 |
| IMPULSIVITY | 3.97 | 3.09 | 4.40 | 2.64 | -0.47 | .64 |
| NURTURANCE | 9.53 | 2.50 | 10.93 | 2.31 | -1.83 | .07 |
| ORDER | 10.63 | 3.98 | 10.60 | 2.41 | 0.02 | .98 |
| PLAY | 7.41 | 2.75 | 7.60 | 2.85 | -0.22 | .83 |
| SENTIENCE | 8.16 | 3.25 | 8.13 | 2.03 | -0.03 | .98 |
| SOCIAL RECOGNITION | 7.31 | 3.04 | 7.47 | 3.42 | -0.16 | .88 |
| SUCCORANCE | 5.09 | 2.84 | 5.00 | 2.73 | 0.11 | .92 |
| UNDERSTANDING | 9.84 | 2.97 | 9.33 | 2.44 | 0.58 | .57 |
| INFREQUENCY | 0.41 | 0.67 | 0.07 | 0.26 | 1.90 | .06 |
| DESIRABILITY | 12.91 | 1.63 | 13.13 | 1.25 | -0.48 | .64 |

SUBJECTS ALL OBTAINED DESIRABILITY SCORES
GREATER THAN 4 AND LESS THAN 16: PREVIOUS
FLYING EXPERIENCE NOT CONSIDERED.

TABLE 4

RESULTS OF T-TEST USING
GROUP 1 (PASSED FLYING TRAINING) AND GROUP 2 (FAILED FLYING TRAINING)
FOR EACH OF THE PRF TRAITS

| | GROUP 1 PASS N = 18 | | GROUP 2 FAIL N = 13 | | | |
|---|---|---|---|---|---|---|
| PRF TRAIT | $\bar{X}$ | SD | $\bar{X}$ | SD | t | P |
| ABASEMENT | 6.78 | 3.06 | 8.38 | 2.69 | -1.52 | .14 |
| ACHIEVEMENT | 12.44 | 3.13 | 13.38 | 1.76 | -0.98 | .34 |
| AFFILIATION | 10.67 | 3.03 | 11.54 | 3.23 | -0.77 | .45 |
| AGGRESSION | 7.06 | 3.27 | 6.38 | 2.90 | 0.60 | .56 |
| AUTONOMY | 7.33 | 2.25 | 6.69 | 2.78 | 0.71 | .48 |
| CHANGE | 10.78 | 1.80 | 10.92 | 2.43 | -0.19 | .85 |
| COGNITIVE STRUCTURE | 10.06 | 2.98 | 10.77 | 3.35 | -0.63 | .54 |
| DEPENDENCE | 4.39 | 2.57 | 5.15 | 4.24 | -0.62 | .54 |
| DOMINANCE | 12.78 | 2.53 | 12.92 | 2.66 | -0.15 | .88 |
| ENDURANCE | 12.17 | 3.19 | 12.77 | 2.35 | -0.58 | .57 |
| EXHIBITION | 9.33 | 3.69 | 7.54 | 3.18 | 1.41 | .17 |
| HARM AVOIDANCE | 4.56 | 2.94 | 4.69 | 1.75 | -0.15 | .88 |
| IMPULSIVITY | 4.22 | 3.64 | 4.62 | 2.79 | -0.33 | .75 |
| NURTURANCE | 9.61 | 3.03 | 11.08 | 2.33 | -1.46 | .16 |
| ORDER | 10.83 | 4.27 | 10.23 | 2.39 | 0.46 | .65 |
| PLAY | 8.17 | 2.71 | 7.69 | 3.04 | -0.46 | .65 |
| SENTIENCE | 7.72 | 3.01 | 8.00 | 2.00 | -0.29 | .77 |
| SOCIAL RECOGNITION | 7.11 | 3.18 | 7.31 | 3.52 | -0.16 | .87 |
| SUCCORANCE | 4.78 | 2.84 | 5.15 | 2.64 | 0.37 | .71 |
| UNDERSTANDING | 9.11 | 2.78 | 9.15 | 2.51 | -0.04 | .97 |
| INFREQUENCY | 0.33 | 0.69 | 0.08 | 0.28 | 1.27 | .21 |
| DESIRABILITY | 13.39 | 1.85 | 13.15 | 1.28 | 0.39 | .68 |

ALL SUBJECTS HAD NIL PREVIOUS FLYING EXPERIENCE:
ALL DESIRABILITY SCORES ARE INCLUDED.

TABLE 5

RESULTS OF T-TEST USING
GROUP 1 (PASSED FLYING TRAINING) AND GROUP 2 (FAILED FLYING TRAINING)
FOR EACH OF THE PRF TRAITS

| | GROUP 1 PASS N = 36 | | GROUP 2 FAIL N = 15 | | | |
|---|---|---|---|---|---|---|
| PRF TRAIT | $\bar{X}$ | SD | $\bar{X}$ | SD | t | P |
| ABASEMENT | 6.53 | 2.92 | 8.27 | 2.58 | -2.00 | .05 |
| ACHIEVEMENT | 13.08 | 2.61 | 13.67 | 1.80 | -0.79 | .43 |
| AFFILIATION | 10.08 | 2.86 | 11.33 | 3.39 | -1.35 | .19 |
| AGGRESSION | 7.64 | 2.81 | 6.60 | 2.85 | 1.20 | .24 |
| AUTONOMY | 7.25 | 2.17 | 7.13 | 2.85 | 0.16 | .87 |
| CHANGE | 10.42 | 1.94 | 10.93 | 2.25 | -0.82 | .41 |
| COGNITIVE STRUCTURE | 10.39 | 3.04 | 11.00 | 3.16 | -0.65 | .52 |
| DEPENDENCE | 4.55 | 2.41 | 4.93 | 4.01 | -0.42 | .68 |
| DOMINANCE | 13.36 | 2.32 | 13.13 | 2.59 | 0.31 | .76 |
| ENDURANCE | 12.25 | 2.79 | 12.87 | 2.20 | -0.76 | .45 |
| EXHIBITION | 8.97 | 3.18 | 7.67 | 3.11 | 1.35 | .19 |
| HARM AVOIDANCE | 4.89 | 3.58 | 4.67 | 2.09 | 0.22 | .82 |
| IMPULSIVITY | 3.86 | 2.93 | 4.40 | 2.64 | -0.62 | .54 |
| NURTURANCE | 9.42 | 2.58 | 10.93 | 2.31 | -1.97 | .06 |
| ORDER | 10.86 | 3.89 | 10.60 | 2.41 | 0.24 | .81 |
| PLAY | 7.31 | 2.84 | 7.60 | 2.85 | -0.34 | .74 |
| SENTIENCE | 8.08 | 3.08 | 8.13 | 2.03 | -0.06 | .95 |
| SOCIAL RECOGNITION | 7.31 | 2.96 | 7.47 | 3.42 | -0.17 | .87 |
| SUCCORANCE | 5.06 | 2.77 | 5.00 | 2.72 | 0.07 | .95 |
| UNDERSTANDING | 9.72 | 2.90 | 9.33 | 2.44 | 0.46 | .65 |
| INFREQUENCY | 0.36 | 0.64 | 0.07 | 0.26 | 1.72 | .09 |
| DESIRABILITY | 13.25 | 1.83 | 13.13 | 1.25 | 0.23 | .82 |

NO RESTRICTION IS IMPOSED FOR DESIRABILITY
SCORE OR PREVIOUS FLYING EXPERIENCE.

AD P001387

van Rijn, Paul, US Office of Personnel Management, Washington, DC. (Wed. P.M.)

## Self-Assessment in Employee Selection and Placement

A self-assessment questionnaire was administered to 474 firefighter applicants in conjunction with a written test of cognitive abilities. The applicants were asked to make self-assessments of important cognitive and noncognitive abilities required to perform firefighter work. They were also asked to rate their willingness to perform some less desirable yet critical firefighter tasks and to rate the extent of their pre-employment knowledge about firefighting. The self-assessments, although inflated as expected, demonstrated considerable variance. An analysis of the results showed that there are few differences in the mean self-assessments of abilities for black and white applicants and that the differences are limited largely to noncognitive components of the firefighter job. More significant differences were found, however, in the expressed willingness of white and black applicants to perform less desirable firefighter tasks. Correlations between the self-assessed abilities and corresponding scores on the written test were low. Although not without promise, caution is advised in the use of self-assessments, particularly in employee selection.

PREVIOUS PAGE IS BLANK

# SELF-ASSESSMENT IN PERSONNEL SELECTION AND PLACEMENT

Paul van Rijn

U.S. Office of Personnel Management
Office of Personnel Research and Development

This paper describes the first phase of a study to investigate the psychometric properties of self-assessment procedures and to determine to what extent self-assessments made by job applicants can be useful in decisions concerning personnel selection and placement. To date there have been few studies of self-assessment in the employment context and evaluation of its utility in that context have had to come for the most part through inferences from studies conducted in other contexts.

This study was conducted in part to comply with Uniform Guideline requirements to investigate alternative selection procedures, to determine their validity, and to evaluate their adverse impact in a personnel selection context. Other contexts have already demonstrated the utility of self-assessments (e.g., Davey, 1980; Thornton, 1978; and Shraugher & Osberg, 1981), however, their use in personnel selection has received little attention. Part of this lack of interest in self-assessment in personnel selection is concern about leniency or overestimation of the rater's own abilities. Although there is considerable support for this concern, overestimation is not always found (Heneman, 1974) and does not significantly affect the validity of the self-assessments if the overestimations are relatively constant from rater to rater (Primoff, 1979).

Another reason self-assessments have received so little attention in the past has been the assumption that when a self-asssessment is at variance with another measure of the same ability, it is the self-assessment that is inaccurate. This assumption is being increasingly challenged as it is being recognized that self-assessments may be alternative perspectives of the same phenomenon and that total agreement is not always likely, nor necessary (e.g., Schneider, 1977). Additionally, it is being realized that job applicants possess a much larger data base about themselves from which to draw inferences than do most external evaluators or records. Not only is this data base larger, but it also contains all the important private information about situational or motivational determinants that can have such a significant bearing on future performance (e.g., Monson & Snyder, 1977).

---

Paper presented at the 23rd Annual Conference of the Military Testing Association in Washington, D.C., October 1981.

The opinions expressed in this paper are those of the author and do not necessarily represent those of the U.S. Office of Personnel Management.

## DESIGN OF THE STUDY AND RESULTS

Thirty-three (33) self-assessment questions were prepared and administered to 456 firefighter applicants in a large metropolitan jurisdiction. The questions were derived from a thorough analysis of the entry level firefighter job and represented an extension of the questions used in an earlier study by van Rijn and Payne (1980). In that study, self-assessments were made on fourteen variables or dimensions. Although the correlations of the self-assessments and the criterion measures were generally low and not significant, it was postulated that this may have been due to the absence of a clearly defined reference group against which the raters could make their self-evaluations. The relatively small differences in the self-assessments made by black and white participants in that study suggested that self-assessments may significantly reduce adverse impact in comparison to some traditional assessment procedures.

Practical considerations and the complexity of the abilities to be assessed prohibited extensive definition of the abilities themselves (although that would have been desirable), however, the reference group against which the applicants were to evaluate themselves was carefully defined to include "persons of your own age and general background, such as students who went to school with you, persons who served with you in the armed forces, fellow workers, friends and acquaintances." Applicants were asked to compare themselves to this reference group and to rate themselves on 22 abilities, using the following scale:

A. Very much below average
B. Somewhat below average
C. Average
D. Somewhat above average
E. Very much above average

The letters A through E corresponded to spaces on the computer-scored answer sheet and were later converted to the numbers 0 through 4 for statistical analysis.

For another 10 variables the applicants were asked to compare themselves to the same refence group in terms of their "willingness to" perform various activities of the firefighter job. A similar scale was used, ranging from (A) Extremely unwilling to (E) Extremely willing. The last question of the self-assessment questionnaire asked applicants to specify their level of knowledge about the firefighter job.

Ideally in a study of this type, self-assessments should be solicited as an operational part of the entry level examination. This was not possible, and it was announced to the applicants at the beginning of a 3-1/2 hour testing session that <u>some</u> portions (unspecified) of the examination would be used for experimental purposes only. It is not known to what extent applicants perceived the self-assessment questionnaire, which was administered last, to be <u>the</u> experimental portion, but an omit rate of nearly 10% toward the end of the questionnaire suggests that at least some applicants may not have been motivated to the same extent as might be expected from an operational test part.

Table 1 shows the composition of the applicant sample. Of the 456 applicants, 75% were black, since the jurisdiction had recently applied a residency requirement

for service in the fire department. Most (88%) of the applicants who were already residents were black, although 11% of the white applicants were residents in the jurisdiction. As might be expected, most (94%) applicants were male and all of the female applicants were black.

The middle portion of Table 1 shows that in general, black applicants scored about one standard deviation lower on the cognitive abilities test than white applicants, a difference that is not without adverse impact. An extensive validity study reported by van Rijn and Payne (1980) shows, however, that the test is highly predictive (r=61) of job performance and is equally valid for black and white applicants.

In terms of prior knowledge about the firefighter job, the results of the knowledge question reported at the bottom of Table 1 confirm previous findings that black firefighter applicants have significantly less familiarity with the job of firefighting than white applicants. Such a difference significantly affects the amount of self-selection that occurs prior to application for the job and suggests that the applicant group used in this study consists of a highly self-selected group of white applicants and a much less self-selected group of black applicants.

The mean ability ratings for the total sample are shown in Table 2. Although there is some leniency in the ratings, but the effect is not excessive. An "average" ability rating is 2.00, while a mean rating of 3.00 is "slightly above average." Most ratings are within that range. Consistent with discussions in van Rijn (1981), the more verifiable ratings received slightly less inflated ratings. Most significant of the ratings in Table 2 is the virtual absence of a difference between the mean ratings made by black and white applicants. Similarly, passing rates above the arbitrary "passing" point of 2 ("Average"), show almost no adverse impact.

Table 3 shows the "willingness to" ratings. Consistent with the hypothesis that less verifiable variables will manifest greater leniency than variables that can be independently checked, the mean willingness ratings are higher than the ability ratings shown in Table 2. Like the ability ratings, there are few differences in the willingness ratings for black and white applicants.

Table 4 shows the intercorrelations of the major self-assessed cognitive abilities. Except for one, all correlations are significant at the .01 level, although their magnitude is generally low to moderate. The highest correlations are among abilities that were expected to correlate highly, e.g., Remember and Quick Recall (r=.51) and Problem Identification and Judgment (r=.52).

The much higher intercorrelations of the "willingness to" variables shown in Table 5 suggest that these variables are much less distinct and are much more susceptible to halo effects. Of the 45 correlations above the diagonal, 35 are .40 or higher.

Although later phases of this self-assessment study can more directly address the validity of self-assessments for predicting actual job performance, there are some indirect comparisons that are possible here. Table 6 shows the correlations between the six cognitive ability subparts of the entry level predictor test and selected self-assessed variables. Although some of the self-assessed variables correlate significantly with tested scores of the same ability--and in a pattern consistent with the hypothesis that self-assessed abilities will correlate higher

with tested scores of the same or related abilities than different abilities--this was not always the case. Self-assessed Reading ability correlates highest (r=.23) with tested Reading Comprehension, (although there is no statistically significant difference between this correlation and those of self-assessed Reading with the other parts of the test). Self-assessed Oral Directions also correlates highest (r=.17) with the tested score of that ability. Calculations correlates (r=.35) with Using Formulas, but the effect is much less clear for the self-assessed Problem Identification and Judgment abilities. Problem Identification correlates as significantly with other tested variables as it does with tested Problem Identification, while self-assessed Judgment fails to correlate with any tested ability.

Not only did the self-assessed cognitive abilities not correlate very highly with the tested cognitive ability measures, but they failed to correlate much higher than did some of the non-cognitive self-assessed abilities or willingness ratings, i.e., stress, react quickly, and risk/life. It must be noted, however, that these variables are but three from a questionnaire of thirty-three questions and may have reached statistical significance, in part, through chance. The large number of research participants for the study yields statistical significance at the .05 level for correlations as low as .10, however, correlations of such a low magnitiude tend to have little utility.

## CONCLUSIONS

The results of the study suggest that self-assessment procedures may virtually eliminate adverse impact in some situations and resulted in only minimal and acceptable levels of overestimation or leniency in the quasi-operational employment context of this study. However, the important evidences for the validity of the self-assessment procedure are meager. More extensive definition of the self-assessed dimensions, use of numbered rather than lettered scales, use of independent non-cognitive as well as cognitive measures of the rated dimensions, and the use of more directly job-related criterion measures might significantly improve the predictive validity of the self-assessment procedure.

Potential users of the self-assessment procedure must exercise great caution in the use of self-assessment procedures or in the interpretation of the data from this study. The results of this study are based on a unique and atypical applicant group, involve the self-assessment of relatively abstract dimensions, and use a rather incomplete and indirect test of the validity of self-assessment. The use of more concrete, clearly observable, and verifiable dimensions (e.g., typing) may significantly enhance the validity of self-assessment, particularly if self-assessments are evaluated against actual job performance rather than a predictor test.

## REFERENCES


Davey, B. The use of candidate self-ratings as test validation criteria. Paper presented at the Annual Conference of the International Personnel Management Association Assessment Council (IPMAAC), 1980.

Heneman, H. G. Comparison of self and superior ratings of managerial performance. Journal of Applied Psychology, 1974, 59, 638-642.

Monson, T. C. & Snyder, M. Actors, observers, and the attribution process. Journal of Experimental Social Psychology, 1977, 13, 89-111.

Primoff, E. S. The use of self-assessments in examining (PS-79-1). Washington, D.C.: Personnel Research and Development Center, U.S. Office of Personnel Management, 1979.

Schneider, G. E. Multiple rater groups and performance appraisal. Public Personnel Management, 1977, 6, 113-120.

Shrauger, J. S. & Osberg, T. M. The relative accuracy of self-predictions and judgments by others in psychological assessment. Psychological Bulletin, 1981, 90(2),322-351.

Thornton, G. C. III Psychometric properties of self-appraisal of job performance. Paper presented at the annual meeeting of the American Psychological Association in Toronto, Canada, 1978.

van Rijn, P. Self-assessment for personnel examining: An overview. (Personnel Research Report 80-14). Washington, D.C.: Personnel Research and Development Center, U.S. Office of Personnel Management, 1980.

van Rijn, P. & Payne, S. S. Criterion-related validity research base for the District of Columbia Firefighter Test (Personnel Research Report 80-28). Washington, D.C.: Personnel Reserach and Development Center, U.S. Office of Personnel Management, 1980.

TABLE 1

Comparison of Black and White Firefighter Applicants

| Background Variable | Total Sample | | Black | | White | |
|---|---|---|---|---|---|---|
| RACE | | | | | | |
| -number | 456 | | 344 | | 112 | |
| -percent | 100 | | 75 | | 25 | |
| HISPANIC (number) | 2 | | 1 | | 1 | |
| RESIDENT (percent) | 69 | | 88 | | 11 | |
| SEX (percent) | | | | | | |
| -male | 94 | | 92 | | 100 | |
| -female | 6 | | 8 | | 0 | |
| FIREFIGHTER SELECTION TEST | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| (Total Test T-score) | 50.0 | 10.0 | 47.2 | 8.9 | 58.6 | 8.2 |
| KNOWLEDGE ABOUT JOB (percent)[a] | | | | | | |
| -About as much as the general public | 19 | | 24 | | 6 | |
| -A little more than the general public | 27 | | 33 | | 10 | |
| -A lot more than the general public | 28 | | 33 | | 14 | |
| -As much as some firefighters/volunteer work | 17 | | 5 | | 50 | |
| -As much as a trained firefighter/a paid firefighter | 9 | | 5 | | 21 | |

Note. Slight discrepancies in the data are due to rounding. All differences between the black and white subgroups on the FST and KNOWLEDGE variables are significant at the .01 level.

[a]Based on an N of 386. Sixty blacks and 10 whites did not answer this question.

TABLE 2

Comparison of Self-Assessed Ability Ratings of Black and White Firefighter Applicants

| Self-Assessed Variable | Total Sample | | Black | | White | | B/W "Pass" Ratio[a] |
|---|---|---|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | |
| ABILITY TO: | | | | | | | |
| 1. Read | 2.63 | .75 | 2.62 | .77 | 2.65 | .72 | .90 |
| 2. Follow instruct. | 2.57 | .81 | 2.72 | .86 | 2.82 | .62 | .80 |
| 3. Remember | 2.57 | .81 | 2.57 | .84 | 2.56 | .74 | .94[b] |
| 4. Quick recall | 2.63 | .79 | 2.65 | .81 | 2.57 | .71 | .91[b] |
| 5. Calculate (math) | 2.41 | .89 | 2.39 | .90 | 2.45 | .87 | .93 |
| 6. Identify problems | 2.81 | .81 | 2.76 | .85 | 2.96 | .68 | .75 |
| 7. Make judgements | 2.72 | .83 | 2.73 | .84 | 2.69 | .81 | .96[b] |
| 8. Work with people | 3.34 | .86 | 3.36 | .88 | 3.31 | .82 | .99[b] |
| 9. Work with stress | 3.06 | .87 | 3.00 | .90 | 3.24 | .77 | .88 |
| 10. Work under orders | 3.05 | .88 | 3.07 | .91 | 3.00 | .78 | .99 |
| 11. Work independently | 3.09 | .93 | 3.10 | .96 | 3.04 | .80 | 1.00 |
| 12. Act responsibly and dependably | 3.15 | .87 | 3.13 | .90 | 3.22 | .74 | .92 |
| 13. Drive a car | 3.16 | 1.04 | 3.15 | 1.07 | 3.17 | .93 | .93 |
| 14. React quickly | 3.09 | .84 | 3.05 | .88 | 3.21 | .68 | .88 |
| 15. Sustain physical activity | 3.01 | .85 | 3.08 | .87 | 2.77 | .75 | .82[b] |
| 16. Exert maximum force | 2.75 | .84 | 2.76 | .86 | 2.71 | .77 | .95[b] |
| 17. Coordinate body | 3.00 | .87 | 3.05 | .88 | 2.86 | .80 | .84[b] |
| 18. Maintain balance | 2.88 | .87 | 2.91 | .90 | 2.77 | .74 | .97 |
| 19. Find way | 2.76 | .86 | 2.70 | .88 | 2.90 | .79 | .79 |
| 20. Concentrate | 2.73 | .82 | 2.74 | .82 | 2.70 | .81 | .93 |
| 21. Work/high places | 2.69 | .95 | 2.58 | .94 | 3.01 | .89 | .69 |
| 22. Work/confined | 2.78 | .91 | 2.70 | .91 | 3.03 | .83 | .74 |

Note. Since every applicant did not answer every question on the Self-Assessment Questionnaire, the data in the table are based on n's that vary slightly from 448 to 456.

[a]"Passing" represents all self-ratings above 2 or "Average."

[b]Passing rate was greater for blacks than whites.

TABLE 3

Comparison of Self-Assessed Willingness Ratings of Black and White Firefighter Applicants

| Self-Assessed Variable | Total Sample | | Black | | White | | B/W "Pass" Ratio[a] |
|---|---|---|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. | |
| WILLINGNESS TO: | | | | | | | |
| 23. Work outside | 3.21 | .81 | 3.18 | .82 | 3.33 | .78 | .97 |
| 24. Work shifts | 3.29 | .80 | 3.23 | .81 | 3.50 | .75 | .94 |
| 25. Risk injury/ property | 2.41 | 1.14 | 2.32 | 1.14 | 2.67 | 1.11 | .81 |
| 26. Risk injury/ life | 3.35 | .86 | 3.28 | .87 | 3.57 | .77 | .91 |
| 27. Work on high places | 3.10 | .87 | 3.00 | .88 | 3.38 | .78 | .89 |
| 28. Work in confined places | 3.14 | .80 | 3.06 | .78 | 3.38 | .81 | .92 |
| 29. Work in dark, unfamiliar places | 3.04 | .87 | 2.91 | .85 | 3.43 | .79 | .82 |
| 30. Perform routine maintenance | 3.28 | .83 | 3.26 | .86 | 3.36 | .73 | .94 |
| 31. Work at maximum strength for long time | 3.21 | .86 | 3.16 | .91 | 3.35 | .70 | .89 |
| 32. Accept hazing or practical jokes | 2.82 | 1.06 | 2.69 | 1.11 | 3.19 | .81 | .79 |

Note. Although the Self-Assessment Questionnaire was administered to 456 firefighter applicants, the number of applicants responding to the "willingness to" questions varied slightly from 406 to 417.

[a]"Passing" represents all "Slightly willing" and "Extremely willing" ratings, i.e., ratings of 3 and 4.

TABLE 4

Intercorrelations of Self-Assessed Cognitive Abilities

| Variable | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 1. Read | .41 | .30 | .30 | .31 | .33 | .29 |
| 2. Follow instructions | --- | .26 | .24 | .29 | .36 | .27 |
| 3. Remember | | --- | .51 | .26 | .29 | .35 |
| 4. Quick recall | | | --- | .17 | .37 | .42 |
| 5. Calculations (math) | | | | --- | .20 | .11 |
| 6. Problem identification | | | | | --- | .52 |
| 7. Judgment | | | | | | --- |

Note. All correlations are significant at the .01 level, except the .11 correlation between Calculations(5) and Judgment(7).

TABLE 5

Intercorrelations of Self-Assessed Willingness Variables

| Variable | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1. Work outside | .59 | .33 | .45 | .56 | .62 | .60 | .55 | .51 | .40 |
| 2. Work shifts | --- | .31 | .45 | .52 | .57 | .55 | .55 | .46 | .38 |
| 3. Risk/property | | --- | .41 | .32 | .32 | .43 | .29 | .28 | .20 |
| 4. Risk/life | | | --- | .45 | .42 | .50 | .42 | .42 | .33 |
| 5. High places | | | | --- | .66 | .63 | .52 | .53 | .36 |
| 6. Confined places | | | | | --- | .65 | .58 | .46 | .34 |
| 7. Dark places | | | | | | --- | .58 | .51 | .43 |
| 8. Routine work | | | | | | | --- | .50 | .44 |
| 9. Maximum strength | | | | | | | | --- | .46 |
| 10. Hazing | | | | | | | | | --- |

Note. All correlations are significant at the .01 level.

TABLE 6

Intercorrelations of Selected Self-Assessed and Tested Variables

| Self-Assessed Variables | | FST TEST SCORES | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Read. comp. | Use Formu-las | Judg-ment | Rea-son. | Prob. Ident. | Oral dir. | Total |
| 1. | Read | .23 | .18 | .20 | .21 | .20 | .09 | .22 |
| 2. | Follow instructions | .08 | .04 | .10 | .12 | .08 | .17 | .12 |
| 3. | Remember | .02 | .05 | .08 | .10 | .08 | .03 | .07 |
| 4. | Quick recall | .04 | .01 | .08 | .05 | .11 | .00 | .07 |
| 5. | Calculations (math) | .12 | .35 | .15 | .34 | .16 | .21 | .23 |
| 6. | Problem identification | .19 | .09 | .15 | .12 | .18 | .06 | .17 |
| 7. | Judgment | .00 | -.09 | -.01 | .00 | .00 | -.06 | -.03 |
| 8. | Stress | .15 | .08 | .13 | .07 | .20 | .09 | .16 |
| 9. | React quickly | .15 | .04 | .08 | .06 | .14 | .08 | .12 |
| 10. | Risk/life | .13 | .06 | .15 | .18 | .08 | .11 | .13 |
| 11. | Firefighting knowledge | .21 | .10 | .18 | .12 | .12 | .12 | .18 |

Note. Correlations equal or greater than .10 and .12 are significant at the .05 and .01 level, respectively. All self-assessed variables not shown did not correlate significantly (p less than .01) with the Total FST score.

AD P001388

# AN MTA PUBLISHING CHALLENGE: A TEXTBOOK CONCEPT AND PROSPECTS

Raymond O. Waldkoetter
US Army Research Institute for the Behavioral and Social Sciences[1]
Fort Sill Field Unit, P.O. Box 33066, Fort Sill, Oklahoma 73503

## Summary

Each year the customary proceedings are published for distribution following the annual Military Testing Association (MTA) Conference. Only a modest amount of editing and transition discussion among subject-matter sections and papers are practical since the proceedings conform to a topical program plan and size. With a growing diversity of material found within topical areas and across the proceedings, there is an immediate challenge to edit selected material to show a wider scientific application with an integrated structure for key subject-matter sections.

A textbook plan, as a response to this challenge, is in progress to work toward a definitive volume dealing with military personnel assessment research (1958-1983), paralleling reporting activities through 25 years of MTA programs. Several characteristics will make an MTA text noteworthy, largely because most presentations to be drawn on are well done and currently original in view of the reported topics. Major sections will describe interservice and international research priorities, military and related research and program management results, mutual and dissimilar problems and solutions, and present a critical analysis of future directions.

The growing scope and quality of MTA participation indicate there is a continuing need to report the proceedings as efficiently and professionally as possible. Certain implications from deliberations of the MTA Publishing Review Group can furnish advisory assistance for innovative MTA publication practices as well as preparing a scholarly foundation for editing one or more longitudinal texts.

## INTRODUCTION

Publication of annual proceedings is and has been a strong motivation to produce MTA papers and presentations. The sharing of professional and technical experiences gives the participants a ready audience to propose,

---

[1]The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.



PREVIOUS PAGE IS BLANK

review and critique research or studies in a working atmosphere. After 23 years MTA has extended its invitation and participation to annually attract nearly 300 or more military and civilians to explore current topics having specific and general interest in the field of personnel assessment psychology. Coordination responsibility to plan an annual conference, direct a program, and collect the material to publish the proceedings is, to say the least, a formidable yet rewarding challenge.

In the early years of MTA these gatherings were arranged with more of the atmosphere of a college class reunion. It was never expected then that our neat little stack of papers and banquet speech would signal a trend toward a two volume proceedings of 1095 pages, our Canadian conference produced last year (MTA, 1980) nor the 995 pages the US Navy host published the year before (MTA, 1979). A discussion late this summer (Birdsall, 1981) helped estimate, for example, the contrasting length of proceedings for the second annual MTA Conference (1960) which was about 84 pages. With a publication trend that seems to encourage a rather large volume or volumes, we should seriously begin to think through the various publication choices that can guide the development of particular features regarding quantity and quality control in what is selected for MTA format and content.

Publication decisions affecting the length of MTA Proceedings or structure of articles must be identified so that greater advantage may be taken of the rather prolific reporting achieved each year. Where the proceedings furnish an opportunity to deliver papers or present programs, the emphasis is largely on working level actions instead of fully examined and assimilated technical products. There are several other possible ways proceedings might vary in format and content arrangements depending on the objectives selected. As minimal manuscript guidance this year our MTA (1981) host, the Army Research Institute, has for the first time distributed "instructions to authors for MTA manuscripts." A step taken thereby confirming that if we need to structure manuscripts, perhaps due regard for proceedings composition has become of timely concern also.

Should any proposed alternatives indicate inherent constraints in publishing proceedings, then it would seem reasonable to ask: Could there be other publications conceived in MTA and published through MTA or independently? And I would say categorically, another abstracting approach alone would be unacceptable since the article content could not usually be available through a selective retrieval system.

Certain constraints are natural, I feel, with any organizational proceedings. First, when an open call for papers is announced the program size and eventually plan for topical order are dependent on nominal title, topics and degree of paper or panel technical preparation and documentation. Secondly, only a modest amount of editing and transition discussion will be practical to bring papers together more coherently within topical areas and when areas or sections are put in a given program order.

As the diversity of material has increased even in subject-matter sections and across the proceedings, an immediate challenge is perceived to have the proceedings or some alternative derived from proceedings selectively edited to offer a more generalized application in the field of psychology. An integrated structure of key topical areas using independently composed articles or chapters is conceivable with the synthesizing effect of introductory and transitional comments. An attempt to edit a group of papers simply taken from a sample of MTA Proceedings which would include largely the initial papers does not achieve the concept I have in mind, and would most likely be of limited interest beyond the military personnel and training situations. Program scope and material diversity are not in and of themselves handicaps. But the great bulk of material assembled by program sequence can serve most often only as an authentic history and technical reference for participants and colleagues normally familiar with the purpose of the subject-matter content.

Making the effort to see beyond the cumulative proceedings annually inspired under the obvious pressure of on-going operational requirements, does call for an altered perspective, if not a state of consciousness. This adjustment is needed to decide what publication alternative would appear new, instructive, and distill the principal technical trends observed through MTA Conferences. An alternative concept has gradually emerged, which is being refined, that suggests a textbook structure to describe military personnel research from the perspective of MTA experiences and then generalize to academic and other settings.

## A TEXTBOOK PLAN

The MTA Publishing Review Group was designated following the 21st Conference and was the result of an Ad Hoc committee created to explore the development of a publication, preferable a textbook in hardcover (MTA, 1979). Our first objective, I determined as appointed chairman, was try to mold a concept which would document the major professional contributions contained in the proceedings of previous conferences. At the 22nd Conference in Toronto a plan of action (1980) was presented including a global definition of the task and tentative text outline with basic author guidelines. This committee plan was the basis for selecting prospective editors to further develop the text concept and help revise the preliminary content outline. Along with the tentative outline certain crucial milestones were projected to develop article or chapter outlines for the subject-matter areas or major textbook sections.

A textbook plan is being carried out where the book sections are defined by co-editors who obtain the interested authors, and we all industriously strive to put flesh on a definitive volume dealing with military personnel assessment research (1958-1983) paralleling the MTA programs through 25 years of reporting activities. Four major sections are identified in the text outline: Occupational Analysis and Research, Personnel Measurement and Evaluation, Training Methods and Programs, and Organizational Assessment and Technology. Two optional sections were identified to add an introductory MTA background perspective and a concluding

summary of military personnel research utility and future development. To date our endeavor, not unlike the proceedings in part, has impressed us also with the grave magnitude of bringing so much digested material into one manageable book that would adequately treat these sections of complex subject-matter.

An MTA text should be noteworthy because of several characteristics derived from the background of the organization and proceedings. The material presented in almost every instance is well done since it describes the direct experience of individuals and organizations, candidly confronting their research and study problems before conclusions and recommendations are rationalized to fit any inevitable operational compromises or changes in product functions. Of course this written material may profit from some editing, yet a raw originality is evident and the current importance of the reported topics increases their technical relevance. From the MTA organizational affiliations an atmosphere is encouraged to deliver information having a realistic problem solving thrust, rather than exploiting findings which may or may not guarantee rigorous application due to overly elegant technical conditions and logistical support.

Although there is a diversity of material, strong interest groups have tended to become visible and establish communication to delineate prevailing issues. So that there are distinctive groups of interest in spite of the problem of sequencing the proceedings material in the most reasonable order. It is out of this familiarity participants have with so many content areas and their timely individual technical products that new articles or chapters can develop to show a longitudinal analysis of military personnel research and future requirements. Editorial initiative and author discipline are the prime ingredients to translate our plan into a cohesive arrangement of selectively reviewed material extracting the best proceedings highlights with a synthesis of critically discussed MTA contributions and their wider scientific application.

During the 21st MTA Conference (1979) our efforts were spent on just learning whether enough Steering Committee members would have an interest in the textbook concept to justify development of a plan for publication. The notion of publishing something in addition to the proceedings was and is supported, but molding an editorial board to breathe life into this concept took most of the year before the 22nd Conference (1980). And in fact a designated board of editors and an acceptable book prospectus were not prepared in depth until after several good working sessions as recorded in the closing minutes of the 1980 Steering Committee meeting. Draft working papers had been prepared for Steering Committee review before the conference and a survey conducted to focus on the decision points needing thorough discussion and agreement. Fortunately, during the brief time available the book concept was revised, the editorial board and Publishing Review Group confirmed, the preliminary milestone dates selected, and the Steering Committee approval given to circulate our prospectus to publishers.

The 1981 Conference will attend to reviewing the updated status of the editorial board as to sections or subject-matter areas, outlining details, and the commitment received from prospective authors. Again, the first step is to see where we are in reaching better definition of our task ahead during this conference and the coming year. Needless to say, the agreement on new milestones must decidely be an urgent part of our agenda here to assure that there is ample productivity to write and edit the planned sequences of articles or chapters.

Also this year after contacting a relatively respectable number of prospective publishers, it has been advised that a text, or series of texts optimistically suggested, clarify features that are distinctively promoted by the MTA programs. Major book sections and topical areas can describe interservice and international research priorities from the military and related research experiences while reviewing the impact on program management results. Mutual and dissimilar problems and solutions will permit a contrast of research systems and functions and explain the differences observed. A critical analysis of what we did, currently do, and intend to do in future activities will present a definitive military personnel research record and verify the effective continuity achieved in the reported proceedings.

## REPORTING AND PUBLISHING

While a singular effort is planned to attain the textbook publishing objective, it is not meant in anyway to detract from the annual reporting of the MTA proceedings. Reporting of the papers and programs given each year lends both incentive and feedback in specific areas at this working-level conference to expand our military and professional skills (Ellis, 1972). The content of the proceedings in a sense has to capture the concurrent task requirements of each participating individual and organization where personnel problems are assigned for investigation. Publishing in my discussion, then, is an alternative means by which the Military Testing Association can review and evaluate its contribution to the long-term promotion of military and professional personnel assessment activities (Abramson, Tittle, & Cohen, 1979).

As was mentioned at the beginning of this paper, the scope of MTA and the enduring quality of participation attract deserved attention in regard to their prodigious capacity for communication and interaction. Membership is voluntary from year to year but indicates persisting work orientations, which act as vehicles to expand presentations without necessarily obtaining the priority of issues to secure the most tangible benefits. Numerous variables operate to cause the proceedings to be affected by contemporary problems that do not always encourage authors to find a perspective to make their articles integrate past and future circumstances. In reporting the proceedings, I think that most program chairpersons would

agree, it was extremely difficult to anticipate and organize presentations to expressly group and sequence papers by anything like a tightly structured set of subject or topical categories.

In spite of constraints imposed by the reporting format, there are the fundamental criteria of efficiency and professionalism which should recommend exploring potential modifications in publication of the MTA Proceedings. I would like to solve these questions of what and how to immortalize in print, however there is more to driving the solutions than merely setting up new publication aims. Decisions to change the size, use print or direct microfilm, or design edited articles will require, perhaps, at least a new concept of what our proceedings should aspire to. To stimulate this adjustment toward positive reporting changes, the textbook plan for publication continues to stimulate questions about content, style, and size which are answerable only in reference to our library of proceedings and its relationship to the field of personnel assessment psychology.

When the MTA Publishing Review Group has finished discussions with the Steering Committee and completes its working agenda assignments, these deliberations may well have implications generally for MTA publication practices. Certainly wherever the textbook concept experience takes us, innovative planning results can furnish advisory assistance to conceive of new program management techniques in guiding suggested structural changes in the proceedings. Some members of MTA may not see any reason in modifying the proceedings as a key organization activity. Yet it is because of the growing mass of proceedings information that a textbook concept has triggered the collective imagination. Both the prospect of digesting the best of MTA and that of having a commercial text appealing to a larger domain of personnel research and assessment users can precede a new phase of MTA growth, professional identity and publication.

Whenever such an organizational challenge is detected and accepted, there may follow a few unavoidable consequences. Obviously, one consequence may cause a constructive difference in MTA program content and related membership. Another is to verify a realistic and scholarly foundation from which the editing of a military personnel research text follows as an important product, signifying a functionally sound and continuous basis for our affiliation. And still another is that a longitudinal review of the organization is invited where its principal reporting and publishing products are examined.

Our committee was tasked to develop a textbook plan and soon discovered that we had to inventory and review the history and purpose of MTA Proceedings. Being a new publication group we have attempted to design a longitudinal appreciation and critique of the proceedings to create a new type of document. It has been necessary to better understand this organization and analyze the potential risks in producing a marketable

textbook. Many details remain before our plan will finally unleash the designated organizational and professional resources to publish one or more texts after the 1983 Conference. It is felt that no matter what our final published product is, the prospects are highly favorable we shall have progressively modified MTA's publishing concepts, products and organizational objectives.

## REFERENCES


Abramson, T., Tittle, C. K., & Cohen, L. (Eds.). Handbook of vocational education evaluation. Beverly Hills, CA: Sage Publications, Inc., 1979.

Birdsall, W. W. Personal communication. Pensacola, FL: US Naval Education and Training Program Development Center, July 1981.

Ellis, H. C. Fundamentals of human learning and cognition. Dubuque, IA: Wm. C. Brown Co. Publishers, 1972.

Military Testing Association. Proceedings of the second annual conference. Indianapolis, IN: Fort Benjamin Harrison, September 1960.

Military Testing association. Proceedings of the twenty-first annual conference. San Diego, CA: US Navy Personnel Research and Development Center, October 1979.

Military Testing Association. Proceedings of the twenty-second annual conference. Toronto: Canadian Forces Personnel Applied Research Unit, October 1980.

Military Testing Association. Instructions to authors for MTA manuscripts. Alexandria, VA: US Army Research Institute, 1981.

AD P001389

# THE OFFICER CANDIDATE PREPARATORY SCHOOL
## An Intense Program that Doubles as a Personnel Assessment Center

Commander John B. Washbush, Ph.D., Head, Professional Development, Officer Accession Programs Division, Chief of Naval Education and Training, N-122, Naval Air Station, Pensacola, FL 32504, AV 922-4291

Summary: This paper summarizes the development of the Navy's Officer Candidate Preparatory School. Inaugurated on a pilot basis in 1980, two classes have been conducted at the University of North Carolina, Chapel Hill during two consecutive summers. Designed as an initiative to increase the number of minority group members qualified for admission to the Navy's Officer Candidate Schools, this program provides, during an 8-week period, instruction in English, mathematics, chemistry and physics, and naval science. In addition, general military training is provided, including drill, physical training, and general military indoctrination. As the program has evolved during the pilot development phase, it can be considered to be akin to a personnel assessment center, producing an atmosphere very much like an Officer Candidate School and providing for in-depth analysis and assessment of student learning, skill building, and motivation. Seventeen students from the 1980 class qualified for the Officer Candidate School at Newport, RI, and 15 have successfully earned commissions and are presently serving as officers on active duty. Twenty-four graduates from the 1981 class qualified for the October 1981 OCS class. The pilot program is presently under evaluation, but the favorable results achieved in 1980 have led the Navy to decide to establish the school on a permanent basis.

## BACKGROUND

In early 1980, the Chief of Naval Operations (CNO) asked the Chief of Naval Education and Training (CNET) to evaluate the concept and feasibility of establishing an Officer Candidate School (OCS) Preparatory School in 1980. The idea of such a school was rooted in the commitment of the Navy to improve the representation of minority group members in the officer corps. The Navy Recruiting Command had noted that it was not able to select a considerable number of highly qualified candidates for OCS because of low scores on the Officer Aptitude Rating (OAR) examination (Petho). The OAR is an aptitude test, designed for use in screening applicants for OCS, developed for the Bureau of Medicine and Surgery (BUMED) by the Naval Aerospace Medical Research Laboratory (NAMRL). The OAR consists of two sections, an Academic Qualification Test (AQT) and a Mechanical Comprehension Test (MCT). Failure to achieve qualifying scores on the OAR was determined to impact negatively on efforts to increase minority enrollment at OCS.

Relying on experience gained in recent years with the Naval Enlisted Scientific Education Program (NESEP) and the Broadened Opportunity for Officer Selection and Training (BOOST) Program, and aided by the consultation of BUMED psychologists, CNET was able, for evaluative purposes, to posit two assumptions: (1) that the student who would attend the school lacked an appropriate background and development in language skills, science, mechanical

comprehension, and mathematics skills required for success on the OAR and in OCS; and (2) that the OCS Preparatory School experience would positively affect student performance on the OAR and at OCS.

The assessment also determined that a pilot school could be established at a Naval ROTC unit during the summer period, relying in part on personnel assigned to the unit, with academic instruction being provided by temporary active duty (TEMAC) Naval Reserve officers who were civilian educators appropriately qualified in the discipline areas. The school would provide, during an 8-week period, instruction in English, mathematics, chemistry and physics, naval science, and general military training. On 25 April 1980, CNO approved the CNET proposal and established a class convening date of 15 June. The NROTC Unit, University of North Carolina-Chapel Hill was chosen as the school site, and, with barely a month's worth of working days, a crash effort began to implement the CNO directive. On 15 June 1980, 31 students began instruction at the first Officer Candidate Preparatory School (OCPS).

## THE 1980 OCPS

The 1980 OCPS was conducted at the University of North Carolina-Chapel Hill during an 8-week period commencing 15 June. The host command for the Navy was the Naval ROTC unit on that campus. Unit personnel serving in key roles included the commanding officer as officer in charge (OIC), the executive officer as assistant OIC, the Marine staff sergeant as military/drill instructor, and unit clerical/administrative personnel. The Naval Military Personnel Command provided the TEMAC services of an enlisted yeoman for administrative support and three Naval Reserve officers as English, mathematics, and chemistry/physics instructors. CNET provided an active duty minority officer, from the NROTC unit at Savannah State College (an historically black college), who acted as naval science instructor and division officer/counselor for the students enrolled.

Because of the short lead time, curriculum development had to occur simultaneously with instruction. The curricular guidance provided to the OIC was predicated on an academic day commencing at 0800 and ending at 1630, Monday through Friday, and at 1200 on Saturday. Instruction was to comprise these academic segments: English including grammar, reading, writing, and listening skills; mathematics including basic arithmetic skills, progressing through college level algebra; science including basic principles of physics, chemistry, and mechanical comprehension; and naval science including naval orientation, introduction to ships and weapons systems, relative motion fundamentals, and naval vocabulary. Military drill and physical fitness training were also to be included. Primary attention was directed to be placed on the academic work in English, mathematics, and science.

The students attending were selected by the Navy Recruiting Command. They were all college graduates, were considered to be excellent candidates for OCS, and had all scored in the range of 24-37 on the OAR (at that time the Recruiting Command was normally requiring scores in excess of 40 for OCS selection). They were all minority group members (24 Black, 2 Hispanic, 5 other), and 8 were female. They were enlisted by the Recruiting Command in pay grade E-3 and designated as Officer Candidates Under Instruction (OCUI). Those who completed the school and were selected would be ordered on to OCS for the October 1980

class convening. Those who failed to complete the school, disenrolled at their own request, or were not selected for OCS would be discharged from the naval service.

The stated objective of OCPS was to attempt to produce an academic and military environment which would not only lead the student to OCS, but would also assist in greatly improving their chances for success at OCS and later. Consistent with this mission, and in light of the stated curricular guidance, the strategy was adopted to conduct OCPS as an intense and comprehensive review of English and mathematics and, since few of the students had taken a college level physics or chemistry course, to concentrate efforts in this area on those concepts which would have the greatest application to naval systems and which would perhaps make the greatest contribution toward success at OCS.

The students were organized into one company divided into two platoons. The typical training day included a physical training period prior to breakfast, a formal colors ceremony, 2 hours of instruction in each academic area, 1 hour of naval science, a drill/athletics period, and supervised study. A formal military environment was established, and students were rotated through a variety of positions of responsibility in the company structure. Students were issued appropriate uniforms, and these were worn throughout the training day. Normal military courtesies were observed. All staff members were also in uniform. Billeting and messing were contractually provided by the university, and the NROTC unit served as the classroom site.

## The English Curriculum

The English instructor administered the Gates-MacGinitie Reading Test, survey F, and an in-class written essay to determine, at the beginning of the course, the general degree of development and skill of the students. Most of the students were evaluated as possessing a competency generally adequate to the ordinary functions of a naval officer. The instructor was therefore able to shift from an intended complete grammar review to a review of common trouble areas. Students with exceptional deficiencies were assigned appropriate individual work. Other major elements of the curriculum included vocabular building, listening and note taking skills, and Navy correspondence. In-course testing was by instructor prepared examinations, and periodic essay papers were assigned and graded.

## The Mathematics Curriculum

The mathematics curriculum was designed to provide students with a thorough review to the level of college algebra. Self-paced, individualized instructional material was used in the form of the Navy's Mathematics correspondence courses and supporting texts. The course was designed to provide a balance of emphasis in these areas: symbols, definitions, review of arithmetic and algebraic operations, problem solving, vectors, and trigonometric functions. The Navy's Basic Machines text was used to provide additional work in practical applications. A positive effort was made to coordinate instruction with the science instructor. Students exhibited a considerable deficiency in understanding geometric concepts, they lacked familiarity with mathematical terms, concepts, and symbols, and they found translation of word problems into equations difficult. Significant amounts of instructional time had to be devoted to these areas of difficulty. Tests were instructor prepared.

## The Science Curriculum

Nearly all of the students had little or no college physics or physical science, and over two-thirds were similarly deficient in chemistry. This area, then, was one in which a substantial amount of new learning had to occur. The course was designed to emphasize physics because of its importance to the study of ships engineering and weapons systems in the OCS curriculum. The chemistry portion of the course was conducted during the first two weeks of OCPS, and the remaining weeks were devoted to physics. Mechanical comprehension was emphasized throughout physics, and the Navy's Basic Machines correspondence course was used to supplement instruction. Physics instruction included vectors, kinematics, laws of motion, static equilibrium, work, energy, fluids, wave motion, electromagnetic waves, the photon, nuclear energy, heat and temperature, and thermal expansion. Considerable emphasis was placed on the topics of force and acceleration, equilibrium problems, torque, work, power, simple machines, pressure, and the gas laws because of the relation to mechanical comprehension. Vectors, trigonometric concepts, and problem solving proved to be persistent areas of difficulty for a majority of students during the physics portion of the course. Considerable flexibility was necessary to allow for reviewing and slower pacing in areas where students might experience difficulties. Only four-fifths of the originally intended topics in physics were covered. Testing was by means of instructor prepared examinations and quizzes.

## The Naval Science Curriculum

Naval science instruction was designed to provide an introduction to the major areas of instruction encountered at OCS. Instruction included naval orientation, ships and weapons systems, relative motion, naval vocabulary and terminology, and piloting. Every attempt was made to integrate naval science instruction into the entire scope of OCPS instruction. An example of this was the use of a film on the basic steam cycle as a training aid in an English class on note taking. A major objective was to relate both OCPS and OCS instruction and training to the duties, responsibilities, knowledge, and skills required of and used by the junior naval officer. Tests were instructor designed and were intended to assess familiarity rather than mastery. A significant problem encountered in this instructional area was the perception held on the part of the students that this material was not of significance in preparing for a retake of the OAR nor in the determination of whether or not the student would continue on to OCS.

## The Military Program

The military program was intended to provide an appropriate military environment that would both improve the capability of the students to successfully complete all OCS military and physical requirements and would enhance the OCPS academic instruction. Training in this area encompassed physical fitness, drill, marksmanship, personnel/barracks inspections, watches and duties, holding organizational billets, and peer evaluation and counseling. This training was integrated, to the extent possible, with naval science instruction. In addition, an objective of the military program was a carry-over of physical conditioning and growing mental confidence in the classroom.

## Results and Evaluation

Twenty-six students completed OCPS and graduated. Of the original 31, four disenrolled at own request and one was academically disenrolled. Seventeen students were recommended for acceptance at OCS and were approved for enrollment by the Recruiting Command. The remaining nine students were processed for discharge. Fifteen of the students ordered to OCS completed the course, were commissioned as ensigns, and are now serving on active duty. One student was academically disenrolled from OCS, and one student disenrolled at own request. The majority of OCPS graduates attending OCS finished in the lower half of their graduating class; however, their performance was subjectively rated by OCS administrative personnel as being equal to that of minorities attending the school. It is interesting to note that one student, who received an OAR score of 28 at the end of OCPS, finished OCS standing 110 in the class of 356, and compiled an academic average of 3.612 (on a 4 scale). An analysis of costs incurred in conducting OCPS determined that the cost per commission of an OCPS graduate, in addition to the normal OCS cost to train, was approximately $6600. The 1980 OCPS program cost approximately $100K.

Detailed statistical analyses were made of OCPS student performance and scores achieved on administrations of the OAR and the Gates-MacGinitie Reading Test. These indicators were evaluated by means of correlational analysis for the students who were commissioned, comparing OCS performance data. While the sample size was too small to enable global conclusions to be drawn, OCPS tests administered and academic performance were not predictive while several OCPS military performance indicators appeared to have been predictive. The OAR scores and its individual subscores did not predict either OCPS or OCS academic and military performance grades. This conclusion is logical in that all OCPS entrants had achieved scores clustered in the lower end of the spectrum. This being the case, the constraint of restriction in range was expected, and did prove, to remove predictive validity.

In spite of the somewhat ad hoc nature of the initial OCPS, a close review of the conduct, operations, and accomplishments of the school led to the inevitable conclusion that the essential utility of OCPS lay in that fact that it provided a suitable environment for the in-depth assessment of potential OCS candidates who are considered marginal by traditional screening criteria, particularly the OAR. The most critical and useful aspects of OCPS are the establishment and maintenance of an OCS-like character. The presence of a pervasive military environment, the highly structured schedule, and the pressures placed on the students were considered to be essential to establishing, building, and assessing the potential and motivation of each student for selection for OCS and, more importantly, for effective functioning as a naval officer. The inevitable conclusion is that, somewhat serendipitously, OCPS is de facto an OCS candidate assessment center.

The assessment center concept is grounded in evaluation of the behaviors of people in environments and facing situations which can logically be encountered in a job or position aspired to. Furthermore, assessment centers have been found to be both valid and legally accepted methods for screening candidates for organizational entry—assuming they are properly designed and supervised by

appropriately trained personnel. In the case of OCPS, the curriculum and training environment was designed based on the best assessment available of the needs of the students for success at OCS. In addition, the supervisory and teaching staff was composed of military personnel who had been both professionally trained and were professionally experienced in the tasks assigned to them at OCPS. These staff members conducted daily, intensive evaluations of the students, and frequent evaluative discussions and reports were shared among the members of the staff. This potential found in the inaugural OCPS had major impact on the design of the second convening in 1981.

## THE 1981 OCPS

Subsequent to completion and evaluation of the 1980 OCPS, CNET recommended, and CNO approved, extending the pilot development period through a second convening of the school at the University of North Carolina. A thirty-student loading was again approved. A number of personnel who participated in the initial OCPS were available for the second convening including the OIC, assistant OIC, English and mathematics instructors, and several administrative support people from the NROTC unit. As a result of having a considerably longer lead time, and with the benefit of hindsight, a number of changes were introduced to the focus and structure of the school. These changes gave naval science equal instructional time compared to the academic areas. In addition, naval science was intentionally made the focus of instruction to the extent possible. This was done to identify and exploit the relevancies of OCPS to the curriculum and training program the students could anticipate at OCS. The changing focus of OCPS is shown in the statements of purposes, goals, and objectives of the program as developed prior to the 1981 convening. These statements are listed in Appendix A.

Except for the changes just noted, the 1981 program was staffed and conducted in a manner essentially consistent with that of 1980. Naval Reserve TEMAC officers taught English, mathematics, and science. A minority officer from the NROTC unit at Southern University and A&M College (an historically black college) was assigned on temporary additional duty as division officer/counselor. However, he was freed of naval science instructional responsibility because of the decision to make greater instructional use of the officer staff of the host unit. The host unit enlisted Marine again served as military/drill instructor. As in 1980, an enlisted yeoman was assigned under TEMAC orders for administrative support. Additional administrative support was provided by the host unit.

On 14 June, 33 students commenced instruction at the 1981 OCPS. All were minority group members (31 Blacks and 2 Hispanics). Twenty-three were male and ten were female. Their OAR scores ranged from 26-40. Those completing the school and selected would attend the October OCS class. Others would, as in 1980, be processed for discharge. Compared to the 1980 class, this group was somewhat older, several being over 30. The students were again organized into a company of 2 platoons, wore uniforms throughout the instructional day, and experienced an encompassing formal military environment. Contract services from the university included berthing, messing, and academic area consulting services. Classes were again conducted in the NROTC Armory.

## The Naval Science Curriculum

As a result of preconvening planning and in light of the experiences of 1980, the decision was made to enlarge the Naval Science portion of instruction to make it equivalant in emphasis with the English, mathematics, and science courses. In addition, to the extent possible, the academic courses would attempt to orient their objectives toward supporting the student's ability to master the naval science curriculum. The goal was to attempt to expose the student to the subject areas taught at OCS. Subject matter covered included naval orientation, naval engineering, naval weapons, navigation, and naval operations. A major problem noted in this curricular area was the difficulty most students had in making practical application of skills and concepts developed in mathematics and science. Great difficulty was encountered with concepts involving spatial perception. The presence of these problems validated the concern and actions taken to integrate the curricular areas with naval science, and they indicate the need for continuing attention to these problems in OCPS curriculum development.

## The English Curriculum

The English curriculum was closely modeled on that of 1980. Instruction was provided in vocabulary, grammar, briefing exercises, Navy correspondence, composition, and listening and note taking skills. As in 1980, the majority of students were moderately skilled in English; however, about 20% showed significant grammar deficiencies. Results of the administration of the Nelson-Denny Reading Test (forms C and D) indicated that 25% of the students read below the 11th grade level in either vocabulary or comprehension. The addition of the briefing exercises, in which students had to prepare, give, and critique briefs on subjects chosen from naval professional journals, was a very effective adjunct to the English program.

## The Mathematics Curriculum

Substantial modifications were made to the initial mathematics curriculum, although the general objectives and scope of the program were consistent with that developed in 1980. Attention was particularly focused on practical problem solving requiring the construction and use of mathematical models (equations). Students with advanced mathematics abilities were able to pursue advanced topics in mathematics, and two did so; the remainder pursued a course of study in basic algebra at the college level. A persistent problem encountered throughout the course was difficulty in comprehension and translation of word problems. Where practical to do so, problems were related to course work in science and in naval science.

## The Science Curriculum

As in 1980, the students were very deficient in prior exposure to physics, physical science, and chemistry. Therefore, this area was again one in which substantial new learning had to occur. The science course was conducted in substantially the same manner as in 1980, an introductory overview of important concepts from chemistry, major emphasis on physics, and, within the physics course, emphasis on topics related to mechanical comprehension. A major change in the curriculum was in the addition of 10 hours of physics laboratory experiments. The Basic Machines correspondence course was again used in support of

physics instruction. Students displayed the same difficulties with mathematical concepts, problem conceptualization, and problem solving that were noted in 1980. These difficulties again required the exercise of considerable flexibility for reviewing and slower pacing. Because of some reduction in classroom time caused by the increased time devoted to naval science and because of time spent in laboratory sessions, some subjects (e.g., magnetic fields, light, and nuclear physics) were forced off the schedule. The importance of physics to the technical nature of many of the studies undertaken at OCS and the problems experienced in OCPS with subject matter crowding, slow rates of student progress, and sequential integration with mathematics and naval science subject matter underscore the fact that the science curriculum will, of necessity, undergo considerable scrutiny and modification as OCPS evolves.

## The Military Program

The military program was essentially the same as that of 1980. The program was supervised by both the assigned division officer, who had no naval science teaching assignments, and the enlisted Marine attached to the NROTC unit. The concept of a pervasive military environment was once again reestablished, and military training included drill, physical training, counseling and evaluation, marksmanship, inspections individual and group competition, watch standing, leadership training, and selected topics from naval administration. The generally positive aspects noted in the 1980 military program were again seen, particularly in the development of positive motivation and personal confidence.

## Results and Evaluation

Twenty-nine students completed the 1981 OCPS, four having withdrawn at their own request. Twenty-four students were approved by the Recruiting Command for continuation to the October OCS class. The remaining five graduates were processed for discharge. Although the OAR was readministered at the end of the course of instruction, these results were virtually unused in the final evaluation of students. In the 1980 class, improved performance on the OAR was given the status of a program objective; in 1981 it was not. The difference in philosophy had no major impact on the performance of either group. In 1980, the class improved an average of 3.25 points and five students scored 40 or higher. In 1981, the class declined an average of 2.54 points and four students scored 40 or higher. At the completion of both classes, therefore, the OIC and the staff of the school based their evaluations on whole-person assessments using all OCPS performance data. The recommendations of the OIC were the product of these detailed and lengthy staff deliberations. Thus, the assessment center concept, postulated after the completion of the 1980 class, became the primary assessment tool in 1981. The incorporation of the concept was made even more valid by the change in curricular focus toward naval science and the conscious effort to make the entire education/training experience clearly relevant to the environment of OCS.

In an effort to provide a greater basis for analysis, a number of different types of tests were administered to OCPS in 1981. These included the Nelson-Denny Reading Test (form C and D), the Strong-Campbell Interest Inventory (merged form of the Strong Vocational Interest Blank, form T325), the Armed Services Vocational Aptitude Battery (ASVAB), the OAR, and the Objective Test and the Activity Inventory under experimental development by American College

Testing (ACT) Program in its College Outcome Measures Project (COMP). Data from all these sources will be analyzed in detail, including an assessment of relationships to both preparatory school and OCS performances. This analysis will be completed in March of 1982, subsequent to completion of the OCS class entered by 1981 OCPS graduates.

While members of the Military Testing Association are generally familiar with most of the tests identified, the ACT COMP program is not likely so well known, and, therefore, some descriptive comments are in order. The College Outcome Measures Project was organized in 1976 by ACT for the purpose of designing, developing, validating, and implementing assessment instruments that would measure general education outcomes (general knowledge and skills), with particular emphasis on the ability of the student to apply knowledge and skills to adult life situations. Participating in the development of COMP were Alverno College, Florida International University, Governors State University, Mars Hill College, Our Lady of the Lake University, the State University System of Florida, the Tennessee Higher Education Commission, Brigham Young University, Delaware County Community College, and Tennessee Technological University. The outcomes to be assessed in COMP included:

-Development of assessment instruments to measure outcomes for which no adequate tests currently existed

-Initial focus of measurement on cognitive characteristics

-Assessment of the more generic college outcomes and not the more discipline- or content-specific outcomes

-Assessment of the learning outcomes expected of the general education components of postsecondary curricula

-Assessment of outcomes relevant to effective functioning in a variety of adult roles

The development of COMP has identified and focused on six major areas of general education: Functioning Within Social Institutions; Using Science and Technology; Using the Arts; Communicating; Solving Problems; and Clarifying Values. Greater elaboration on these outcome areas is provided in Appendix B.

There are three COMP instruments available for use: the Composite Examination; the Objective Test; and the Activity Inventory. The Composite Examination is a series of 15 simulation activities based on realistic stimulus materials. For some questions, examinees must provide their own answers in written or oral form; for others, they must select from among several answers. Scoring is done by faculty raters using standardized scales. Although this examination is modularized, it is lengthy to administer and requires special faculty training for scoring. It was therefore considered too cumbersome and time consuming for use at OCPS.

The Objective Test was designed as a proxy measure for the Composite Examination. ACT has determined that probable group performance on the Composite Examination can be predicted accurately from performance on the Objective Test. This test has the advantages of being less expensive, less time consuming, easier to administer, and it can be computer scored. The Objective Test

has not been evaluated as a tool for making decisions about individuals, and it does not generate four subscores in communicating (writing, speaking, computing, reading) which are provided with the Composite Examination. This test also consists of 15 simulation activities, and all questions are posed in a multiple-choice format. The test is modularized, and it yields a General Education total score plus subscores in the six COMP areas. For experimental purposes, this test was administered to the 1981 OCPS students during the first week of classes.

The COMP Activity Inventory is specifically designed to measure the quality and quantity of individuals' participation in various key activities in each of COMP's six major outcome areas. By implication, it measures acquired general knowledge and skill as well as capabilities and motivational predispositions. This Activity Inventory is modularized, provides scores for General Education and subscores in the six COMP areas, and is computer scored. For experimental purposes, this Inventory was also administered to the 1981 OCPS students.

Analysis of COMP data is incomplete at this time; however, several interesting findings are apparent. On the Objective Test, as a group the 1981 OCPS students performed at the lower end of the spectrum with respect to the norming group (3562 seniors at 41 institutions). Percentiles ranged from 24 (total score) to 33 (communicating) in all seven scoring areas. This is consistent with the scores obtained on the OAR, also at the lower end of the spectrum. Differentiating the OCPS students by major category, social science majors (30th percentile) outscored arts/humanities majors (22nd percentile) outscored natural science majors (15th percentile) on the basis of group mean total score. Most interesting was the fact that, as a group, the 24 students selected for OCS outscored the nine who withdrew or were not selected. This relationship is even more pronounced in the case of seven of those nine. Thus, it appears that it may be possible to establish expectancy tables for successful performance at OCPS using the COMP Objective Test. This finding suggests that further experimental evaluation with that test is highly desirable.

In the case of the Activity Inventory, mean scores ranged from the 26th percentile (using the arts) to the 52nd percentile (communicating) compared to a norming group of 266 seniors at 11 institutions. The mean total score was at the 34th percentile. On a total score basis, arts/humanities majors (46th percentile) outscored social science majors (35th percentile) outscored natural science majors (24th percentile). Data on the Inventory did not suggest predictive potential for a success/nonsuccess criterion. Continued experimental use of the Inventory is doubtful; however, the data from both the Objective Test and the Activity Inventory will be compared to OCS performance criteria before final continued use decisions are made.

## PROGNOSIS

The Chief of Naval Operations has determined that the OCPS will be established as a permanent school in Fiscal Year 83. It is also intended that an interim school will be conducted in 1982. At this writing, siting recommendations are being evaluated. It is anticipated that the permanent school will see an annual input of 400 students in six 8-week classes. While considerable

curricular revision and development will occur with increasing experience, the basic model developed in the pilot phase will continue. It is this writer's hope and recommendation that the primary objectives of the permanent OCPS will remain (1) training/education and (2) evaluation. To that end, the assessment center nature of OCPS needs to be documented, evaluated, and exploited in the long term. If this is done, it will provide the greatest value and most significant potential in OCPS.

## REFERENCES


COMP: College Outcome Measures Project, Summary Report of Research and Development, 1976-1980. Iowa City: American College Testing Program, 1980.

Petho, Frank C., Ph.D., LT, MSC, USN, A Brief Description of the United States Navy and Marine Corps Aviation Selection Tests. Pensacola, FL: Naval Aerospace Medical Research Laboratory, 1980.

# OFFICER CANDIDATE PREPARATORY SCHOOL

## PURPOSES AND GOALS

To prepare candidates academically, militarily, and motivationally, and to determine their overall qualifications for admission to the Navy's Officer Candidate Schools

To establish a sound basis for success in officer candidate training and in a naval career

## OBJECTIVES

To provide the candidate with an academic program of instruction in English, mathematics, and physical science which will prepare him/her for a comprehensive program of study and training in naval science, particularly in its technical aspects

To prepare the candidate with a basic orientation to the naval service; naval customs, traditions, and values; and to Navy career options and opportunities

To provide the candidate with a basic understanding of the role and function of the officer in naval organizations and to foster the development of motivation for assuming officer responsibilities

To provide the candidate with the ability to comprehend:

- -Basic concepts of naval ship systems, including the steam cycle, propulsion systems, ship design, stability and bouyancy, and damage control
- -Basic concepts of naval weapons systems, including radar and sonar fundamentals, weapons propulsion and guidance, warheads, and weapon types
- -Basic concepts of piloting and celestial navigation, including the ability to successfully perform a comprehensive piloting exercise
- -The elements of relative motion and to enable him/her to successfully perform practical maneuvering board problems

To conduct general military training to prepare the candidate physically and militarily for admission to the officer candidate schools

To establish a military environment which will promote discipline, provide for opportunities in the exercise of military organizational responsibilities, and foster learning

To develop the study, verbal, and communication skills necessary to success in the officer candidate schools and as an officer

To provide positive interaction with officer role models

## APPENDIX B

### THE SIX MAJOR AREAS OF GENERAL EDUCATION AS DEFINED IN COMP

1. Functioning within Social Institutions. Can identify those activities and institutions which constitute the social aspects of a culture; understand the impact that social institutions have on individuals in a culture; analyze one's own and others' personal functioning within social institutions.

2. Using Science and Technology. Can identify those activities and products which constitute the scientific/technological aspects of a culture; understand the impact of such activities and products on the individuals and the physical environment in a culture; analyze the use of technological products in a culture and one's personal use of such products.

3. Using the Arts. Can identify those activities and products which constitute the artistic aspects of a culture; understand the impact that art, in its various forms, has on individuals in a culture; analyze use of works of art within a culture and one's personal use of art.

4. Communicating. Can send and receive information in a variety of modes, within a variety of settings, and for a variety of purposes.

5. Solving Problems. Can analyze a variety of problems; select or create solutions to problems; and implement solutions.

6. Clarifying Values. Can identify one's personal values and the personal values of other individuals; understand how personal values develop; analyze the implications of decisions made on the basis of personally held values.

AD P001390

Weeks, Joseph L., Air Force Human Resources Laboratory, Brooks Air Force Base, Texas. (Tues. P.M.)

## The Development and Application of Measures of Occupational Learning Difficulty

The optimal allocation of talent requires, among other considerations, the measurement of both enlistee aptitudes and job aptitude requirements. Although objective procedures are available to measure aptitudes accurately, job aptitude requirements have traditionally been determined by global judgment. The Air Force Human Resources Laboratory has developed a measurement procedure which provides an empirical basis for inferring relative aptitude requirements. It is based on occupational information collected at the task level. The measure issuing from this procedure is referred to as learning difficulty and represents the time it takes to learn to perform the occupation satisfactorily. This presentation will focus on the development of the measurement procedure, the derivation of learning difficulty measures for enlisted occupations, and potential management applications of the data.



PREVIOUS PAGE IS BLANK

# The Development and Application of Measures of Occupational Learning Difficulty

Joseph L. Weeks
Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

One of the major organizational goals of the Air Force and one of the greatest challenges for Air Force management is the optimal allocation of talent. The most talented enlistees must be assigned to the most demanding occupations. Historically, Air Force occupational classification has been accomplished by reference to measures of enlistees' aptitudes. Today, entry into enlisted occupations is largely governed by minimum aptitude requirements expressed in terms of percentile score cut-offs on the mechanical, administrative, general, or electronics aptitude index of the Armed Services Vocational Aptitude Battery (Weeks, Mullins, and Vitola, 1975). The problem which serves as the subject of this presentation concerns misalignments in the levels of minimum aptitude requirements as presently established by Air Force Regulation (Note 1, Air Force Regulation 39-1).

Originally, aptitude requirements were established on the basis of expert judgment; however, once they were available more rigorous and defensible procedures were adopted. Today aptitude requirements are established and modified on the basis of studies of the relationship between enlistees' aptitudes and their performance in Air Force technical training. Correlations between different aptitude indexes and measures of training performance are examined to identify the most appropriate aptitude index to use for stating minimum aptitude requirements. The aptitude index having the greatest relationship with training performance is considered most indicative of training success and therefore most appropriate for specifying occupational aptitude requirements. After the aptitude type is determined, the minimum level of aptitude is established by reference to training course pass/fail rates (Maginnis, Uchima, & Smith, 1975a, 1975b, 1975c). For example, high course failure rate is considered a sufficient basis for increasing the level of the associated aptitude requirement. Although this approach appears to be both reasonable and defensible, it has serious deficiencies. The major problem is that standards of successful training performance appear to vary from course to course. Nevertheless, pass-fail rates are considered comparable when used for establishing minimum aptitude requirement levels. To further aggravate the problem, unsuccessful trainees may be recycled through the training course until they are successful. This transforms a training failure into a training success and directly changes the course pass/fail rate so that it is confounded with training time. In summary, course pass/fail rate is both a non-standardized and a confounded measure of training performance and, as such, is a totally unsatisfactory basis for setting minimum aptitude requirement levels.

## Scope of Problem

The problem has both an extensive and critical impact on the Air Force personnel system. Its effect is not limited to the training arena. There are immediate effects on personnel procurement. For example, when aptitude requirements are lowered the number of eligible applicants for a given

occupation typically increases. Alternatively, when aptitude requirements are raised, the number of eligible applicants decreases. The adjustment of aptitude requirements has a direct effect on occupational selection ratios.

In addition, inappropriately established aptitude requirements can have an adverse impact on job attitudes. Locke (1976) suggests that individuals assigned to jobs that do not fully utilize their talents tend to experience boredom. On the other hand, individuals assigned to jobs that require more talent than they have tend to experience frustration. To the extent that boredom and frustration contribute to job dissatisfaction, a lack of correspondence between the demands of the job and the associated aptitude requirement may indirectly contribute to increased attrition and decreased retention. Clearly, occupational aptitude requirements are critical parameters in personnel procurement, training, and utilization. Every effort must be made to insure that they are established systematically and with reference to the most accurate, empirical information available.

## Approach

In the early 70's, the Air Force Manpower and Personnel Center requested that the Air Force Human Resources Laboratory (AFHRL) conduct research to develop a more objective means of establishing relative aptitude requirements. In response to this request, research was initiated which was based on the philosophy that the occupation, rather than the training course, was the most desirable source of information for making decisions concerning aptitude requirements. The proposed approach to setting relative aptitude requirements involved the derivation of measures of the learning difficulty associated with each enlisted occupation. These measures of occupational learning difficulty were to serve as the frame of reference for inferring relative aptitude requirements for occupations in the same aptitude area; that is, occupations having a common aptitude requirement type. Notice that the emphasis is on relative levels of aptitude requirements; the determination of aptitude requirements in absolute terms was not considered to be an attainable goal given the state of the art.

In this context, occupational learning difficulty was defined as the time it takes to learn to perform an occupation satisfactorily. This definition was adopted only after research (Mead, 1970a, 1970b; Mead and Christal, 1970; and Lecznar, 1971) indicated that work supervisors achieved a high level of agreement about learning difficulty when it was defined in terms of learning time. After this definitional problem was resolved, the next major issue involved the relationship between learning difficulty and aptitude. The importance of this relationship originates with the proposal that learning difficulty be used as the standard for inferring relative aptitude requirements. Clearly, this assumption is basic to the entire approach. Christal (1976) argues that indirect evidence in support of the relationship issues from training research. He indicates that when training time is constant, the amount of material mastered is positively related to aptitudes. Alternatively, when students are trained to some standard of performance and are allowed to progress through training at their own rate, training time is negatively related to aptitudes. Direct evidence of the relationship between time to learn occupational tasks and required task aptitude was provided by Fugill (1972). He reported correlations that range from .89 to .93 between supervisors' judgments of time to learn occupational tasks and behavioral scientists' judgments of the aptitude required to insure satisfactory performance of the tasks. Fugill (1972) concluded that relative task aptitude

was conceptually inseparable from relative task difficulty when difficulty is defined in terms of learning time. Additional evidence in support of the relationship between aptitude and learning rate is readily available in the area of eductional research. Studies by Krumboltz (1965), Cronbach and Snow (1977), Block and Anderson (1975), and Gettinger and White (1979) all lend support to the notion that aptitudes are related to learning time.

Once the general strategy was adopted, the next step was the development and application of a procedure to obtain measures of occupational learning difficulty. The procedure that was developed is both highly technical and relatively complicated. Although only a brief summary of the procedure will be provided, detailed technical discussions are available in a formal report by Burtch, Wissman, and Lipscomb (1981), and in a paper presented by Weeks and Wissman (Note 2, 1980) at the Third International Occupational Analysts Conference.

The initial step in determining the learning difficulty of an occupation is to perform an occupational analysis. The occupational analysis technique employed was the task inventory approach supplemented by the Comprehensive Occupational Data Analysis Programs (CODAP). This technique consists of first having job experts analyze occupations into distinct functional areas of interrelated activities (e.g., duties) and then having them further analyze duties into meaningful units of work recognizable to the worker (e.g., tasks). This technique is routinely employed by the Air Force Occupational Measurement Center (USAFOMC) to develop task inventories descriptive of Air Force occupations. Task inventories are then used to survey occupational incumbents to obtain ratings of the time spent on tasks they perform and to survey occupational supervisors to obtain task by task ratings of learning difficulty.

In deriving measures of occupational learning difficulty, task time spent ratings and task difficulty ratings collected by USAFOMC were two of three different types of occupational information employed. The third type of occupational information was obtained through contract with a private research firm. Private contractors also provided ratings of task learning difficulty. These additional task ratings were needed because work supervisors' task difficulty ratings were not comparable across occupations. Supervisors' task ratings provided information concerning the relative order of tasks within occupations. This stems from the fact that supervisors used simple, relative rating scales in rating task difficulty. Contractors, on the other hand, provided ratings of task learning difficulty which were based on task anchored, benchmark rating scales (Burtch, Wissman, and Lipscomb, 1981). The benchmark rating scales were designed to capture the range of learning difficulty characteristic of all tasks in all occupations in an aptitude area and, as a result, yielded ratings of task learning difficulty that were comparable across occupations. Benchmark task ratings were collected for the ultimate purpose of adjusting supervisors' task ratings so that they would be comparable across occupations.

In obtaining benchmark task difficulty ratings, an unusually rigorous set of requirements were imposed. The contractor was required to select raters who had expert knowledge of occupations in each aptitude area. Next, the raters were intensively trained in the application of the benchmark rating scales. To help in the training process, procedural guides were developed which provided detailed information concerning the steps to be followed in the actual rating process, the criteria to be used in assessing task learning

difficulty, and the interpretation of the anchor tasks which define the level of difficulty associated with each point on the benchmark rating scales.

In the application of a benchmark scale to tasks from a selected occupation, the first step was to interview instructors of training courses associated with the occupation. Interviews were conducted to clarify misunderstandings concerning task statements and to collect information concerning the equipment used in task performance. After training instructor interviews were completed, the contractor raters were divided into two teams. Each rating team visited a different operational site to interview occuational incumbents and gather additional information concerning the task assessment criteria. After becoming thoroughly familiar with both the tasks to be rated and the benchmark scale anchor tasks, each team member independently provided benchmark ratings of task learning difficulty.

The final step in the rating process was to analyze the ratings for reliability and validity. If established standards of reliability and validity were attained, benchmark task ratings were then averaged across raters and used to adjust average task ratings of learning difficulty provided by occupational supervisors. The final product was an adjusted learning difficulty rating for each task in the job inventory. Adjusted learning difficulty ratings are, therefore, based on judgments from both occupational supervisors and independent experts and can be used to compare the difficulty of tasks both within and across occupations.

## Results

Examples of the reliability of both benchmark and supervisor ratings are provided in Table 1. Using the intraclass correlation technique described by Lindquist (1953), estimates of interrater reliability were calculated. The Spearman-Brown formula was then employed to obtain estimates of reliability for 14 raters per task for the benchmark ratings and for 40 raters per task for the supervisor ratings. Table 1 presents the range and median $R_{kk}$ values for several occupations in each aptitude area. These data indicate that high agreement can be achieved by both occupational supervisors and independent experts when scaling tasks along the dimension of learning difficulty.

Table 1. Interrater Reliability of Supervisor and Benchmark Ratings of Task Learning Difficulty

| | Summary Statistics | | | |
|---|---|---|---|---|
| Aptitude Area | Benchmark Ratings | | Supervisor Ratings | |
| | Range of $R_{kk}$ | Median $R_{kk}$ Value | Range of $R_{kk}$ | Median $R_{kk}$ Value |
| Mechanical | .89 - .98 | .95 | .93 - .97 | .96 |
| Electronics | .94 - .98 | .97 | .93 - .99 | .96 |
| General/ Administrative | .93 - .98 | .96 | .93 - .98 | .96 |

Note. Reliability statistics are provided for 25 mechanical occupations, 18 electronics occupations, and 55 general/administrative occupations.

The accuracy of the benchmark rating procedure was confirmed by convergent validation (Campbell and Fiske, 1959). Correlations between relative ratings of task learning difficulty provided by occupational supervisors and benchmark ratings of task learning difficulty provided by the independent experts were derived for occupations in each aptitude area. The range of correlations and median correlation obtained in each aptitude area is presented in Table 2. These data suggest convergence of judgments of task learning difficulty by independent methods and from independent sources and provide evidence in support of the validity of the benchmark rating procedure.

Table 2. Correlations between Supervisor and Benchmark ratings of Task Learning Difficulty

| Aptitude Area | Range of r | Median r | N Occupations |
|---|---|---|---|
| Mechanical | .58 - .88 | .81 | 25 |
| Electronics | .54 - .96 | .84 | 18 |
| General/ Administrative | .54 - .95 | .77 | 55 |

Once adjusted ratings of learning difficulty were available for all tasks in the inventory, the next step was to develop a procedure for combining them to represent the learning difficulty of the entire occupation. It was decided that a weighted sum of all tasks in the inventory was the most desirable approach. Also, because tasks are performed different amounts of time by different incumbents within the same occupation, learning difficulty was derived on an incumbent by incumbent basis. In so doing, the adjusted learning difficulty rating for a given task is weighted by the percent of time spent performing that task by a given incumbent. The sum of these products for all tasks performed by the incumbent is divided by a scaling factor and is considered the index of learning difficulty for that incumbent's position. To derive a global measure of learning difficulty representative of an occupation, learning difficulty indexes can be averaged across the incumbent population or any specified incumbent sample. For the purposes of this research, estimates of occupational learning difficulty were produced by averaging across incumbents in the first term of service. This procedure was dictated by the objective of the research and the management policy of assigning occupational aptitude requirements so that they correspond to entry level positions.

To date, measures of occupational learning difficulty have been produced for over two hundred enlisted occupations or Air Force specialties (AFSCs). For example, Figure 1 presents the relative order of learning difficulty of sample occupations in the general/administrative aptitude area and the associated minimum aptitude requirement as they are listed in Air Force Regulation 39-1 (Note 1, 1981). The horizontal bar next to each occupation represents $\pm$ 1 standard deviation about the mean learning difficulty for that occupation. The vertical hashmark on each bar represents the mean. The position of the bar along the horizontal dimension indicates the relative

learning difficulty of the occupation with bars located on the right side of the figure being higher in learning difficulty than bars located on the left. Comparisons of the relative order of learning difficulty with the relative order of aptitude requirements indicate misalignments of aptitude requirements with respect to learning difficulty. For example, weather specialist is assigned a higher aptitude requirement than contracting specialist even though the learning difficulty of the weather specialist occupation is lower than that for contracting specialist. Also, note that the learning difficulty for air passenger specialist and materiel facilities specialist is the same; yet, the associated aptitude requirements are different. This example is useful in demonstrating the major limitation associated with the use of occupational learning difficulty for inferring relative aptitude requirements. Although learning difficulty provides a valuable reference point for evaluating relative aptitude requirements, it provides no basis for inferring absolute aptitude requirements. Accordingly, we can conclude that the aptitude requirement for air passenger specialist and materiel facilities specialist should be the same but we can make no legitimate inferences with respect to whether the aptitude requirement should be either 40 or 50.

In addition to revealing inaccuracies in the relative levels of aptitude requirements, Figure 1 also indicates that the variability in learning difficulty is quite different from one occupation to the next. For example, the variability of learning difficulty for weather specialist is nearly twice the variability of learning difficulty for contracting specialist. This suggests the possibility of restructuring the weather specialist occupation into two separate jobs, each having different aptitude requirements, and/or redesigning the occupation or occupational tasks so that the associated learning difficulty is less variable.

## *Conclusions*

A number of different conclusions can be based on these results. First, knowledgeable judges can reach high levels of agreement concerning the relative learning difficulty of work tasks when learning difficulty is defined in terms of learning time. This was demonstrated in situations where judgments were provided by work supervisors and in situations where judgments were provided by independent experts. Second, benchmark rating scales were designed to capture the range of learning difficulty of all tasks in an aptitude area. The data indicate that the proper application of the benchmark rating scales produced valid measures of task learning difficulty. This was demonstrated by showing substantial convergence between supervisors' judgments of task difficulty based on relative rating scales and independent experts' judgments of task difficulty based on benchmark rating scales. Third, there is some degree of inaccuracy in the occupational aptitude requirements stated in Air Force Regulation 39-1 (Note 1, 1981). This was demonstrated by showing that some occupations assigned high aptitude requirements were found to be of lower learning difficulty than other occupations assigned lower aptitude requirements. Fourth, the variability in learning difficulty of entry level positions in some occupations is so great that occupation specific investigations appear to be warranted to determine the desirability of job restructuring and/or job redesign.

In general, it appears that the talent available to the Air Force is not being allocated in the most optimal manner. This may have been a relatively minor problem in the past; the Air Force has encountered few difficulties in recruiting its share of the available talent. However, projected decreases in

the number of available applicants in primary recruiting age groups suggest a growing urgency of this problem and the need for a more refined method of assigning aptitude requirement minimums.

## Reference Notes


AF Manual 39-1. Airman Classification Manual. Washington D.C.: Department of the Air Force, April 1981.

Weeks, J. L. & Wissman, D. J. The Use of Occupational Survey Data to Develop Measures of Job Difficulty. Unpublished paper presented at the Third International Occupational Analyst Conference at Randolph AFB, TX, 21-13 May 1980.


## References


Block, J. R., & Anderson, L. W. Mastery Learning in Classroom Instruction. New York: Macmillan, 1975.

Burtch, L. D., Wissman, D. J., & Lipscomb, M. Suzanne Aptitude Requirements Based on Task Difficulty. AFHRL-TR-81-34 (In Press). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Campbell, D. T., & Fiske, D. W. Convergent and disciminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Christal, R. E. What is the value of aptitude tests? Proceedings, 18th Annual Conference of the Military Testing Association, Gulf Shores AL, October 1976, 36-51.

Cronbach, L. J., & Snow, R. E. Aptitudes and Instructional Methods. New York: Irvington Publishers, Inc., 1977.

Fugill, J. W. K. Task difficulty and task aptitude benchmark scales: A feasibility study in mechanical, electronic, administrative, and general job areas. Proceedings, of 4th Annual Conference of the Military Testing Association, Lake Geneva WI, September 1972.

Gettinger, M., & White, M. A. Which is the stronger correlate of school learning? Time to learn or measured intelligence? Journal of Educational Psychology, 1979, 71(4), 405-412.

Krumboltz, J. D. (Ed.). Learning and the educational process. Chicago: Rand McNally, 1965.

Lecznar, W. B. Three methods for estimating difficulty of job tasks. AFHRL-TR-71-30, AD-730 594. Lackland AFB, TX: Personnel Division, Air Force Human Resources Laboratory, July 1971.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

Locke, E. A. The nature and causes of job satisfaction. In Dunnette, M.D. (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Maginnis, E. B., Uchima, A., & Smith, C.E. Establishing aptitude requirements for Air Force jobs: Historical review of aptitude levels and impact on the personnel system. AFHRL-TR-75-44(I), AD-A023 250. Lackland AFB TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, October 1975. (a)

Maginnis, E. B., Uchima, A., & Smith, C. E. Establishing aptitude requirements for Air Force jobs: Some personnel system actions to offset negative impacts of aptitude changes. AFHRL-TR-75-44(II), AD-A022 206. Lackland AFB TX: Occupational and Manpower Research Divsion, Air Force Human Resources Laboratory, October 1975. (b)

Maginnis, E. B., Uchima, A., & Smith, C. E. Establishing aptitude requirements for Air Force jobs: Methodological approaches. AFHRL-TR-75-44(III), AD-A022 250. Lackland AFB TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, October 1975. (c)

Mead, D. F. Development of an equation for evaluating job difficulty. AFHRL-TR-70-42, AD-720 253. Lackland AFB TX: Personnel Division, Air Force Human Resources Laboratory, November 1970.(a)

Mead, D. F. Continuation study on development of a method for evaluating job difficulty. AFHRL-TR-70-43, AD-720 254. Lackland AFB TX: Personnel Division, Air Force Human Resources Laboratory (AFSC), November 1970.(b)

Mead, D. F. & Christal, R. E. Development of a constant standard weight equation for evaluating job difficulty. AFHRL-TR-70-44, AD-720 255. Lackland AFB TX: Personnel Division, Air Force Human Resources Laboratory, November 1970.

Weeks, J. L., Mullins, C. J., & Vitola, B. M. Airman classification batteries from 1948 to 1975: A review and evaluation. AFHRL-TR-75-78, AD-A026 470. Brooks AFB TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1975.

AD P001391

# VALIDATION OF RCMP SELECTION PROCEDURES

L. WEVRICK and C.K. HUNG
Ministry of the Solicitor General
Government of Canada

This paper describes a study of the selection procedures used by the Royal Canadian Mounted Police, with emphasis on the concurrent validation of two paper and pencil tests. The first, the Education Test, consists of five subscales: mathematics, geography and science, social studies, general knowledge, and language. The second, the Psychometric Test adapted from the Army "M", has been used by the RCMP since 1945. It consists of a verbal and a non-verbal scale. Both tests were translated into French. Other components making up a weighted selection standard were also examined. Two criteria were chosen: academic results during the six-month training program, and yearly performance appraisals. Both tests had satisfactory reliabilities. As for the validities, test scores correlated with academic training results, but not job performance. Differential reliabilities and validities were found for the English and French versions. The implications of this study for selection of policepersons are discussed.

---

The views expressed in this paper are those of the authors and not necessarily those of the Solicitor General of Canada.

## INTRODUCTION

This study consists of a validation of two paper and pencil tests, the Education Test and the Psychometric Test, which are being used by the Royal Canadian Mounted Police in the selection of recruits. Since both tests have been in use for some time (the Psychometric since 1945, the Education since 1964), it was felt that they should be evaluated. The impetus for this was partly derived from a need to have defensible instruments in case of claims of test bias, and partly from a desire to see if any improvements could be made to the selection process.

### Selection procedures

In selecting recruits, the RCMP draws from a potential candidate pool which largely consists of 18-20 year old high school graduates who are in good physical condition, and who meet certain other basic criteria such as Canadian citizenship, possession of a driver's licence, etc. Applicants who meet these basic requirements are asked to take the Education Test. If they achieve a minimum score, they return to take the Psychometric Test, and to be interviewed. The interview, which can last for up to three hours, covers major aspects of the applicant's life history such as school, family, work, and leisure. In addition to the interview, an extensive background investigation is carried out by RCMP field staff.



PREVIOUS PAGE IS BLANK

Each component of the selection process is given a weight, and points assigned to each candidate. This point system is known as the Weighted Selection Standard. In addition to the components previously described, it includes variables such as height, level of education, second language capability, and previous police experience. The largest single contributor to the point total is that portion derived from the interview.

As a result of this process, those applicants who are found to be acceptable are so informed, and placed on a waiting list where they remain for a period of up to approximately two years. Successful applicants are then called in when needed, and given a rigorous medical and a further field character investigation. Finally, they are sent to the training program.

**Training program**

All recruit training is carried out at RCMP Depot, Regina, Saskatchewan. This consists of a six month residential program, made up of classroom lectures, workshops, drill, firearms use, etc. A six-month on-the-job training period follows this initial program. After successful completion of training the recruit is posted to a regular unit.

**Test in use**

The first test, the Education Test, was designed in 1964 and later translated into French. It consists of five subscales; mathematics, geography and science, social studies, general knowledge, and language, and is administered on a timed basis - 3 hours. There are 100 items in all.

The second test, the Psychometric Test, was constructed in 1945 (Ferguson, Note 2) and also translated into French at a later date. It is based on the earlier Army "M" test, and consists of 60 verbal, and 40 non-verbal (figure analogy) items. Twenty-five minutes are allowed for the first part and 20 minutes for the second. This test was standardized on a sample of 1250 RCMP members in 1945. Only one version of each test is in use.

**Previous studies**

A number of studies (BMC, Note 1; Fraser, Note 3; Monsbraaten, Note 5) were carried out in the period 1945 to 1976 describing various aspects of the tests. These studies found reliability coefficients ranging from about .40 to about .80. All of the studies were done using serving members rather than applicants. None of the studies attempted a rigorous validation although one, the BMC study, reported a validity coefficient of 0.20 for recruit academic training. No study took into account the fact that restriction of range was present due to the tests being used as part of the selection process.

### Intent of present study

The present study was designed in order to meet the following major objectives:

(a) to determine the reliability of the two tests;
(b) to determine the validities of the tests in terms of training results and job performance;
(c) to check on the equivalence of the English and French version of the tests; and
(d) to examine other aspects of the selection process to see whether any modifications were required.

## METHOD

The fact that the tests had been used in the selection process for some time (35 years for the Psychometric Test) imposed a certain number of constraints on the study design. First, the Education Test was not administered to all candidates, but only to those that passed the preliminary screening; second, the Psychometric Test was only administered to those who scored above a certain point on the Education Test. Both of these circumstances lead to restriction of range problems at various points in the study. In addition, organizational constraints made it impossible to suspend the use of the tests in the selection process, necessitating the use of a concurrent rather than a predictive validation design.

For these and other reasons, the study was divided into two parts: (I) test reliability and (II) test validity; the data for each coming from a different subject pool.

It should be noted that the number of subjects used in each part of the study varies. This resulted from the use of as many cases as possible for each stage of the analysis.

### Part I. Reliabilities

### Subjects

Part I, the reliability study, used as subjects a random sample from all applicants tested during the years 1978 and 1979. Data were drawn from 646 applicants files which contained the original answer sheet for the Education Test and the marking sheet of the Psychometric Test. After cases containing missing data were deleted and the files separated into English and French language tests, a total of 627 cases remained. Since there were few females, these data were pooled with the male applicants after a preliminary analysis showed no differences.

**TABLE 1**
Descriptive statistics and reliabilities of Education Test (applicant data)

| | English | French |
|---|---|---|
| Sample Size | 279 | 342 |
| Mean | 63.9 | 63.1 |
| Standard Deviation | 12.0 | 10.4 |
| Odd-Even Reliability | 0.89 | 0.85 |
| Kuder-Richardson 20 | 0.88 | 0.83 |

**TABLE 2**
Descriptive statistics and reliabilities of Psychometric Test (applicant data)

| | English | French |
|---|---|---|
| Sample Size | 148 | 209 |
| Mean | | |
| Verbal scale | 40.3 | 33.1 |
| Non-verbal scale | 28.9 | 26.9 |
| Total score | 69.2 | 60.0 |
| Standard Deviation | | |
| Verbal scale | 8.1 | 8.9 |
| Non-verbal scale | 5.3 | 6.2 |
| Total score | 12.1 | 13.5 |
| Odd-Even Reliability | | |
| Total | 0.89 | 0.92 |
| Kuder-Richardson 20 | | |
| Verbal scale | 0.86 | 0.86 |
| Non-verbal scale | 0.77 | 0.81 |
| Total score | 0.89 | 0.90 |

After the tests had been re-scored, summary descriptive statistics, test score distributions, item parameters, and reliability coefficents for each language version of the two tests were calculated. Two reliability coefficients were used: Kuder-Richardson 20 and the split half (odd-even), corrected by the Spearman Brown formula. While both tests were administered on a timed basis, an examination of the item statistics showed that most people attempted all items thus lessening the effect of speeding on the obtained reliabilities.

A summary of the test statistics and reliabilities are given in Tables 1 and 2. A discussion of these data will be found in a later section.

## Part II. Validities

### Subjects

The data used in this part of the study were derived from files of regular members of the RCMP engaged in the years 1975 to 1979. A total of 3342 files were initially included, with about one-sixth of the members having French as their first language. As in the reliability study, the data from the two language groups were analyzed separately. Data from male and female members were pooled after a preliminary analysis showed no significant differences.

### Criteria

The criteria used in this study are: (1) overall academic performance as reflected by achievement tests during the six month recruit training period; and (2) on-the-job performance as derived from the average yearly performance evaluation scores (PEP) based on supervisory ratings.

Validity coefficients (Pearson product-moment correlations) were calculated for each test (English and French separately) against each of the two criteria. The coefficients were then corrected for restriction of range in the predictors (Gulliksen, Note 4) using variances from the applicant group obtained in the earlier reliability study. No correction was made for unreliability in the criterion since this error factor was unknown.

The usual procedures for cross-validation were not followed. Instead, the stability of the validity coefficients was determined by doing separate calculations for each intake group from 1975 to 1979. This was done since it was felt that any changes in the coefficients because of possible changes in candidate pool, classroom marking, or performance evaluations would be more readily detectable by this method.

**TABLE 3**
Descriptive statistics for test and criterion variables (serving members)

| | English | | | French | | |
|---|---|---|---|---|---|---|
| | Sample Size | Mean | Standard Deviation | Sample Size | Mean | Standard Deviation |
| Education Test Total (maximum = 100) | 2594 | 71.4 | 7.8 | 514 | 68.9 | 8.1 |
| Psychometric Test Total (maximum = 100) | 2785 | 70.0 | 11.0 | 558 | 64.4 | 11.0 |
| Training Score (maximum = 1000) | 2451 | 807 | 55.1 | 462 | 754 | 62.7 |
| Average Performance (maximum = 100) | 2383 | 68.1 | 6.9 | 421 | 66.1 | 7.1 |

**TABLE 4**
Validity coefficients , uncorrected, and corrected for restriction of range

| | English | | French | |
|---|---|---|---|---|
| | Training | Performance | Training | Performance |
| Uncorrected | | | | |
| Education test | 0.37 | 0.11 | 0.21 | 0.05 |
| Psychometric test | 0.41 | 0.03 | 0.31 | 0.04 |
| Corrected | | | | |
| Education test | 0.53 | 0.16 | 0.27 | 0.06 |
| Psychometric test | 0.44 | 0.04 | 0.38 | 0.04 |

Note: Validity coefficients are Pearson product-moment correlations

A summary of test and criterion statistics for the validity portion of the study is shown in Table 3.

## RESULTS AND DISCUSSION

### Reliability component

From Table 1 it can be seen that the English and French versions of the Education Test are quite similar with regard to means and standard deviations. The obtained reliabilities, ranging from .83 to .89, are also similar for the two versions. An analysis of the data from the Psychometric Test (Table 2) showed somewhat different results. The mean score on the English version was approximately 9 points higher (most of the difference being on the verbal scale) than that on the French, indicating some possible test bias or language related problem. The reliability of the psychometric test ranged between 0.89 and 0.92, slightly higher than the Education test.

Although the magnitude of the reliability coefficients leaves some room for improvement, these are sufficiently high to be used as estimates when calculating the reliabilities of the tests.

### Validation results

An examination of Table 3 shows that English members have higher means on all 4 scores (2 tests, training, and PEP) than do the French members. It should be noted, however, that while the training score means showed a substantial difference, the on-job performance of the two groups was quite similar, leading support to the presence of a language problem rather than an ability one. This is consistent with the fact that most of the academic tests taken by the French group at that time were in English.

Since the two language groups had significantly different means for all scores, validity coefficients were calculated separately for each of the groups. These are presented in Table 4, both without and with a correction for test reliability. The data here are pooled for 5 year period.

Table 5 shows the validity coefficients for each of the 5 years from 1975 to 1979. While these fluctuate from year to year, the overall pattern shows the relatively consistent predictive power of the two tests.

### Education test

An examination of the score distribution on the English and French versions of the Education test showed that the two versions could be considered as equivalent forms for most purposes. However, the obtained reliability coefficients indicate that the test scores should be interpreted with some degree of caution, and that the tests could profitably undergo revision.

**TABLE 5**
Validity Coefficients for individual years, corrected for restriction of range

| | Education Test Total/Training Score | Psychometric Test Total/Training Score | Education Test Total/Performance Score | Psychometric Test Total/Performance Score |
|---|---|---|---|---|
| | ENGLISH | | | |
| All Members | .53 | .44 | .16 | .04 |
| 1975 | .60 | .46 | .13 | .02 |
| 1976 | .64 | .49 | .19 | .10 |
| 1977 | .65 | .50 | .13 | .00 |
| 1978 | .49 | .46 | .11 | -.03 |
| 1979 | .64 | .51 | .08* | .11* |
| | FRENCH | | | |
| All Members | .27 | .38 | .06 | .04 |
| 1975 | .31 | .40 | -.05 | -.11 |
| 1976 | .55* | .45* | .35* | .16* |
| 1977 | .52 | .51 | -.11 | .07 |
| 1978 | .37 | .31 | -.02 | .14 |
| 1979 | .29* | .42* | -.03** | -.02** |

* indicates 40 to 99 cases
** indicates fewer than 40 cases

The Education Test correlated significantly with the training criterion, but failed to show any relationship with job performance. This finding is quite consistent with a number of other studies which show that the predictors of trainability and job performance tend to be different.

**Psychometric Test**

The most striking finding on the Psychometric Test was the difference between the two language versions with regard to mean scores. The spread of approximately nine points of score indicates that either the two forms are not equivalent, or that the groups being tested are dissimilar with regard to the ability being measured.

The obtained reliabilities of 0.89 and 0.90 for the English and French versions were quite satisfactory, even allowing for the fact that the test was designed at the end of WW II.

The validity of the Psychometric Test was found to be 0.44 for the English Version and 0.38 for the French version, with academic training as a criterion. As in the case of the Education Test, no correlation was found with measures of job performance, a not unexpected finding. It should be noted that the obtained validity coefficient for this test is probably an underestimate of its true value. This situation arises because a correlation of approximately 0.50 exists between the two tests. Since the Education test is used as part of the process which determines who will take the Psychometric Test, a further restriction of range problem was created which could not be corrected for.

Further analyses of the data (Wevrick & Hung, Note 6) were carried out in order to explore the relationship between other components of the selection process and the criteria. Two findings relevant to the present study are worth noting. The first was that the relationship between the criteria, was quite low. For the English group the correlation was 0.21 between training and job performance; 0.15 for the French trainees. The second finding was a general lack of correlation between other components of the weighted selection standard and the criteria. Only the two test consistently demonstrated predictive potential for one of the criteria, academic achievement in training.

**SUMMARY**

Both the Education and the Psychometric Tests are sufficiently reliable and valid to justify their continued used in the selection process as predictors of performance in the academic portion of the training. However, the mean score difference between the English and French versions of the Psychometric Test indicates that consideration should be given either to adjusting the score on the French version to bring it into line with the English or to revising the test so that both versions yield the same results. These findings emphasize the fact that an intelligence test cannot be translated into another language without changing its psychometric properties.

The failure of the tests to predict subsequent job performance points out the need for the development of separate tests or procedures for this purpose. At the same time, the criteria used to assess police performance should be examined, since the obtained lack of correlation may, in part, have been due to poor criteria.

## REFERENCE NOTES


Bureau of Management Consulting, Government of Canada. Recruit selection program Vol. 2. (Project No 5-1127), Ottawa, 1976.

Ferguson, G.A. RCMP classification test. Technical Report, DND (Army), Ottawa, 1945.

Fraser, A.W. A study of the standards and methods of selection used by the Royal Canadian Mounted Police. Unpublished thesis, University of Alberta, Department of Philosophy, 1949.

Gulliksen, H. Theory of Mental Tests. Wiley, New York, 1950.

Monsebraaten, A. Review of the Royal Canadian Mounted Police Test. Public Service Commission, Ottawa, 1968.

Wevrick, L., & Hung, C.K. R.C.M.P. Selection Study. Technical report, Ministry of the Solicitor General, Ottawa, 1980.


## ACKNOWLEDGEMENT


The authors wish to thank Dr. S.J. Piccinin of The University of Ottawa, and Insp. W.R. Spring of the R.C.M.P. for the many contributions they made to this study.

AD P001392

Witmer, Bob G., Kristiansen, D. M., US Army Research Institute for the Behavioral and Social Sciences, Fort Knox, Kentucky. (Thurs. A.M.)

## A Systematic Approach to Training Program Evaluation

A training program for transitioning tank crewmen from the M60A1 to the M1 tank was evaluted during the operational test for the M1. The training program was evaluated systematically using a series of job aids and data collection forms developed by the Army Research Institute (ARI). The evaluation method used was thorough--lesson plans, training procedures, and testing procedures were all evaluated--yet simple. The data collection forms were completed by mid-level noncommissioned officers whose only experience in evaluation consisted of a three-day workshop. Completed data collection forms were returned to the training analyst, an ARI researcher, who identified deficiencies in the training program from the information appearing on the forms. When soldiers could not pass course embedded tests, the method provided objective information about what had gone wrong during training and how to modify the training program to improve it.

# A SYSTEMATIC APPROACH TO TRAINING PROGRAM EVALUATION

Bob G. Witmer
Research Psychologist
US Army Research Institute
Fort Knox Field Unit

and

D. M. Kristiansen
Team Chief
US Army Research Institute
Fort Knox Field Unit

## Introduction

The Army, in response to new Warsaw Pact weapon system development, new technology in NATO, and searches for economies in force structure, is constantly upgrading its weapon system inventory. Within the past few years, the development of a large number of new weapons systems has been started. Many of these systems are in, or are approaching, the operational testing phase.

Attending each of these systems are training programs designed to transition soldiers to the new equipment and to train newly accessioned soldiers. Transition training programs, and to some extent programs for new soldiers, are developed concurrently with the hardware. The first real opportunity to test the effectiveness and efficiency of transition training comes during operational testing. Historically, little attention has been paid to training during operational tests. Cost and Training Effectiveness Analyses have been long on cost analysis and short on training analysis. The problem has been that a useful, easily administered, high face valid set of procedures for gathering information on training effectiveness and efficiency has not been available.

The lack of a useful set of procedures for measuring training effectiveness and efficiency has resulted in many training problems going undetected. Poor training practices used in the transition training program for a given weapons system are often incorporated in the design of other training programs for that weapons system. Once routine training begins, analysis of the effectiveness of that training is rarely considered. Without rigorous formal evaluation, poor training can be given over and over again producing large numbers of soldiers who operate or maintain weapons systems with considerably less than optimal proficiency.

To ensure that training programs for new weapon systems (or old weapon systems for that matter) are effective and efficient, a methodology was needed whereby factors that adversely affect training effectiveness and efficiency could be identified. Such a method has been developed by researchers at the Fort Knox Field Unit of the Army Research Institute (ARI) and given a field trial during Operational Test III of the M1 tank.

## Training Program Evaluation Methods and Materials

The methodology for assessing training program effectiveness is the elaboration and refinement of a concept developed by Harless Performance Guild, Inc. The concept assumes a training program should be based on a set of explicitly stated training objectives. The training objectives must state the tasks to be learned in performance terms to include the conditions under which the performance must occur and the standards that must be met. The object of the training is to arrange a series of training events that enable soldiers who receive the training to meet the training objectives. The purpose of the test is to ensure that soldiers can perform the tasks stated in the training objectives. The method does not question the adequacy of the training objectives to meet the soldier's training needs. It assumes that the objectives were established through a thorough front-end analysis. The method seeks to answer two questions: "Has the training met its own goals as stated in the training objectives?" and "Is the training as effective and efficient as possible, given the constraints under which it must operate?"

To answer these questions, the method requires that data be collected on the training and testing processes as they unfold. Preliminary information on the training and testing processes is obtained through an evaluative review of the lesson plans. The adequacy of the lesson plans is assessed by evaluating each

lesson plan against a set of preestablished criteria. The criteria basically determine if the training objectives as stated are measurable, if the test as designed is a good measure of whether or not the training objectives were met, and if the lesson plan specifies the necessary training events in sufficient detail to allow the instructor to train the soldiers to meet the training objectives. Examples of the criteria used in evaluating lesson plans are included in Table 1. The detailed procedures for evaluating lesson plans and a complete list of the criteria may be found in ARI Research Product 81-15 (Kristiansen and Witmer, Note 3).

TABLE 1. Sample Items for the Systematic Evaluation of Lesson Plans

1. Does the training objective specify what the soldier must do after having been trained?
2. Does the training objective specify the standards to which the soldier must perform?
3. Are the standards clearly spelled out so that the soldier, the instructor, and a training evaluator can tell the difference between performance at or above standard from performance that is below standard?
4. Do the test items derive directly from the training objective?
5. Are the instructions for administering the test such as to ensure standardization across instructors?
6. Does the lesson plan prescribe how demonstrations should be conducted?
7. Is practice called for in the lesson plan?
8. Does the lesson plan call for practice to take each soldier up to the training standard?

Data on the training process is also collected as the instruction is being delivered. Training observers monitor the training as it is being conducted, recording their observations and comments about the training process on specially designed training observation worksheets. Data is recorded about certain critical training events such as the explanation of unfamiliar terms and concepts, demonstrations, practice activities, and testing for task proficiency. While monitoring training, training observers also record information about the training environment on a worksheet designed for collecting such information. Data collected on the training environment includes information about the availability of training resources (i.e., equipment, materials, personnel) and other factors in the training environment (e.g., noise, temperature, lighting) that might influence training effectiveness. Typically a given block of instruction will include a test to determine if the soldiers can demonstrate task proficiency. A testing observation worksheet is used to record data regarding the conduct of the test to include information about test administration, scoring procedures, and contamination of test results. Sample items from the training observation, training environment and testing observation worksheets are listed in Table 2. Notice that the items listed are in question format so that they can be answered with a simple YES or NO response. Additional information may be obtained by asking training ovservers to record their comments when training problems are observed. A complete list of the worksheet items and a detailed explanation of how they are used may be found in the observer's job aid (Witmer, Note 4).

TABLE 2. Sample Items by Worksheet for the Structured Observation of Training

Training Observation Worksheet

1. Were soldiers told the training objectives including tasks, conditions, and standards?
2. Did the instructor tell the soldiers how the equipment worked and label the parts?
3. Could demonstrations be seen and heard by all soldiers?
4. Did all soldiers practice?
5. Did the instructor follow the lesson plan?

Training Environment Worksheet

1. Were enough instructors present to provide adequate supervision and assistance?
2. Did the training equipment work properly?
3. Did each soldier receive a copy of handouts or other materials used?
4. Were the weather conditions so uncomfortable that soldiers were distracted from the training?

Testing Observation Worksheet

1. Were soldiers tested on any tasks that were not taught?
2. Were the standards specified in the training objectives used to score test performance?
3. Did the examiner help the soldiers in any way during the test?
4. Were soldiers told what they did right and wrong on the test?

Data collected using items like those in Table 2 are useful in identifying training program deficiencies. When tests given at the end of a block of instruction produce unacceptably high failure rates, the data collected during training are used by the training analyst to determine what changes to make in the training program to reduce the performance deficiency. When test performance is much better or much worse than anticipated, the data collected during testing may identify irregularities in the administration of the test that account for the unexpected test scores. Kristiansen (Note 1) has developed a modifications job aid for assisting the analyst in making recommendations for training program changes from the data collected during training and testing. Kristiansen and Witmer (Note 2) are presently developing an additional job aid, Research Product 81-18, that will provide overall guidelines to the analyst on how to organize and conduct a training program evaluation.

## Planning the M1 Transition Training Program Evaluation

Preliminary versions of these job aids were used in conducting the evaluation of the transition training program for the M1 tank at OT-III. The transition training program was designed by the M1 NET team to transition tank crewmen and mechanics from the M60A1 tank to the new M1 tank. Three tank companies were to be transitioned in succession to the new tank. For these companies both individual and collective crew-level skills would be taught. Organizational maintenance training would also be conducted to transition track vehicle and turret mechanics from the M60A1 to the M1 tank. Because the M1 transition training program was new and unproven, there existed a need to evaluate the effectiveness of the program. ARI fulfilled that need by providing a methodology and evaluation materials for determining effectiveness of the transition training.

Considerable preparation and planning preceded the field trial of the TPE job aids and procedures. The planning process was complicated by the fact that several different organizations with different goals and different data needs were involved. Among the organizations that were to have the greatest impact in determining the ultimate shape of the M1 training program evaluation were the TRADOC Combined Arms Test Activity (TCATA), the Office of Armor Force Management and Standardization (OAFMS), the Directorate of Training Developments (DTD) at the US Army Armor Center, the TRADOC Systems Analysis Activity (TRASANA), the Operational Test and Evaluation Agency (OTEA), and the Army Research Institute. Through a series of meetings and other interactions, these agencies jointly made the plans and decisions that would determine how the training evaluation for the M1 OT-III would be conducted. Some of the important issues discussed during these meetings include the following: 1) what data would be collected; 2) how it would be collected; 3) who would collect it; and 4) how the data would be distributed to the various participating agencies. Despite some initial disagreements among the agencies involved, each of these issues was resolved prior to the beginning of the OT-III.

Basically it was decided that two types of data would be collected - training data and performance data. The training data would be collected using the procedures and worksheets (see Table 2) developed by ARI. In addition to the worksheets developed by ARI, several other data collection instruments (e.g., student questionnaire, instructor questionnaire, training aids data sheet) that were of special interest to other agencies were to be used for collecting training data during the M1 OT-III. Performance data on the tests given at the end of each block of instruction were to be recorded for each soldier on individual score sheets. Diagnostic tests designed to measure selected soldier skills on the M60A1 tank just prior to M1 transition training and the same skills on the M1 following transition training were developed by TRASANA and DTD. TRASANA also constructed special score sheets for recording these diagnostic data. Although sampling of training and performance data had been considered, it was decided to collect complete training and performance data by monitoring all training sites.

The training data would be collected by a team of data collectors organized and controlled by TCATA. The data collection team would be composed of a company team chief data collector (captain), three platoon team chief data collectors (lieutenants) and one tank data collector per tank (mid-level noncommissioned officers). Data collection responsibilities were to be shared by the platoon and tank data collectors with the company data collectors acting as supervisor for the other data collectors and reviewing completed worksheets to ensure that the data collectors had responded appropriately to each worksheet item.

Performance data would be collected by the M1 new equipment training team (NETT) instructors for both end-of-block tests and diagnostic tests. However, tank data collectors would monitor the collection of performance data and complete testing observation worksheets describing the testing process for each test given.

Training data were to be collected as the training was being conducted for each block of instruction. Data collectors would be assigned to a particular block of instruction at least 24 hours prior to observing the instruction. The data collectors would observe each part of the instruction as it was being conducted. Guided by the items listed on the worksheets, the data collectors would look to determine whether certain training events (e.g., demonstration, practice, test) occurred satisfactorily and record comments on events that were not satisfactory.

Because of the large amounts of data that were expected to be generated by the evaluation of the M1 transition training program and the diverse data requirements of the agencies involved, a system was devised for handling the data. The data handling system called for completed worksheets and score sheets to be checked for omissions and inconsistencies by the company team chief data collector before being forwarded to the field test center. At the field test center, the worksheets and score sheets were to be reproduced and copies distributed to ARI and other

participating agencies. After copying, the original data forms would be placed in a master file for future reference. The data on the worksheets and score sheets were to be analyzed by ARI as received, and any resulting recommendations for training program changes were to be forwarded by ARI to the NETT through DTD. This system was designed to allow ARI to receive the data, analyze it, and make recommendations to the NETT within 24 hours of the time the training was conducted.

Prior to the beginning of the M1 OT-III ARI obtained copies of the lesson plans for the transition training program for tank crewmen. Each lesson plan was evaluated using the procedures and criteria described in a preliminary version of ARI Research Product 81-15 (Kristiansen and Witmer, Note 3). This job aid for evaluating lesson plans seemed to work well, producing 43 pages of comments and suggestions regarding problems associated with the lesson plans. However, the NETT did not revise the lesson plans as recommended prior to the beginning of the M1 OT-III.

In preparing for the M1 OT-III evaluation, ARI conducted workshops in order to train the soldiers who were to collect the data. In the workshops, the soldiers would learn about objective observation procedures and be familiarized with the observer's job aid and the training evaluation worksheets. During the workshop, each item on training environment, training observation, and testing observation worksheets were explained and relevant examples were given where applicable. After this initial familiarization with the worksheets, the soldiers were required to use the worksheets to record observations on an actual class as a practice exercise. Following practice in using the worksheets, the soldiers discussed their observations of the class and received feedback from the workshop leader.

## Field Testing the TPE Method and Materials During the M1 OT-III

The M1 OT-III commenced at Fort Hood, Texas in September of 1980. ARI's early involvement in the operational test consisted primarily of training data collectors to use the TPE worksheets to collect training evaluation data. Aside from training data collectors, ARI's role during the M1 OT-III was limited to that of the training analyst. As the training analyst ARI analyzed the training and performance data generated during the OT-III and recommended changes to the training program based on this analysis. ARI's activities during the M1 OT-III provided an excellent opportunity for field testing the TPE method and materials. Analyzing the transition training allowed the ARI researcher to assess the adequacy of the TPE system as a method for identifying training problems. Along with teaching the workshop, performing the analysis also provided information on the ability of noncommissioned officers to use the TPE worksheets to collect the training data.

In conducting the field test of the TPE system, ARI confined its analysis activities to evaluating the effectiveness of the M1 transition training given to tank crewmen. ARI evaluated the effectiveness of this training for each of the three tank companies participating in the M1 OT-III. Data was systematically collected by task on observable training process variables detailed in the observer's job aid (Witmer, Note 4). For most of the training program, training was conducted at multiple sites, which often required as many as thirteen training observers watching the same training simultaneously, but at different sites. The bulk of the training data was collected by the tank data collectors. Contrary to what was planned, the platoon data collectors collected very little data and served primarily as supervisors for the tank data collectors. The company data collector reviewed each completed worksheet and after receiving a complete set of data for a given block of instruction, forwarded the data to the field test center for reproduction and distribution. At the field test center, the ARI analyst received copies of the Observation of Training Events Worksheet., the Training Environment Worksheet, the Observation of Test Events Worksheet and individual test performance score sheets for each block of instruction. Several days often elapsed between the time when the training was conducted and the time when ARI received the data for a given block of instruction. The delay in the receipt of the data most often was

due to inefficient reproduction of the data resulting from problems with both the copying machine and the machine operator. Frequently the copies when received were of such poor quality that it was necessary to go back to the originals to read a comment that was illegible on the copy.

For ease of analysis, the data for each block of instruction were transferred onto summary data forms. The summary data forms showed the total number of occasions on which a worksheet item was judged satisfactory or not satisfactory and included all the comments made regarding a particular block of instruction. The summary data forms allowed the training analyst to quickly identify the training problems for each block of instruction.

For analysis purposes, the analyst also summarized the data from the test performance score sheets. The score sheets recorded individual soldier performance by task for each task tested. Each soldier received either a GO indicating that the task was performed to standard or a NO-GO indicating that the soldier failed to meet the standard. The analyst reviewed the score sheets for each block of instruction to obtain the percentage of soldiers receiving NO-GO's for each task. Tasks for which 20 percent or more of the soldiers tested received a NO-GO were considered to represent performance deficiencies. Possible causes of these deficiencies were identified from the training data recorded on the Observation of Training Events, Observation of Test Events and Training Environment Worksheets. From these causes, the analyst, with the help of the modifications job aid (Kristiansen, Note 1), suggested changes in the training for eliminating the performance deficiencies.

For some blocks of instruction, review of the score sheets did not turn up any performance deficiencies. Nevertheless the observation of test events worksheet was reviewed to determine if there were any irregularities in the testing procedures. When problems were identified in the testing process that could adversely affect the validity of the test, changes to the test were suggested based on the guidance provided in the modifications job aid.

The training analyst summarized his findings and recommendations for each block of instruction in a memorandum to DTD. The memorandum identified the tasks for which performance deficiencies were found and specified changes in the training program designed to correct the deficiencies. The memorandum also identified problems with testing procedures when they existed and suggested changes as appropriate.

Because of a number of problems unique to the M1 OT-III at Fort Hood, ARI memorandums were not forwarded to the NETT until the first company had completed much of its training. Day-to-day changes were not made as planned.

Even through the NETT received the ARI's memorandums following transition training for a given company, there was still time to modify the transition training before training the next company. Upon completion of the transition training for the first company, DTD issued a memorandum instructing the NETT to modify the transition training given to the next two companies so as to eliminate some of the recurring training problems identified by ARI for the first company. Some of the modifications suggested by ARI appeared to have been made prior to conducting transition training for the second and third companies. Evidence that some changes were made comes from the training and performance data collected for the second and third companies. However, changes were not extensive and were not always documented in the lesson plans. Failure to document changes in the transition training from one company to the next precluded establishing direct

relationships between the changes made in the transition training for each block of instruction and performance on the tests given following that block of instruction. However, considerable information was gathered that indirectly demonstrated the usefulness of the TPE methodology in evaluating the effectiveness of training programs.

The TPE methodology provided the largest pool of objective training data that was available during the M1 OT-III. The usefulness of this data is attested to by the fact that virtually every organization involved in evaluating the effectiveness of the transition training made use of this data. The primary users were OAFMS, TACTA and ARI. OAFMS used the data for certifying the readiness of the OT players as M1 qualified crewmen and mechanics. OAFMS also used the data in certifying the effectiveness of the M1 transition training program. TCATA kept detailed records of all the TPE training data and used the data and the memorandums generated by ARI as input for their own independent training effectiveness analysis. DTD and TRASANA also used the data in conducting a Cost and Training Effectiveness Analysis (CTEA) of the M1 transition training program, although TRASANA relied more heavily on pre- and post-diagnostic test data. All of the organizations involved in evaluating the effectiveness of M1 transition training seemed to find the data useful.

Among the changes made in the transition training for the second and third companies were the addition of demonstrations to some lessons and closer adherence to the lesson plans in conducting the instruction. Evidence that these changes were made came from comments recorded by the data collectors on the worksheets and were verified by informal contacts with the NETT and OAFMS. The number of comments indicating that tasks were not demonstrated or that lesson plans were not followed dropped sharply from the first company to the second and remained at the lower level for the third company. This indicates that the TPE methodology was sensitive to changes made in the training process.

Performance data indicating soldier performance on the end-of-block tests shows improvement from one company to the next. This may be interpreted as indicating increased training effectiveness. For example the proportion of tasks having 100 percent first-time GO rates increased from 24 percent for the first company, to 34 percent and 53 percent for the second and third companies, respectively. While such increases may be due in part to other factors (e.g., reduction of standards for some tasks and the elimination of some task requirements), the trend toward higher first-time GO rates constitutes indirect evidence that the changes made in the training from one company to the next increased training effectiveness and thus supports the usefulness of the TPE methodology responsible for these changes.

Although the data provided by TPE methodology was considered useful by the organizations employing it and the recommendations suggested by ARI based on the data seemed to increase training effectiveness, ARI encountered some problems in using the TPE methodology for evaluating the M1 transition training. The biggest problem came in getting the NCO's who were collecting the data to use the TPE worksheets the way they were designed to be used. The worksheets listed specific items which required the data collectors to observe the training to determine if it met specific criteria described in the observer's job aid and discussed in the TPE workshop. When these criteria were not met, the data collectors were supposed to record a comment detailing what they observed. However, many of the data collectors treated the worksheets as a simple checklist, responding to the items subjectively based on their general impressions of the items rather than the objective criteria specified in the observer's job aid. Furthermore the number of comments recorded on the worksheets were far fewer than were expected based on independent observations of the training by the ARI analyst and others.

Any one of several factors might account for the failure of some data collectors to use the worksheets as the worksheets were designed to be used.

The NCO's collecting the data were tankers who were perhaps more interested in the new M1 tank than in making objective observations about the training and testing processes. For these tankers, the task of collecting training and testing process data may have been considered menial or meaningless, especially in contrast to learning how to operate the new tank. The data collectors were required to complete several other forms in addition to the TPE worksheets; this additional workload may have diminished the amount of effort devoted to the TPE worksheets. On several occasions, the TPE analyst observed that data collectors were standing around in groups of four or five while the training that they should be monitoring was being conducted on the tanks. In these instances, the data collection team supervisors were not supervising and the data collectors were not collecting data. Such incidents and informal discussion with the data collectors indicated that many of the data collectors were not approaching the data collection task seriously. This apparent lack of motivation in the data collectors may have resulted in part from the factors listed above, but may also be due in part to the adverse working conditions. For example, data collectors were often required to collect data all day in wet or cold and windy weather. Often they were called upon to collect data right through (and long after) normal meal hours. Many of the "creative comforts" provided to the participating units were not given to the observers and they were treated as unnecessary by the NETT and the units.

The problems encountered in using the TPE methodology during the M1 OT-III are instructive. They indicate that the TPE analyst must select the data collectors more carefully and personally oversee and control the data collection effort. The number of forms to be completed by any one data collector and the number of hours spent completing these forms must be limited to a reasonable level. In order to ensure the timely flow of data from the data collector to the analyst and from the analyst to the persons responsible for instituting changes in the training program, a direct line of communications should be established from the analyst in both directions. The originals of completed worksheets and score sheets should go directly from each data collector to the analyst. The analyst would then analyze the data and results of this analysis would be made available to other organizations. If other persons or organizations needed the raw data, they would have to obtain copies through the TPE analyst. The recommendations for changes in the training program would then be forwarded directly to a member of the team responsible for making changes in the training program. This person would check into the possibility of making each of the changes suggested by the analyst and would inform the analyst which changes were made and which were not.

In order to reduce the tendency of data collectors to respond subjectively to the items on the TPE worksheets, the worksheets and the items used during the M1 OT-III were modified in the new version of the observer's job aid. The old worksheets required data collectors to make judgments (i.e., OK or Not OK) regarding their observations according to criteria listed for each item. On the revised worksheets, items are worded more precisely and data collectors only need respond objectively with a "YES" or "NO" indicating whether or not the event occurred as described in the item.

In conclusion, the TPE system has proven to be a useful method for evaluating training programs such as the transition training program for the M1 tank. The method provides objective training and testing process data not heretofore available. The TPE system has been well documented in a series of easy-to-use job aids (Witmer, Note 4; Kristiansen, Note 1; Kristiansen and Witmer, Note 2; and Kristiansen and Witmer, Note 3). The quality of these job aids has been much improved through the lessons learned from the M1 OT-III experience, and provide detailed information on how to collect and analyze training data in order to evaluate the effectiveness and efficiency of training programs.

## Reference Notes


1. Kristiansen, D. M. A job aid for modifying ineffective or inefficient training programs (ARI Research Product 81-17 Fort Knox, KY. Fort Knox Field Unit, US Army Research Institute for the Behavioral and Social Sciences, Manuscript submitted for publication, September 1981.

2. Kristiansen, D. M. & Witmer, B. G. Guidelines for conducting a training program evaluation (ARI Research Product 81-18) Fort Knox, KY. Fort Knox Field Unit, US Army Research Institute for the Behavioral and Social Sciences, Manuscript in preparation, October 1981.

3. Kristiansen, D. M. & Witmer, B. G. A job aid for the systematic evaluation of lesson plans (ARI Research Product 81-15) Fort Knox, KY. Fort Knox Field Unit, US Army Research Institute for the Behavioral and Social Scieces, Manuscript in preparation, October 1981.

4. Witmer, B. G. A job aid for the structured observation of training (ARI Research Product 81-16) Fort Knox, KY. Fort Knox Field Unit, US Army Research Institute for the Behavioral and Social Sciences, Manuscript submitted for publication, September 1981.

AD P001393

Yard, Gilbert F., Royal Canadian Mounted Police, Ottawa, Ontario, Canada. (Wed. A.M.)

## Training Evaluation - A Pragmatic Overview

Evaluation is a common term, however, familiarity does not breed understanding. There are about as many meanings to the term "evaluation," as there are evaluation tasks to be performed and personnel charged with evaluation responsibilities. To the manager "evaluation" may be tied to personnel efficiency or answering the question: "Does it work?" To government "evaluation" most often translates into terms of cost/efficiency (inefficiency) or accountability. To the individual line-worker "evaluation" usually means a critique or judgment of personal worth conducted by someone in authority. All of these definitions are, of course, partially correct. Training evaluation could be carried out by session or course, however, in large organizations in particular, the more global concerns of identifying the fragmented components that contribute to the evaluation process and organizing such components within a formal systematic approach of program evaluation is often ignored. This paper reports the progress of the Royal Canadian Mounted Police in addressing systematic program evaluation as it pertains to the Force's training program. At issue are the objectives of such evaluation; the melding of methodological and pragmatic concerns; and the development of an approach that will meet the needs of both internal efficiency and external scrutiny.

# TRAINING EVALUATION - A PRAGMATIC OVERVIEW

GILBERT F. YARD

ROYAL CANADIAN MOUNTED POLICE
TRAINING AND DEVELOPMENT BRANCH
PSYCHOLOGIST UNIT

## Introduction

Evaluation is a common term, however, familiarity does not breed understanding. There are about as many meanings to the term "evaluation", as there are evaluation tasks to be performed and personnel charged with evaluation responsibilities. To the manager "evaluation" may be tied to personnel efficiency or answering the question: "Does it work?". To Government "evaluation" most often translates into terms of cost/efficiency (inefficiency) or accountability. To the individual line-worker "evaluation" usually means a critique or judgment of personal worth conducted by someone in authority. All of these definitions are, of course, partially correct. Training evaluation could be carried out by session or course, however, in large organizations in particular, the more global concerns of identifying the fragmented components that contribute to the evaluation process and organizing such components within a formal systematic approach of program evaluation is often ignored. This paper reports the progress of the Royal Canadian Mounted Police in addressing a systematic program evaluation as it pertains to the Force's training program. At issue are the objectives of such evaluation; the melding of methodological and pragmatic concerns; and the development of an approach that will meet the needs of both internal efficiency and external scrutiny.

## Background

The communication of complex concepts is often difficult but as such difficulty is expected it is usually compensated for by devoting additional time and effort in both explanation and interpretation. This is not the case, however, when common terminology is used to describe an area where the expectations are that the majority will understand or at least have an appreciation of the subject matter. In such cases, there can be a diversity of interpretations and the ingredients are present for widespread misuse, abuse and misunderstanding.

"Evaluation" is such a concept. We have all evaluated and ourselves been subjected to evaluation. Depending on our individual orientation and the context in which we are thinking, however, the term "evaluation" can take on a number of meanings.

## The Problem

Questions of accountability, efficiency and a commitment to delivering the "best" training product attainable gave rise to a Training Evaluation Unit within the Royal Canadian Mounted Police in 1979. The pressure to evaluate training came almost equally from internal and external sources. Internally, the need was for a systematic approach which would formally allow for monitoring and quality control. Externally, through Government, the pressure was to accountability and the best value for the dollar spent. While these pressures are, of course, relatively universal, the response is worthy of examination.

By far the greatest volume of evaluation literature deals with the individual case and transposed to the training milieu, we find that there has been much written on "course" evaluation strategies and methods. The R.C.M.P. Training Evaluation Unit, however, was given the responsibility for keeping track of several hundred different training courses offered to R.C.M.P. personnel by both internal and external sources. The question was: "How do you apply relatively scarce resources in the most meaningful way?". The mandate, with those same scarce resources, was to become accountable to Government and to assess the commonly held belief that our training product was the highest quality available or alternatively to institute a positive change in that direction.

## The R.C.M.P. "Systems" Approach To Training

Trainers within the R.C.M.P. have long prided themselves on their systematic and thorough approach to training. They are committed to providing training on an identified need basis only and work exhaustively to assess the value of each request for training and then to develop clear cut, behavioural "student-performance objectives". The Force relies heavily on work by Robert Mager (1967, 1970, 1972, 1973, 1975) in developing this behaviourally-based, criterion-referenced program. The focus on training to need and training by objectives is the anchor which provides a logical strategy for both student assessment and program evaluation.

## The Solution

Many of the components of an evaluation system existed within the R.C.M.P. training program prior to 1979. It has taken until this year, however, to delineate a formal evaluation system. Some of the existing components were identified as:

- needs analysis - used to determine if training is the solution to specific problems.

- student performance objectives - used to base student testing and later to assess behavioural change on the job.

- informal feedback - used as an indicator of where attention is needed or objectives met.

- formal course critiques - a long used catalyst for positive change.

The need was to tie these components together and provide a formal structure which would answer the need for accountability as well as provide the most efficient use of the relatively scarce human resources available for more traditional evaluation research.

The identification of these "existing components" was the first step of a systems approach to evaluation. To add structure and formalize the feedback component an interlinked series of forms and policy was developed. The result of this exercise was the "end-of-course report" which provides a measure of the following factors on a scale of 1 to 7:

- need for the course
- timing of the course
- prior knowledge
- course level
- success of course

Opinions are provided on these issues by both the students and the course co-ordinators. Additionally, the co-ordinator is asked to comment on:

- whether the selection criteria for the course had been met.

- whether temporary changes were necessary to the course standards or objectives.

- whether permanent changes are recommended in the course content.

This information is gathered in a computerizable format and provides a gross but formal measure on every course used by the R.C.M.P.

## Evaluation Research

Evaluation research is instituted on a need basis in accordance with the following indicators:

- informal feedback
- formal feedback (end-of-course report)
- as directed by the Officer i/c Training and Development Branch.

By adopting this approach, the normal maintenance of our programs is encouraged along with a systematic review and access to further research. In cases where further research is called for, the Force once again institutes a systematic approach.

Where in-depth evaluation research is called for, questionnaires are developed based on four levels of questioning:

(a) the general level covers the questions of need, timing, prior knowledge, course level and success noted in the end-of-course report. These questions are once again responded to by indicating a position on a seven point scale. The repetition of these questions on all evaluation research projects provides a relative measure between evaluation research projects.

(b) Recognizing the frustration experienced by being forced to answer to a seven point scale, the second level allows for narrative expansion of the questions covered in level 1.

(c) The third level consists of questions dealing with course specific information and is directly based on the student performance objectives of the course under review.

(d) The fourth and final level deals with specific questions on matters which we may suspect a priori are cause for concern.

This "level" approach to evaluation research is directed at students, superiors, peers or subordinates as circumstances may dictate, although the student is the most usual target. It should also be pointed out that the use of telephone or direct interview serve to supplement what might otherwise be a purely "mail out" exercise.

Post-course testing is consistently applied to guarantee the training objectives have been met. The use of pre-course testing, an often valuable evaluation tool, is used to a much lesser degree and then most often at the course development stage. This is not to say that a pre/post course test sequence would not be used where a specific need was identified but the nomination of course candidates on a need basis and the time and rigor necessary for an adequate pre/post test program make usage infrequent.

## Conclusion

The evaluation program as described has allowed the R.C.M.P. to provide Government with a strategy that meets the requirement of accountability. The policy necessary to make the system operational is only just being put into place, however, early returns are most promising. There is little doubt that "fine tuning" will be necessary but the main components appear to be in place. In restrospect, the system appears to be straightforward, however, the 14 months necessary as development time would indicate that there are many areas where consensus must be obtained before operational status can be achieved.

## References

MAGER, R.F., and BEACH, K.M. - Developing Vocational Instruction, Belmont, Calif., Feardon Publishers, 1967.

MAGER, R.F., and PIPE, P. - Analyzing Performance Problems, Belmont, Calif., Feardon Publishers, 1970.

MAGER, R.F. - Goal Analysis, Belmont, Calif., Feardon Publishers, 1972.

MAGER, R.F. - Measuring Instructional Intent, Belmont, Calif., Feardon Publishers, 1973.

MAGER, R.F. - Preparing Instructional Objectives, Belmont, Calif., Feardon Publishers, 1975.

# PANEL DISCUSSIONS



PREVIOUS PAGE
IS BLANK

# Psychometric Considerations for Adaptive Testing Systems: Test Development and Calibration

Bejar, Isaac I., (Chair); Dorans, Neil J., Dwyer, Carol A.,
Educational Testing Service, Princeton, New Jersey.

Much of the existing literature on adaptive computer-assisted testing has been based on small-scale samples. The large-scale implementation of that methodology presents problems that have seldom been considered and are probably not well understood. This panel discussion will focus on issues on test development and calibration from the prespective of Item Response Theory as applied to adaptive testing. The four presentations will discuss the creation and maintenance of an item pool, the estimation of item parameters, the linking of item parameters, and a discussion of these issues from a military testing perspective.

PREVIOUS PAGE IS BLANK

AD P001394

# Why and How to Insure that Items in the Same Pool Are Appropriately Credentialed: Data Collection Strategies and Analytical Methods for Item Linking[1]

Neil J. Dorans
Educational Testing Service

[1]Presented at the 23rd Annual Conference of the Military Testing Association, Alexandria, Virginia, October 27, 1981.

# Why and How to Insure that Items in the Same Pool Are Appropriately Credentialed: Data Collection Strategies and Analytical Methods for Item Linking[1]

After designing an adaptive test and obtaining item parameters via some calibration procedure, it is necessary to place item parameters on a common metric prior to testing in an adaptive mode. In the adaptive testing context, several subpopulations of single individuals are involved; hence, common metric considerations are critical. Several data collection strategies and analytical techniques for placing parameters on scale will be surveyed.

Prior to this survey, a contrived illustrative example will be used to persuade you that a common metric is essential for valid comparisons in any investigation of individual differences and similarities. Although it borders on the absurd, this fabrication clearly demonstrates the need for a common metric in interindividual comparisons.

## A Contrived Illustrative Example

In the spirit of peaceful coexistence, an American scientist and a Soviet scientist were told to collaborate in an investigation of the effectiveness of a new weight reduction drug. Although the project was hampered by communication problems due in part to divergent training, external pressures mounted, pushing it along to the experimentation stage. An experiment was conducted. The American administered a daily noncarcinogenic dose of the drug to a group of rats. The drug was mixed in with the rats' daily breakfast for a one-month period. Meanwhile, on the other side of the world, the Soviet scientist administered the same daily breakfasts without the drug to a control group of rats from the same litter. At the end of the month, the rats on both sides of the world were weighed. Appropriate statistical analyses were performed to assess the effectiveness of the drug.

The mean difference between the experimental and control groups was remarkably large. In addition, the variance of weight in the experimental group was much smaller than the variance in the control group. Apparently, the drug not only reduced weight but also reduced variability in weight. (The Americans were most interested in the former result, while the Soviets were excited by the latter.) More importantly, the discovery of the new miracle drug was the result of a joint Soviet-American effort. Consequently, the result was immediately publicized, and the press played a major role in arousing public expectations throughout the world.

In the meantime, replications of the original investigation were hindered by bureaucratic red tape (a double dose in this case). Eventually,

approval for a limited number of replications was given. By the time the results of these replications became public, however, the new drug had firmly established an identity as a miracle drug. As a consequence, the results of these new experiments shocked the world: The miracle drug had no effect on weight reduction. Disbelief, the immediate reaction, developed into hostile rejection of these new results. Accusations filled the air and additional studies were commissioned, some by the pharmaceutical companies that had acquired rights to manufacture the drug. Results from these new studies varied. The preponderance of evidence, however, indicated that the so-called miracle drug was useless.

As tensions mounted, each of the original investigators accused the other of fraud and deception, further straining already tenuous diplomatic ties between their countries. In an effort to avert international disaster, a panel of influential nonexperts and a few experts was commissioned by the United Nations to effect a resolution of the controversy. The two scientists and their experiment were the foci of this investigation. The credentials of both scientists were questioned and established. Incidents from their private lives that might have had a bearing on the controversy were scrutinized.

Eventually, the investigation focused on the actual experiment itself. The experimental design was reviewed and deemed adequate. Then, the statistical analyses that were performed came under scrutiny. Although minor flaws were detected, the panel concluded that the analyses met professional standards. (The researchers went well beyond the magical .05 level of significance.)

Suddenly, with the answer to a simple question, the controversy was resolved. When questioned about the method of measurement he used, the Soviet revealed that he had measured weight in grams. In another room, the American admitted to measuring weight in pounds. As a result of a difference in metric, the control group weighed much more than the experimental group, and also exhibited more variability. Had either pounds or grams been used by both scientists, the drug would have been deemed useless immediately, and the heated controversy would never have evolved.

Most psychologists immediately recognize the absurdity of not using a common metric in this contrived example. Unfortunately, not enough of these same psychologists readily see the example's implications for their own research. Perhaps they have misinterpreted the dictum that "the choice of metric for psychological variables is often arbitrary" to mean that it is permissible to make comparisons between groups on scores that are expressed in different metrics. Choice of metric is arbitrary. However, the contrived example clearly illustrates that, while the choice of metric may be arbitrary, it is essential to ensure that the arbitrarily chosen metric is maintained across any groups that are the objects of intergroup comparisons. In the contrived example, the choice of grams or

pounds was arbitrary. Arbitrary or not, both researchers needed to use a common metric, either grams or pounds.

## Data Collection Strategies

Having established the need for a common metric, I will now discuss how to place item parameters on the same metric. There are two aspects to item linking: data collection strategies and analytical methods.

There are three basic designs for data collection: single-group, equivalent-group, and anchor-test (Lord, 1975).

### Single-group Method

In the single-group method, all items to be placed on the same scale are administered to the same examinees. This method is limited by the number of items examinees can be administered before fatigue and practice effects begin to have an appreciable impact on item performance.

### Equivalent-group Method

In the equivalent-groups method, two or more equivalent groups of examinees are set up by random sampling or by other methods. Each group of examinees is administered a unique set of items. This method is less susceptible to fatigue and practice effects and allows for the administration of more items than the single-group method. However, the inevitable but small differences among the equivalent groups introduces a small but unknown bias into the linking results.

### Anchor-test Method

In the anchor-test method, different groups of examinees are administered two or more sets of items. Each group of examinees is administered a set of items that is unique to that group only, and a set or sets of items that is also administered to one or more other groups of examinees. The set of items that is administered to two or more groups of examinees, referred to as the "anchor test," is crucial to the linking process because it measures any differences between the groups of examinees and can be used to reduce bias due to group differences. It is essential to use the anchor-test method whenever the groups are not equivalent, or when only small samples of randomly chosen groups are available. This method is particularly useful for linking long chains of test forms.

## The Three-Parameter Logistic Item Response Theory Model

A mathematical model is needed to guide the item linking procedure. Item response theory models hold the greatest promise for computerized adaptive testing systems (Lord, 1980; Urry, 1977). The three-parameter logistic model has been the most widely accepted unidimensional model for use with binary-scored multiple-choice items.

The three-parameter logistic model specifies that the probability of success on an item is a function of ability level $\theta_i$ and three characteristics of the item. For an individual with ability $\theta_i$, the three-parameter logistic model expresses probability of success on item g as

$$P_g(\theta_i) = c_g + (1 - c_g)\left[1 + \exp(-1.7\, a_g(\theta_i - b_g))\right]^{-1} \qquad (1)$$

where $a_g$ is the item's discriminatory power, $b_g$ is the item's difficulty index and $c_g$ is referred to as a guessing parameter. A sample ICC relating $p_g(\theta_i)$ to $\theta_i$ is depicted in Figure 1.

In Figure 1, note that the probability of a correct response is a monotonically increasing function of ability level. As ability increases, success on the item is more likely. The probability of a correct response approaches unity at very high levels of ability. At very low levels of ability, the probability of a correct response approaches $c_g$, the guessing parameter.

In mathematical jargon, $c_g$ is the lower asymptote of the curve (ICC), $b_g$ is the value on the $\theta^g$ continuum corresponding to the point of inflection in the ICC, and $a_g$ is proportional to the slope of the ICC at its point of inflection. As a value expressed on the $\theta$ continuum, $b_g$ is often referred to as the item location parameter. Note that as the location of the curve shifts to the right along the $\theta$ continuum, $b_g$ increases, and greater amounts of ability are needed to maintain the same probability of successful performance on the item. Hence, $b_g$ is known as the item difficulty parameter.

The parameter $a_g$ indicates how well the item discriminates between levels of ability that are slightly above and slightly below $\theta = b_g$. As $a_g$ gets larger, the slope of the curve becomes steeper and the discrimination between ability levels close to $b_g$ increases. In contrast, flatter curves have lower values of $a_g$, reflecting coarser measurement over a broader range of ability. For a more complete description of these three item parameters consult Lord and (1980) and Hambleton and Cook (1977).

## Analytical Methods for Item Linking

After collecting data, an analytical method for item linking is employed to place item parameters on a common scale. Several methods exist. The simplest methods are applied to data collected under the single-group design or the equivalent-groups design. Data collected under an anchor-test design requires more complicated analytical methods.

### Single-Group Data

The analytical method used on single-group data is actually no method at all. Because all items are administered to the same group of individuals, all data can be submitted to a single calibration run. As a consequence, all item and ability parameters are on the same scale, the metric defined by the single calibration run.

### Equivalent-Group Data

In the equivalent-group method, two or more equivalent groups of examinees are administered unique sets of items. Each item is administered to only one group of examinees. Actual equivalence of groups is necessary for this method of data collection. If the groups actually are equivalent in ability, then item parameters can be placed on scale by setting the means and standard deviations of ability scores in the different equivalent groups equal to some fixed arbitrary values. With equivalent-group data, setting the first and second moments of the ability score distributions equal to some fixed values in all equivalent groups of examinees fixes the ability metric and places all item parameters on a common scale.

### Anchor-Test Data

Several methods for item linking exist for data collected under the anchor-test design. Frederic Lord and his ETS colleagues, Martha Stocking and Marilyn Wingersky, are actively involved in item linking methodology for anchor-test data. In this section, some of their developments in this area will be briefly described.

Partial pre-calibration or "fixed-b" method. The first method described in this section differs from the subsequent methods in that only a single set of parameter estimates is obtained for the anchor-test items. In the partial precalibration method, parameter estimates for the anchor-test or common items are obtained from a calibration run on some pre-existing data set. Then those items are included in a calibration run with the items that need to be placed on scale. In this calibration

run, the anchor-test items are used to define the scale. Estimates of difficulty, discrimination, and the lower asymptote of the item characteristic curve are not obtained for these anchor-test items. Instead, their parameters are fixed at the values obtained at the earlier pre-calibration run. Hence, in this method, which is sometimes referred to as the "fixed-b" method ("b" is the symbol employed for item difficulty in item responce theory models), the pre-existing parameter estimates for the anchor-test items are used to fix the scale for the new items.

Standard mean and sigma method. The next three methods are variants of a traditional method of placing item parameters on the same scale, setting the mean and standard deviation of item difficulties for a set of common items equal to some fixed values; e.g., the mean and standard deviation of the difficulty parameters obtained for those same items in some other data set. This method of placing item parameters on scale employs two sets of item parameter estimates for each item in the anchor-test. Typically, the mean and standard deviation of one set of item difficulty parameter estimates is chosen as fixed, and a linear transformation is computed that converts the second set of parameter estimates to the scale of the first by setting the mean and standard deviation of the second set of estimates equal to the mean and standard deviation of the first set of estimates. Then, this linear transformation is applied to all the other items from the second calibration set to convert their parameters to the scale of the first calibration set.

The mean and sigma method is simple to employ. The means and standard deviations of the two sets of difficulty parameter estimates are easy to compute. Once obtained, the equation for converting the difficulty parameters from the second calibration set to the scale of the first set is

$$\tilde{b}_2 = Ab_2 + B \qquad (2)$$

where,

$b_2 \equiv$ item difficulty parameter estimate for an item from the second calibration set,

$b_2 \equiv$ transformed item difficulty parameter estimate on the scale of the first calibration set.

In Equation (2), the slope

$$A = SD(b_{1c})/SD(b_{2c}), \tag{3}$$

and the intercept

$$B = \bar{b}_{1c} - A\bar{b}_{2c}, \tag{4}$$

are defined in terms of the means and standard deviations of the difficulty parameter estimates of the common items in the first ($b_{1c}$) and second ($b_{2c}$) calibration sets respectively; i.e., $\bar{b}_{1c}$, $SD(b_{1c})$ and $\bar{b}_{2c}$, $SD(b_{2c})$. Equation (2) is applied to all item difficulty estimates in the second calibration set to place them on scale. In addition, the item discrimination parameter estimates of these same items are placed on scale by multiplying their original discrimination estimates ($a_2$) by the inverse of the slope A defined in equation (3),

$$\tilde{a}_2 = a_2/A. \tag{5}$$

The occular mean and sigma method. The only statistics required for the mean and sigma method of item linking are means and standard deviations. The mean and particularly the standard deviation are sensitive to the existence of outliers. Outliers are data points that are inconsistent with the pattern of data suggested by the preponderance of data. For example, suppose a difficult item on an anchor-test becomes easy because of a security leak; i.e., somehow the key to the item becomes disclosed. Also, suppose that the other items on the anchor-test remain secure. If we plotted new difficulty estimates versus old difficulty estimates for these anchor-test items, most points would lie on a straight line. To convert the new estimates to scale of the old estimates, we fit a line to the swarm of points and obtain a slope and intercept. One standard method of obtaining this conversion line is the mean and sigma method. In this case, however, because of the insecure anchor-test item, which stands out in our plot of old and new item difficulties, the mean and sigma method would not produce the best conversion line. A better line would be obtained from using the mean and sigma method on the reduced data set obtained by excluding the "outlier" item from the analysis. This use of the mean and sigma method on a data set from which outliers have been eliminated is referred to as the "occular" mean and

sigma method because the data is scanned visually for outliers before application of the mean and sigma methodology.

Transforming b's using standard errors. While the occular mean and sigma method eliminates outliers, it does it in a subjective way. Two researchers examining the same bivariate plot of item difficulties might disagree about which points are outliers. The Lord-Stocking method of transforming b's using standard errors (Petersen, Cook, and Stocking, 1981) is a relatively complicated linking method that attempts to achieve the goal of the occular method without its subjective process.

The Lord-Stocking linking procedure produces robust estimates of location (mean) and scale (standard deviation) of each distribution of item difficulties and an equation based on these robust estimates. This equation is used to convert the parameter estimates of a set of items from one calibration to the base metric of another calibration. We start with two sets of item difficulty estimates, one from each calibration. Each difficulty estimate is weighted by the reciprocal of its squared standard error of estimate; for each item the larger of its two standard errors of estimate (from the two estimates of item parameters) is used. Then the means and standard deviations of these weighted item difficulty estimates are computed and used to obtain the conversion line that converts item parameters from one metric to the other base metric. At this point the process becomes iterative. The perpendicular distances of the item difficulty points from this conversion line are computed, and then biweights (Mosteller and Tukey, 1977) for these distances are obtained. These biweights are then applied to the weighed difficulties and a new conversion line is produced. The distance, biweight, reweighting, and new conversion line cycle is repeated until the maximum change in perpendicular distance is less than some criterion. The last conversion line produced by this iterative process is then used to place the items from the second calibration onto the base metric of the first calibration.

## Summary

It is essential that all item parameters be on a common metric prior to testing in an adaptive mode. The contrived example dealing with the miracle weight reduction drug illustrated the importance of metric considerations. Three basic designs for data collection were discussed: single-group, equivalent-group, and anchor-test designs. The three-parameter logistic item response theory model was introduced as the model for guiding the item linking process. Finally, several analytical procedures for item linking were described briefly.

## References


Hambleton, R., and Cook, L. Latent trait models and their uses in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Lord, F. A survey of equating methods based on item characteristic curve theory. Research Bulletin 75-13. Princeton, New Jersey: Educational Testing Service, 1975.

Lord, F. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

Mosteller, F., and Tukey, J. Data analysis and regression. Reading, Massachusetts: Addison-Wesley Publishing Company 1977.

Petersen, N., Cook, L., and Stocking, M. IRT versus conventional equating methods: A comparative study of scale stability. Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 14, 1981.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

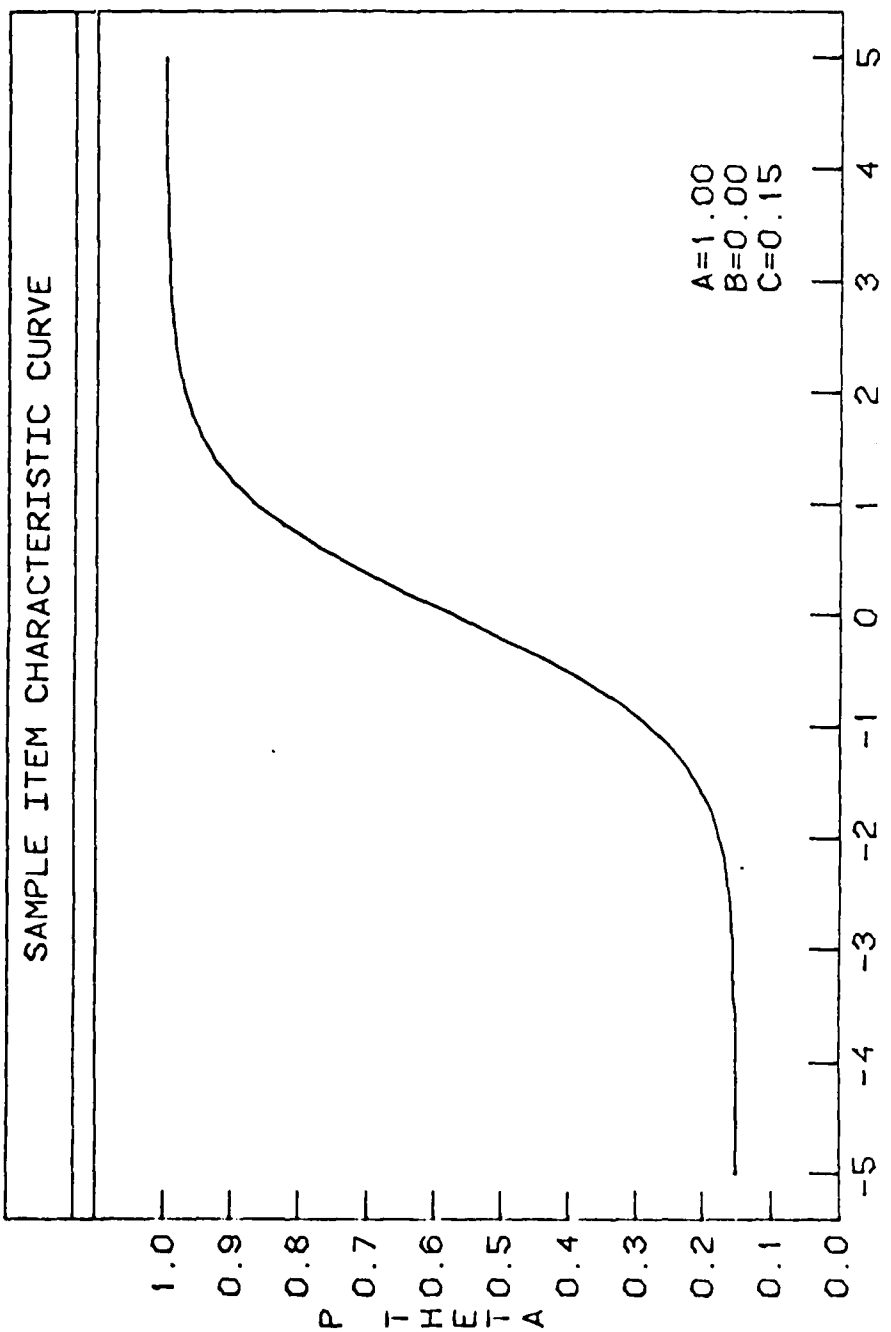

SAMPLE ITEM CHARACTERISTIC CURVE
A=1.00
B=0.00
C=0.15
P
I
H
E
I
A
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-5
-4
-3
-2
-1
0
1
2
3
4
5
THETA

FIGURE 1

AD P001395

# Test Development Issues for Adaptive Testing

Carol Anne Dwyer
Educational Testing Service
Princeton, NJ

Paper presented to the 23rd Annual Military Testing Association Conference, Arlington, VA, October 25-30, 1981.

## Introduction

The goal of this presentation is two-fold: to describe the various approaches to adaptive testing that have been implemented, planned, or considered for use at ETS; and to explain test development elements as practiced at ETS that impact development of adaptive tests.

## What is Test Development?

In general one might say that the test development process is driven by such general measurement concerns as validity, rather than by statistically based concerns in the strictest sense. The test development process as practiced at ETS is lengthy and complex, as it is for all large scale testing programs, but it can for convenience be partitioned into five major areas of activity:

Planning. This includes defining the tests to be developed (for example, developing a normatively-based selection test vs. diagnostic testing); determining the appropriate test format and size of the item pool; and developing specifications that include not only statistical targets, but also characteristics of the test materials themselves such as the nature of the stimulus materials, item context and content, and the skills to be assessed.

Generation of acceptable materials. This element of the test development process includes the writing and review of test materials in sufficient quantity within the allocated resources of time and money. Issues pertinent to this area include choice of writers for materials, quality of items produced in first draft (including achievement of target difficulty levels), and specifications for item-writing algorithms.

Assembly or inventory maintenance of test materials. This area includes issues related to actual or potential groupings of items, and problems associated with these groupings such as item context, item content, key distribution, and word overlap or content redundancy. General administration issues such as directions and sample questions are also part of this area.

Quality control and maintenance. This includes such issues as test security and establishing and maintaining accuracy of keys and item coding.

Review of item data for cyclical improvement of tests or item pools. This includes checking on prior estimates of item difficulty for future efficiencies as well as elimination of items that have become obsolete or whose statistical characteristics are not robust on repeated administrations.

Each of these five areas raises issues having applicability to one or more modes and purposes of adaptive testing.

## Variation in Adaptive Testing

Adaptive testing may be broadly conceptualized as having one of two fundamental aims: increased efficiency of interpretation and administration;

or complex diagnoses of skills, strengths and weaknesses. In general, this division is associated with two corresponding modes of item selection, whose development and selection are related to one's primary purpose for testing. The first of these focuses only on the correctness or incorrectness of an answer to an item at a particular point, with a resulting decision to adjust the difficulty level of the succeeding item based on the correctness or incorrectness of the prior choice. The second major strategy for item selection, generally related to more complex testing needs of diagnosis, focuses on the nature of the option chosen as well as its correctness or incorrectness. In this model, the succeeding item leads to a separate content or skill area depending on the option chosen. This may or may not be accompanied by a shift in the difficulty level of the succeeding item.

Which of these methods will be selected is not only determined by the purpose of testing but is also related to the establishment of test specifications. Specifications are designed differently (and items are written differently) for these two purposes. In addition, practical aspects of delivery mode and administration will differ for these two types of adaptive tests. In practice these two major types of adaptive testing share a number of problems but have some unique aspects as well.

Interest in adaptive testing at ETS appears to stem from three major sources that have influenced selection and development of projects. The first of these, and probably the most widely recognized, is an interest in the psychometric implications of current or projected adaptive testing applications. The second is practical exigencies (for example, limited testing time, or the need for complex diagnostic information about individuals). The third of these, less commonly recognized than the preceding two, is a concern for the morale and motivation of candidates to whom tests are administered. At ETS there is growing concern with candidates' being required to attempt responses to many items for which there is a very low probability of their responding correctly. (The converse is also of concern, the situation in which more able test takers are required to respond to large numbers of items for which there is a very high probability of their answering all correctly).

ETS is currently considering, for example, development of two-level sequential tests for a major admissions program, where currently a score at the 50th percentile may require answering fewer than half of the items presented correctly, and the top 37 percent of the score scale serves only to differentiate among the top 2 percent of test takers.

Projects at ETS have utilized a wide variety of modes of delivery over the past 20 years or so (although I believe the adaptive test first appeared in recognizable form in 1916--Louis Terman's revision of the Binet-Simon scale). These range from traditional paper-and-pencil locator tests developed to meet specific projects' needs (such as those developed for the Sequential Tests of Educational Progress), to computer-delivered branching tests that serve as prototypes for long-range technological planning.

Between these two extremes we have worked with self-paced branching paper-and-pencil test utilizing latent images; and Kodak microfiche-delivered self-guided diagnostic branching tests and supporting materials in mathematics.

## Test Development Issues

In the following sections, it seems appropriate to focus primarily on computer-assisted delivery of items for general branching (not highly diagnostic) use, because this particular combination both shows the most promise and presents some of the most difficult problems for test development in adaptive testing. But what are some of the most salient aspects of the test development process that support such applications?

As mentioned earlier, the first phase of the test development process is planning, which includes determination of test specifications. In a wide variety of contexts, the adaptive testing literature appears to focus primarily on statistical aspects of an item pool. For example, Warm (1978) mentions eleven pieces of information about a group of test items necessary to code and store for certain types of item sharing, but no mention is made in the list of content specifications other than the simple naming of the dimension to be tested.

This is clearly not the equivalent of complete test specifications such as are required for large scale assessment in more traditional modes. Specifications for such programs usually include information about four areas:

- o content and skills--domains to be sampled, relative weight to be given to each domain, skills required, balance of curricular or other content-related differences
- o test and item format--selection of item types most clearly related to content or skills, appropriate level of language or reading ability, directions and sample items, multiple-choice or other format
- o psychometric--required level of difficulty, distribution of item difficulties, target homogeneity of items, equating requirements, number of items and time alloted
- o sensitivity--requirement for materials reflecting diverse cultural backgrounds or special experiences

Many of these specifications areas are, of course, ordinarily specified in discussions of adaptive tests. Among these are the purpose of the test, the number of items contained in the set, and target distribution of item difficulties. Other specification areas are clearly not relevant to most adaptive tests, such as test length, mean difficulty of the item pool, and desirability or degree of speededness. Some specifications areas should be detailed for adaptive testing but usually are not, such as content, sensitivity context, and nature of stimulus materials.

The following remarks will focus primarily on areas that need closer scrutiny than they typically receive. In addition, I will discuss other, more general, test development issues that are not variable across individual tests, such as key distribution, conceptual and word overlap, item ordering, and instructions.

Content, skills, and stimulus materials specifications

As mentioned earlier, little attention has been given to content specifications in adaptive testing. IRT applications of course assume unidimensionality of the items in a pool. While the determination of the degree to which this assumption has been met within a particular pool is difficult, in practice unidimensionality appears usually to be quite broadly construed and widely taken for granted in the construction of item pools for adaptive testing. Perhaps it has been this orientation toward assuming unidimensionality that has deflected attention from development of detailed specifications of item content, such as are usually required of item sets for large-scale assessment in other modes.

Despite the necessary assumption of item unidimensionality, there is some recent evidence (Swinton and Powers, 1981) that tests usually considered to be unifactorial may not in fact be so for subgroups of the test-taking population, and that additional factors present for subgroups may correspond to content specifications. Swinton and Powers found that in the Graduate Management Admissions Test (GMAT), the verbal sections were unidimensional for females but not for males. For males, there were two verbal factors: reading and grammar.

It is quite likely that many large-scale applications of adaptive testing will require a very broad range of content (and possibly of item types as well), if the pool is to be of practical utility for large numbers of subjects with a wide range of ability and preparation. This creates a high probability of people taking different looking tests (at least) and possibly even tests with differing factorial structures. These differences in factorial structure can, to a great extent, be controlled by proper classification and representation of various item types, with corresponding programming for branching.

There are clearly at least two approaches to dealing with this question of factorial structure and unidimensionality. The first is restriction of the item pool to a limited (but clearly specified) range of content, whose known factorial structure does not differ for salient population subgroups. The second is the organization of branching strategies, such that content subareas are proportionally represented to all test takers. This latter approach seems more defensible for most practical applications, especially because the desired range of thetas may in fact rule out the former approach since not every content area and skill is associated with the entire range of item difficulties.

Controlling content through detailed specifications can be expected to add to the general utility of the test, as well as to its psychometric properties. For example, when two test-takers are presented with a sequence of items to test their reading ability, control of content specifications could insure that the less able of the test-takers did not receive only items requiring literal comprehension of newspaper advertisements, while the more able test-taker received only items dealing with inferences concerning scientific expository materials.

Without control of content specifications, it is quite likely that if two test takers did receive and respond to such disparate item sets, one might demonstrate through statistical analyses that the resulting scores could meaningfully be compared. But even the most ardent devotee of IRT applications could sense that the general public of test takers and data consumers might be dissatisfied with this situation.

Moreover, if this situation were to obtain for even a small proportion of the test takers, issues of legal defensibility might well arise, And, unfortunately, there is little in the current literature on parallelism of test forms that would be likely to be of assistance in this situation. There is no generally accepted definition of test parallelism to fall back on; we can only attempt to demonstrate careful control of likely sources of irrelevant variance: and content is certainly one of these sources.

How can one best specify test content? For most adaptive testing applications that test content should be specified as completely as it traditionally has been for other modes of testing: a matrix of major content areas and cognitive skills with proportions specified, coupled with sampling of content subareas within these cells. Moreover, to the degree possible, each cell should contain easy, medium, and difficult items. Item presentation rules should provide each test taker with a constant mix of items, at the difficulty levels appropriate to each individual.

In 1979, ETS commissioned a preliminary study of item classification for a wide range of potential applications, including item library support, computer-assisted test assembly, and item banking. This study sheds some light on the complexity of item classification for adaptive testing.

Poyner, Lippey, and Buntaine (1979), responding to needs created by the size of ETS's total item pool, devised a general classification strategy appropriate to a wide range of subject matter areas. ETS at that time reused 30,000 items annually, and added up to 40,000 items to its files annually. The classification strategy developed by Poyner and his associates included three "dimensions" or categories of classification. General dimensions were those that contained information that all items are expected to have, such as correct answer, security status, and intended difficulty or appropriateness for age levels. Cumulative dimensions were intended to store historical use and statistical data associated with items. These dimensions were characterized by Poyner as being open-ended, in the sense that more recent data augment, rather than replace, previous data. The third category, collection dimensions, is used for classifying content-dependent attributes of items. Collection dimensions are structured much like an outline of a subject area and in Poyner's system can contain up to five levels of subheading. Table 1 shows a sample collection outline for mathematics. The number of levels actually used depends upon the particular collection. Some collections may require that three of the available levels be used, while others may require two, four, or five. Within a single collection, different numbers of levels may be defined in different areas of the collection. In addition, a second field, alternate category, was defined for use with items that could be classified under more than one category.

It is probably apparent from this description that the development of content codes for adaptive testing is of great practical importance but of considerable complexity.

## Issues in item-writing and review

I would like to turn now to another set of issues that is dealt with during the writing and review segment of the test development process. This set of issues relates directly to the much-discussed concept of test-wiseness. Test development has evolved a series of very detailed "rules" for writing and reviewing test materials. These rules are linked primarily by their intention of focusing the test taker on the test content itself, rather than on irrelevant attributes of the test or its format, such as the placement correct answers. Unfortunately, it seems popular to attempt to focus test takers on these irrelevant attributes rather than on the content. For example, a recent popular book advises students that when they are in doubt as to a correct answer they should guess either B or D.

Some item-writing rules pertain to any well-constructed test. These include attention to the length of each option, establishment of a parallel grammatical structure for all options, the avoidance of value judgments or inflammatory language, consistency of the degree of qualification of correct and incorrect answers, and the avoidance of use of negatives, especially the double negative. Other item writing and reviewing rules, however, present issues of particular importance for adaptive testing. These include such questions as key distribution, consistency of style and format across an item pool, the ordering and grouping of items, and avoidance of word or conceptual overlap among items within the same pool.

Key distribution. In ordinary test development procedures, the test assembler uses various ploys to avoid having the position of the correct answers form a pattern that might become a source of distraction to the test taker. For example, test assemblers avoid over-representation of one key position, clusters of items with the same key, and chains of correct answers in sequence (such as A, B, C, A, B, C, A, B, C). In computer-delivered adaptive testing, the test taker does not have present the entire array of choices that he or she has made, but obvious patterns in correct answers will still be perceived by the test taker, and will be even more troubling since answers cannot ordinarily be changed once the next item has been presented.

There are two approaches to the problem of key distribution that readily come to mind: an interactive provision for revising the position of the correct option based on the previous responses of the individual test taker; and selection of the next item to be administered partially on the basis of the position of its key. In my opinion, the latter is by far the preferable alternative. Changing option order may alter the difficulty of the item. Items with the key in position D and E are ordinarily expected to be slightly more difficult than the same item with the correct answer in position A or B. Moreover, many items particularly in mathematics and science have their options arranged in logical order and simply do not lend themselves to option rearrangement.

Consistency of style. When a set of items tests a very wide range of ability, it tends to have a number of characteristic styles associated with ability levels. These stylistic differences may include differences in wording and placement of directions, diction (for example, in reading comprehension is the stimulus material referred to as the "story" or the "passage"?), the number of options associated with each item, and the use of sample questions. Any of these stylistic and format differences can be distracting to the test taker and consequently disruptive to test interpretation when a test taker has answered a set of items incorporating diverse styles. Variation in the number of options may have additional psychometric ramifications in adaptive testing.

Item overlap. Another test development issue that merits special consideration in the case of adaptive testing is item overlap. Certain item types such as definitional items and certain types of verbal reasoning items require that there be no overlap of key words or phrases, in order that knowing one piece of information should not be given what is in effect double credit. With any item type or group of items one must also be concerned with conceptual overlap; that is, does one item inadvertently provide information to help answer another item? These concepts take on special meaning in applications of adaptive testing, particularly in light of concern with local independence of items. With the elements of the item set unknown before the test taker responds to them, word overlap cannot easily be controlled, unless one is willing to prescribe no overlap across the entire item pool. This practice, however, maybe unduly restrictive in large item pools.

Item overlap at the conceptual level may be very subtle and difficult to detect, particularly in items testing higher-level cognitive skills. At ETS a separate review aimed at identifying this type of overlap is conducted by editorial specialists. At the extreme, these individuals may work with groups of very difficult items that number up to approximately 200, and this is probably close to the limit of feasibility for detecting conceptual overlap. When much larger item pools need to be checked for conceptual overlap, it appears that the most straightforward solution is simply to partition the pool into groups of items that are likely to address similar problems or that have been designated as measuring similar content. Overlapping item sets may also be designated as an additional aid to detecting overlap. The test content and skill specifications provide a guide for such partitioning of item pools. These smaller sets of items may then be reviewed for conceptual overlap by subject-matter specialists. This implies, of course, that one is willing to tolerate the remaining risk of overlap across the item subgroups.

The use of item sets. The use of item sets represents a particular concern in adaptive testing for a number of reasons. The first of these is the fact that item sets of some types are suspected of violating the IRT assumption of local independence of items. In addition, there is the practical problem of the length of stimulus materials for certain types of item sets. For example, the stimuli for many reading comprehension sets, particularly those that ask difficult or higher-level cognitive skill questions, are quite lengthy. In some cases, the stimulus material may not all fit on the computer screen at one time, much less be presented along with its associated items.

There is another practical problem area in the use of item sets: Most branching strategies are devised using the single item as the basic decision unit. This implies that the stimulus material, no matter how lengthy, would be presented with only one of the set of items associated with it. In the case of a reading comprehension test, a test taker might be required to read a 700-word passage in order to answer a single question. In the worst instance, the student might at a later point in the testing be required to read the same 700-word passage again and answer a different single question. It is apparent that such a situation would largely negate any gains in efficiency expected because of the adaptive testing mode.

One solution to this set of problems is simply to eliminate the use of item sets in any form. There is a drawback associated with this, however: this practice can unduly restrict the range of subject matter that can be tested. Many complex and sophisticated questions require lengthy stimulus materials and are therefore more efficiently measured in sets. An alternate solution has been proposed.[1] Stimulus material is presented at the appropriate time with the complete set of questions associated with that material. At the initial presentation only one of these items is of immediate interest for the adaptive test. This item is scored in the usual way and the appropriate follow-up item is then presented. Other members of the original item set are given priority for selection in future branching decisions. These items need not reappear to the test-taker since they are retained in memory and will simply be scored at the appropriate time. The test taker will then be branched to the next appropriate item. Certain items in the set will, of course, never be required by the branching rules and will thus never be scored. These represent wasted time for the test taker.

Directions and examples. In designing directions and sample items for adaptive testing, the first difference from traditional testing that becomes apparent is that directions and sample items are no longer applicable to complete blocks of items. By the very nature of adaptive testing, item types and formats are mixed rather than presented together in assembled blocks. In a recent application at ETS, full directions for each item type were presented before the first item of each type. When the second or subsequent item of the same type was later presented, an abbreviated form of the directions was displayed along with the item at the top of the screen. It should be noted, however, that this method implies that some item types are incompatible in item ordering. For example, the format of synonym and antonym items is highly similar, and even with abbreviated directions displayed on the screen above the item it is difficult for a test taker to avoid confusion about which item type is currently being presented.

Feedback to test developers. In any item pool certain items will eventually be discovered to be flawed or to have become obsolete. An important element in identifying such problematic items is candidate feedback. In order to facilitate such feedback, it is desirable that in adaptive testing an item identifier be presented with each item taken so that the test taker can accurately identify any item that he or she wishes to critique. Obviously the simple sequential

[1] W. M. McPeek (ETS), personal communication, September, 1981.

item numbering of traditional paper-and-pencil tests is not appropriate in this situation. This candidate feedback information may also be considered an important step in demonstrating to test takers the test developers' commitment to preventing bias in the testing materials.

It should be clear from the preceding material that there are a great many test development issues still to be resolved for most applications of adaptive testing. However, we are in the midst of no less than a measurement revolution, and the great potential of IRT-supported adaptive testing deserves no less attention to the quality of the materials delivered than do more traditional testing modes. We must resist the temptation to become so embroiled in mechanics or technology that we lose sight of our basic measurement goals.

## References


Poyner, H., Lippey, G., and Buntaine, D. Item Classification Study (TM6746/Contract No. 15405). Santa Monica, CA: Systems Development Corporation, 1979.

Swinton, S. and Powers, D. The Construct Validity of the Graduate Management Admissions Test. Princeton, NJ: The Graduate Management Admissions Council, 1981.

Terman, L. M. The Measurement of Intelligence. Boston: Houghton Mifflin, 1916.

Warm, Thomas A. A primer of item response theory (Technical report US-CG-941278) Oklahoma City, OK: U.S. Coast Guard Institute, 1978.

Table 1

Sample Collection Outline

| Coded Positions | | | | | Subject Matter |
|---|---|---|---|---|---|
| 1-2 | 3 | 4 | 5 | 6 | |
| AL | 0 | 0 | 0 | 0 | I. Algebra |
| AL | 1 | 0 | 0 | 0 | A. Algebraic Operations on Algebraic Expressions with Real Numbers |
| AL | 1 | 1 | 0 | 0 | 1. Addition |
| AL | 1 | 2 | 0 | 0 | 2. Subtraction |
| AL | 1 | 3 | 0 | 0 | 3. Multiplication |
| AL | 1 | 4 | 0 | 0 | 4. Division |
| AL | 1 | 5 | 0 | 0 | 5. Parentheses |
| AL | 2 | 0 | 0 | 0 | B. Translation (except %) |
| AL | 2 | 1 | 0 | 0 | 1. Writing symbol expression for verbal statement |
| AL | 2 | 2 | 0 | 0 | 2. Writing equation for solution of problem |
| AL | 3 | 0 | 0 | 0 | C. Linear Functions |
| AL | 3 | 1 | 0 | 0 | 1. Solution - Linear Equations |
| AL | 3 | 1 | 1 | 0 | a. Integral Coefficients |
| AL | 3 | 1 | 2 | 0 | b. Rational Coefficients |
| AL | 3 | 1 | 3 | 0 | c. Irrational Coefficients |
| AL | 3 | 2 | 0 | 0 | 2. Solution - Linear Inequality |

## Psychometric Considerations for Adaptive Testing Systems: Administration and Validation

Bejar, Isaac I., (Chair); Jones, Douglas H.; Rock, Donald A.; and Wainer, Howard, Educational Testing Service, Princeton, New Jersey; Waters, Brian, and Lee, Gus C., HumRRO, Alexandria, Virginia.

Once the item pools have been developed and calibrated, it becomes possible to administer tests adaptively. Much research has been done with synthetic data to evaluate different procedures for test administration and scoring. In practical applications we are likely to find problems that are not present with synthetic data. Two components of the adaptive test, item selection and ability estimation, are especially vulnerable to empirical disturbances and will be the focus of two of the papers. The validity of adaptive testing is also an inherently empirical question. A framework for validating adaptive testing will be the focus of a third paper. The last paper will discuss the proposals offered by the previous three papers from a legal perspective.

PREVIOUS PAGE IS BLANK

AD P001396

# Foundations for the Mathematical Notion of Information in Item Response Theory and Robust Ability Estimation

Douglas H. Jones
Educational Testing Service

Abstract

Much attention is devoted to the information function associated with a mental test. This is based on the belief that the estimate of ability is approximately normally distributed with mean value equal to the true ability parameter and variance equal to the reciprocal of the information function. Computerized adaptive testing is particularly dependent on the truth of this assertion, since the test is constructed for an individual examinee at the same time the test is taken. This is done in order to maximize the information function for the most recently available estimate of ability.

A mathematical setting based on a statistical sampling probability mechanism will be described. In this setting, a mathematical meaning is given for the information function; and it becomes possible to study the relative merits of various ability-estimating procedures.

Among these estimation procedures that will be compared are the maximum likelihood estimation procedure under the one-, two-, three-parameter logistic response model, and some new procedures that are suggested by resistant/robust theory.

# Foundations For the Mathematical Notion of Information in Item Response Theory and Robust Ability Estimation

Douglas H. Jones
Educational Testing Service
Princeton, New Jersey

Item response theory is an attempt to apply statistical theory and methodology to inferences about a subject's ability. The statistical theory receiving the most attention has been that based on the Fisherian concept of information and the associated principals and methods of maximum likelihood (Birnbaum, 1967). Forsaking Fisherian optimality to gain estimators with high stability, researcher's have recently attempted to extend item response theory to realistic testing situations by employing concepts of robust statistical theory (Wainer and Wright, 1980; Bock and Mislevy, 1981). Although robust statistical theory, as developed by Huber (1981), is primarily concerned with discovering and protecting oneself against the least informative model close by the ideal model that can explain the mechanics of inconsistent observations, item response theory researchers have developed estimators of ability that imitate the way the maximum likelihood estimator, associated with Huber's least favorable models, reduces the impact of maverick observations. While Huber's theory yields variance estimates as an easy by-product of his methods, adequate variance formulas are not available for inferences in item response theory based on robust estimators developed thus far. Extending Huber's elegant concept of robust theory to the item response problem is a worthwhile task for it may afford practitioners with streamlined techniques for making accurate inferences. In the meantime, however, our task is that of meeting the immediate problem of practitioners required to make inferences under adverse conditions, even if this is accomplished with ad hoc estimators.

This article describes an attempt to develop the concept of influence in item response theory that parallels a similar concept in robust theory in order that the methods of robust theory be better applied although the results may be ad hoc. Some of the progress in this direction has resulted in a new class (h-estimators) of robust estimators of ability and a systematic approach to the derivation of bias and variance formulas. Not only may these formulas be used for constructing confidence intervals for a subject's ability, they may be used to compare the efficiency of the h-estimators to that of the standard maximum likelihood estimator. An example will be given showing that the worse the h-estimators can do is lose 10% in efficiency, or one item in ten, relative to the maximum likelihood estimator, while maintaining strong stability in the presence of maverick observations -- a property that is absent from the behavior of the maximum likelihood estimator.

After the preliminary notation and results are presented, which hopefully are adequate for informing the reader about robust methods, attention will be again directed to Huber's robust statistical theory and its applicability to item response theory will be discussed. It should be noted that one must be very careful in assuming forms of the item probability of correct response since not every model can adequately describe observed data (Holland,1981 ).

This article will conclude with a brief discussion on the theory of experimental design and how it provides insightful motivation for some methodology in item response theory, including the important computerized adaptive test.

## Preliminaries

If $p(x;\theta)$ is the probability frequency function of the response, $X=x$, to a given item by a subject with ability $\theta$ , the Fisher information in the response X about the parameter $\theta$ is defined as $I_X(\theta) = E\{[\partial_p (X;\theta)/\partial\theta]\}^2$. The term information is used for this quantity, since the joint information in two independent responses X and Y, $I_{(X,Y)}(\theta)$ is the sum of the individual information values:

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta) , \qquad (1)$$

and the average information per observation is the last quantity divided by two. In general, the average information per observation in n item responses $X_1,\ldots, X_n$ is

$$I_n(\theta) = \Sigma I_{X_i}(\theta)/n \qquad (2)$$

where the sum is from 1 to n. It is easy to show that if only one observation is taken with equal probability from $X_1,\ldots, X_N$ where N is different from n, then $I_N(\theta)$ is the information associated with that observation. This concept also generalizes to choosing observation $X_j$ with probability $g_j$ to give the formula for the information per observation equal to

$$I_g(\theta) = \Sigma\, g_j I_{X_j}(\theta) . \qquad (3)$$

where the sum is from 1 to N.

It is easily observed that the information in n independent observations chosen in this fashion is $nI_g(\theta)$ .

In practice, a response is either right (X=1) or wrong (X=0) and the information formula reduces to a simple form:

$$E[\partial \log p(X;\theta)/\partial\theta]^2 = (P')^2/[PQ] \qquad (4)$$

where $P \equiv P(\theta) = p(1;\theta)$, $Q = 1-P$ and $P' = \partial p(1;\theta)/\partial\theta$ .

For illustration and examples, attention will be directed to the two-parameter logistic response function:

$$P(\theta) = \exp[a(\theta-b)]/[1 + \exp[a(\theta-b)] \qquad (5)$$

where $a > 0$, the discrimination parameter; $-\infty < b < \infty$ , the difficulty parameter. The following results are applicable to the three-parameter models; however, since the formulas have simpler forms, the two-parameter model is used throughout.

Since $P'(\theta) = aP(\theta)Q(\theta)$, the Fisher information simplifies. Using this simplification and denoting by $P_i(\theta)$ the probability that $X_i=1$, the information in n responses is

$$I_n(\theta) = \Sigma\, a_i^2\, P_i(\theta)\, Q_i(\theta) \qquad (6)$$

and the information in n responses, chosen independently from a pool of N items such that $X_j$ has probability $g_j$ of being chosen, is equal to

$$nI_g(\theta) = n\, \Sigma\, g_j a_j^2\, P_j(\theta)Q_j(\theta).$$

The log-likelihood function for n observations is equal to

$$\ell(\theta) = \Sigma[x_i \log P_i(\theta) + (1-x_i) \log Q_i(\theta)] \qquad (7)$$

and the maximum likelihood estimator $\hat{\theta}$ is the value of $\theta$ that maximizes $\ell(\theta)$. The maximum likelihood estimator is found by solving implicitly in the equation $\ell'(\hat{\theta}) = 0$; that is

$$\ell'(\hat{\theta}) = \Sigma\, a_i[x_i - P_i(\hat{\theta})] = 0. \qquad (8)$$

Example: With n = 10, $b_i$ = - 0.8, 1.0 by .2, $a_i \equiv 1$, i = 1, 10

x = (1,1,1,1,1,0,0,0,0,0,), recursive solution of (8) yields the value $\hat{\theta}$ =0.11.

If the number of observations is sufficiently large,

$$[\hat{\theta} - \theta]/[-\ell''(\hat{\theta})]^{-\frac{1}{2}} \qquad (9)$$

has approximately the standard normal distribution thus affording an opportunity for one to construct confidence intervals that are likely to contain the true value θ. A 95% confidence interval has the form

$$\hat{\theta} - 2\,[-\ell''(\hat{\theta})]^{\frac{1}{2}} \;,\; \hat{\theta} + 2\,[-\ell''(\hat{\theta})]^{\frac{1}{2}} .$$

The variance can be shown to satisfy

$$1/-\ell''(\hat{\theta}) = 1/\Sigma\, a_i P_i'(\hat{\theta}) = 1/\Sigma a_i^2 P_i(\hat{\theta})Q(\hat{\theta}) . \qquad (10)$$

Comparing this last value with the formula for Fisher's information we see that the variance of the maximum likelihood estimator is the reciprocal of the Fisher information and this is true in general (Chernoff. 1975). In (9) the denominator may be replaced by $[nI_g(\hat{\theta})]^{-\frac{1}{2}}$ and the normal approximation continues to hold.

The important property of maximum likelihood estimation is its claim to optimality. For if $T_n$ is any other estimator of θ that is unbiased and $V_n(\theta) = E(T_n - \theta)^2$, the mean squared error, then by the Cramer-Rao inequality, $V_n(\theta) \geq 1/nI_g(\theta)$. And therefore one can not do better in estimating than the reciprocal of the Fisher information.

## Maverick Observations (Outliers)

Suppose several responses are inconsistent with the majority of the responses. For example, a subject, whose majority of responses indicate mediocre ability, gets a very difficult item correct. Thus, the maximum likelihood estimator is typically overly

influenced by the inconsistent response as the following example illustrates.

Example: Let the parameters have the same values as the previous example but let $x = (1,1,1,1,1,0,0,0,0,1)$, then $\hat{\theta} = .58$.

The primary reason for this distortion is that the implicit equation (8) weights equivalently the contribution of each item to the sum. Whenever a response is inconsistent with the rest of the responses, the difference $x_i - P_i(\hat{\theta}_{/i})$ is large ($\hat{\theta}_{/i}$ denotes the estimate obtained by deleting the ith response), and overly influencing the value of the sum.

Let $h \geq 0$ and consider the estimator that is the solution to the equation,

$$\Sigma\, a_i[x_i - P_i(\hat{\theta})][P_i(\hat{\theta})Q_i(\hat{\theta})]^h = 0 .$$

These estimators, denoted h-estimators, are not overly influenced by maverick responses. If $x_i - P_i(\hat{\theta}_{/i})$ is large, then, $[P_i(\hat{\theta}_{/i})Q_i(\hat{\theta}_{/i})]^h$ is small, so that the product does not contribute a disproportionate value to the sum.

Example: Let the parameters be as in previous examples. In the following table is displayed values of h-estimators for selected values of h.

| Responses | h | | | | |
|---|---|---|---|---|---|
| x | 0(MLE) | 0.5 | 1 | 2 | 3 |
| 1 1 1 1 1 0 0 0 0 0 | 0.11 | 0.11+ | 0.14 | 0.13 | 0.12 |
| 1 1 1 1 1 0 0 0 0 1 | 0.58 | 0.41 | 0.22 | 0.20 | 0.19 |

Note that h-estimators contain the maximum likelihood estimator as a special case.

The variance of the h-estimator is given by the formula

$$A(\theta;F,h) = \left[ n\Sigma\, g_j a_j \overline{PQ}_j^{\,2h+1} \right] \Big/ \left[ n\Sigma g_j \overline{PQ}_j^{\,h+1} [1+h(Q_j-P_j)/\overline{PQ}_j] \right]^2$$

where $F=(P_j(\theta),\ j=1,\ldots,\ N;g)$ and is used to denote the dependence of the variance on the underlying probability model for response and item selection, $\overline{PQ}$ denotes the product and the explicit dependence on $\theta$ is suppressed for notational convenience. Note that $A(\theta;F,0) = 1/nI_g(\theta)$ .

The term efficiency will be used to denote the ratio of variance of two estimators (or in case of biased estimators, ratio of mean squared errors). If the values of the parameters are the same as in the examples, the calculated efficiency, maximum likelihood to h-estimators, are displayed for selected values of h.

| h | 0.5 | 1.0 | 2.0 | 3.0 | 4.0 |
|---|---|---|---|---|---|
| EFF | .99 | .98 | .95 | .92 | .90 |

These results are most promising for they show that the h-estimators have high efficiency even for large values of h, when h-estimators are least affected by inconsistent observations.

Asymptotic variance formulas are also available for true item response probabilities that are different from the assumed forms. Denote the true item response function for item j by $P_j^*(\theta)$. Then for $F^*=(P^*_j,\ j=1,\ \ldots,\ N,g)$, the variance is given by

$$A(\theta;F^*,h) = n\left[\Sigma g_j a_j^2\, \overline{PQ}_j^{*}\, \overline{PQ}_j^{2h} + \Sigma g_j a_j^2\, (P_j^*-P_j)^2 \overline{PQ}_j^{2h}\right] \Big/ \left[n\Sigma\ g_j a_j^2\, \overline{PQ}_j^{h+1}[1+h(Q_j-P_j)/\overline{PQ}_j]\right]^2 .$$

Under models other than the assumed one, both the maximum likelihood estimator and the h-estimators are asymptotically biased. Formulas for the bias are given by

$$B(\theta;F^*,h) = \Sigma g_j a_j \overline{PQ}_j^{h}(P_j^* - P_j) / \Sigma g_j a_j^2 \overline{PQ}^{h+1} .$$

Combining the variance and square of the bias allows one to determine the relative efficiency under $F^*$. The only decision that must be made is which $F^*$. Meaningful alternatives are difficult to formulate and this is an open area needing much more exploration. More on this will be discussed later.

Example: (Distortion of Difficulty Parameters). Suppose each $P^*_j$ has a difficulty parameter $b_j^* = b_j - 0.1$, then in the following table are displayed the relative efficiencies for selected values of h.

| h | 0.5 | 1.0 | 2.0 | 3.0 | 4.0 |
|---|---|---|---|---|---|
| EFF | 2.24 | 4.83 | 19.15 | 51.46 | 82.62 |

This example illustrates how astonishingly bad the maximum likelihood estimator can be relative to the ad hoc h-estimators.

## Finding Suitable Models

The problem of finding models that appropriately describe maverick responses was touched upon in the last section. Should reasonable models be discovered, then neighborhoods of these models surrounding the ideal model could be investigated for the one yielding the least Fisher information. As protection from this least favorable member one could employ the best possible estimator available for observations generated by it. In this fashion, one would naturally be lead to a "robust" methodology and this is the general program followed in standard statistical literature (Huber, 1981). Thus one is tempted to employ some of those ideas in constructing models for item responses. However, derived models may not be suitable when compared for fit on data aggretated over many subjects.

## Sequential Design

As derived earlier, the Fisher information per observation X, $I_X(\theta)$, depends on the unknown ability parameters. This presents a difficulty when one tries to select an item to maximize the information. One would attempt to use some prior information to locate $\theta$ in some interval, then zero in with repeated items to yield the highest information. A method that parallels a practice in bioassay is to initially administer items widely scattered throughout the difficulty range of the item pool in order to create the prior information. Based on an estimate of $\hat{\theta}$, then, item j would be selected if $I_{X_j}(\hat{\theta})$ were largest among the remaining items.

Several ideas discussed thus far impact on this program. $I_{X_j}(\hat{\theta})$ is the quantity to maximize over j, only if the estimator is the maximum likelihood estimator. If any other kind of estimator is used, such as h-estimators or other "robust" estimators, the variance formula must be minimized by the new item that will be administered. Therefore reasonable approximations to these variance formulas are required to do the job.

Proper comparisons between sequential design schemes and proper inferences of the subject's ability may only be made if the correct repeated sampling variances or information measures are known. The value of $I_{X_i}(\hat{\theta})$ obtained at the end of a sequentially administered test is not the Fisher information of the probability response function; nor would the similarly obtained variance formula be correct for robust estimators. The problem lies in the fact that the item responses are not independent. In fact, the probability frequency function of $X_{i+1}$ depends on the responses $X_1, \ldots, X_i$ since the item is chosen according to these responses through the value of $\hat{\theta}$. The relevant Fisher information for the i+1 observation is the expectation over $X_1, \ldots, X_i$ of the quantity

$$E[\partial_p(X_{i+1}|X_1, \ldots, X_i)/\partial\theta]^2 .$$

Once the Fisher information of an n item sequential test is obtained it should then be compared to $nI_g(\theta)$. If the information value in the sequential test is not appreciably higher than the fixed sample test, then the additional cost of sequential item administration may not be justified on the grounds of improved accuracy alone.

## References


Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability (Part 5). In F.Lord & M. Novick, Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1967.

Bock, R. D., and Mislevy, R. J. Bieweight Estimates of Latent Ability, Manuscript, 1981.

Holland, P. W. When are Item Response Models Consistent with Observed Data? Psychometrika, Vol. 46, No. 1, 1981.

Huber, P. J. Robust Statistics. New York: John Wiley, 1981.

Wainer, H., and Wright, B. D. Robust Estimation of Ability in the Rasch Model, Psychometrika, Vol. 45, No. 3, 1980.

AD P001397

# Validity Considerations for Adaptive Testing Systems

Donald A. Rock and Isaac I. Bejar

Educational Testing Service
Princeton, New Jersey 08541

Almost all tests today are administered in a paper-and-pencil mode. That is, a group of testees gather to take the test, which has been previously printed in a booklet. Responses to each question are recorded on an answer sheet. Generally, everyone responds to the same set of questions unless more than one form of the test is being administered for administrative or scoring reasons. I shall refer to this mode of test administration as conventional.

Perhaps the major reason for the almost exclusive use of conventional testing is economy. It is relatively inexpensive to print tests. Costs may be reduced even further if the printed tests are reusable. The staff needed for administering a conventional test consists of at most a few monitors. Thus the cost of testing per testee is minimal. Although the cost effectiveness of the conventional procedure is undeniable, developments in psychometrics, and computer science, have made it possible to administer tests by computer.

## Advantages of CAT

From a psychometric point of view, there are two distinct advantages to computer administration. One is increased precision of measurement, the other is a more controlled testing environment. To the extent that these advantages improve the validity of test scores, it may be argued that computerized testing is capable of breaking the "plateau" in validity of conventional testing which currently exists These two central advantages of computerized testing go back at least to Binet (cf. Weiss, 1973). Because of the nature of his intelligence test and the population he tested, Binet found it necessary to administer the instrument on an individual basis. It must have become obvious after a short while that there was no point in administering the entire test to every individual. Instead, it was more efficient, and psychologically more judicious, to limit the test to those items which were "appropriate" in difficulty for a given individual. This is precisely what computerized adaptive testing is all about: constructing a test tailored to each individual test-taker. The "tailoring"

is done by the computer on a sequential basis; that is, what item is administered next depends on the testee's performance either on the previous item or on all previous items. The procedures for this tailoring are called adaptive testing strategies.

## Evaluation History of Adaptive Achievement Testing

Although there seems to be unanimous agreement that adaptive testing is, in principle, a good idea, its empirical evaluation has proven difficult in practice--for reviews, see Weiss (1973); Wood (1973); McBride (1976); Kreitzberg, Stocking, and Swanson (1978). On the whole, however, the research suggests that adaptive testing is superior to conventional paper-and-pencil testing in that it measures more precisely over a wider range of performance levels. Nevertheless, for several reasons, the existing research is less than adequate in evaluating the usefulness of adaptive testing for achievement.

Much of the research is based on simulation and analytical results. Monte Carlo and analytical evaluations of adaptive testing are to some extent misleading since they assume that the latent trait model fits responses obtained under the adaptive procedure. The problem is best appreciated by examining the procedure of fitting the model, e.g., estimating parameters, to a pool of items. One procedure would consist of administering a previously calibrated test along with the to-be-calibrated items. Since the number of items that can be given to any sample of testees is limited, additional samples would be required to calibrate the entire pool. The fact that all samples responded to a set of common items allows the estimations of the uncalibrated items on a common metric. Eventually, all the items are mixed in a pool and administered by computer. When an item finally appears on the screen of the computer terminal, there is no guarantee that responses to it can be accurately modeled using the previously obtained parameters, since the context in which the item appeared at calibration time is very different from the context in which it appeared in the adaptive test. Furthermore, in an adaptive test, there is reason to believe (e.g., Yen, 1980; Whiteley & Dawis, 1976) not only are the surrounding items different but also the medium of presentation. In addition, feedback is often given as part of the computerized administration; this raises the possibility of violations of the local independent assumption (see Gialluca & Weiss, 1980).

From a pragmatic point of view, the ultimate triumph or defeat of adaptive testing is not likely to depend entirely on how much information it extracts (i.e., reliability) but also

on internal and external validity considerations. Similarly, it is misleading to judge the content validity of adaptive tests as done by McKinley and Reckase (1980) especially when no attempt was made to have the administration program samples objective in the desired manner. These considerations suggest that a more defensible approach to the validation of adaptive testing, and educational measurement in general, is construct validation. (See Messick, 1975, for a discussion of construct validation of educational measures.)

## A Multi-purpose Validity Model

Psychometric theory suggests that computer adaptive testing should: (1) yield more accurate assessment for a fixed number of test items than paper and pencil linear procedures, and (2) maintain a high level of accuracy across subgroups differing in ability levels. However, the introduction of new assessment procedures such as Computer Adaptive Testing (CAT) which differ significantly from the traditional approaches introduces the posibility that individual performance scores may be effected in a new and/or different way by the presence of method variance that may or may not be independent of the trait content. Furthermore, the method component (if it exists in a non-trivial amount) may or may not be related to external criteria. Skills and/or attitudes that contribute to performance at the computer console (independent of the traits beng measured) may also be related to certain training or job performance criteria. The model proposed here takes these factors into consideration.

This paper will outline a general procedure for comparing CAT with more traditional assessment procedures with respect to: (1) validity generalization across traits and populations of different ability levels and (2) predictive validity. That is, we will apply a factor analytic generalization of Campbell and Fiske's (1959) multitrait-multi-method (MTMM) procedure to the problem of distinguishing method variance from trait variances within both a predictive as well as a construct validity framework.

A classical test theory formalization of the MTMM Model using maximum likelihood confirmatory factor analysis was presented by Jöreskog (1971), and Werts, Linn, and Jöreskog (1971) give a path analytic representation of the same multi-trait multi-method model. The approach suggested here is an application of Jöreskog's confirmatory factor analytic procedure to the problem of estimating and testing the invariance of the trait, method, and error score components across populations

characterized by different educational levels. Procedures will also be presented for examining the relationship between the traits, methods, and an external criterion.

To clarify the model, consider a situation where the validity of arithmetic reasoning, mechanical reasoning, and verbal ability traits are under investigation. Furthermore, assume three methods of assessment under investigation: paper and pencil linear, flexilevel (Lord, 1971), and a fully adaptive procedure such as the maximum likelihood or Owen's Bayesian. External criteria could be ratings on subsequent training performance. It is assumed that all subjects within all sub-populations are assessed on all three traits by all three methods. Through spiralling the order effects of both trait and method can be partially balanced along the two dimensions. It is further assumed that parallel item pools are available for each method within traits. Using item statistics, shorter forms of the various trait measures could be constructed which would keep the total testing time from being excessive. It is suggested that the various forms be constructed in such a way that the average number of items used across all subjects be approximately the same across all methods and traits.

Figure 1 presents the multi-trait multi-method model to be fitted. The path diagram for the MTMM matrix in Figure 1 shows P + M unobserved exogenous factors (where P = the number of variables or traits and M = the number of measurement methods), one for each separate trait and one for each separate method; and PM or nine observed endogenous-variables, one for each trait-method combination. This model suggests that non-random variation in each observed variable is due to exogenous sources of variation, one involving trait content and the other involving method content, and further that each exogenous factor is responsible for covariation among specific observed endogenous variables. As Alwin (1974) points out, this confirmatory factor analytic models allows for random sources of variation in each observed endogenous variable. This random variation includes specific variation unique to a particular measured variable and variation due to unreliability.

The dotted line from the trait and method factors to the external criterion indicates that this association between traits, methods, and criterion will be estimated after the multi-trait-multi-method construct validity model has been fitted. That is, this sequence is consistent with the validation paradigm which first examines whether or not you are measuring what you think you are and then asks what is the predictive validity of what has been measured with respect to meaningful outside criteria.

Figure 1

Path Diagram for a Multitrait-Multimethod Matrix

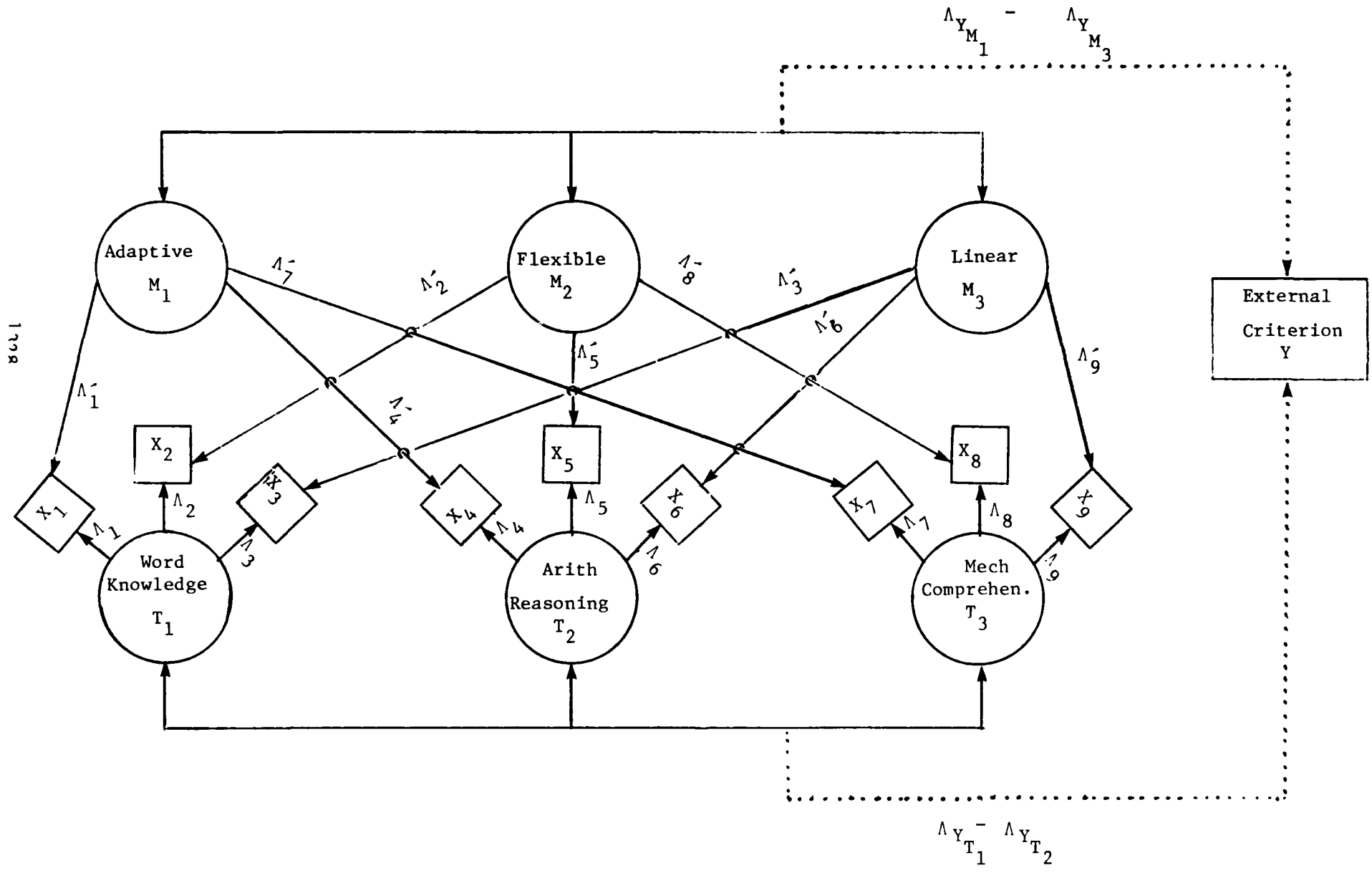


Figure 2

Model I Ho: Method Variance-Covariance is zero

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_4 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} = \begin{bmatrix} \Lambda_1 & 0 & 0 \\ \Lambda_2 & 0 & 0 \\ \Lambda_3 & 0 & 0 \\ 0 & \Lambda_4 & 0 \\ 0 & \Lambda_5 & 0 \\ 0 & \Lambda_6 & 0 \\ 0 & 0 & \Lambda_7 \\ 0 & 0 & \Lambda_8 \\ 0 & 0 & \Lambda_9 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \\ z_8 \\ z_9 \end{bmatrix}$$

and $\phi = \begin{bmatrix} 1 & & \\ \phi_{21} & 1 & \\ \phi_{31} & \phi_{32} & 1 \end{bmatrix}$

$$\underline{x} = \Lambda \underline{f} + \underline{z}$$

Model II Ho:3 Traits and 3-Methods with Traits Orthogonal to Methods

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} = \begin{bmatrix} \Lambda_1 & 0 & 0 & \Lambda_1' & 0 & 0 \\ \Lambda_2 & 0 & 0 & 0 & \Lambda_2' & 0 \\ \Lambda_3 & 0 & 0 & 0 & 0 & \Lambda' \\ 0 & \Lambda_4 & 0 & \Lambda_4' & 0 & 0 \\ 0 & \Lambda_5 & 0 & 0 & \Lambda_5' & 0 \\ 0 & \Lambda_6 & 0 & 0 & 0 & \Lambda' \\ 0 & 0 & \Lambda_7 & \Lambda_7' & 0 & 0 \\ 0 & 0 & \Lambda_8 & 0 & \Lambda_8' & 0 \\ 0 & 0 & \Lambda_9 & 0 & 0 & \Lambda_9' \end{bmatrix} \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ M_1 \\ M_2 \\ M_3 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \\ z_8 \\ z_9 \end{bmatrix}$$

$$\underline{x} = \Lambda \underline{f} + \underline{z}$$

and

$$\phi \begin{bmatrix} 1 & & & & & \\ \phi_{21} & 1 & & & & \\ \phi_{31} & \phi_{32} & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & \phi_{54} & 1 & \\ 0 & 0 & 0 & \phi_{64} & \phi_{65} & 1 \end{bmatrix}$$

Model III Ho same as Model II except traits and Methods free to covary

This construct validity portion of the model can be assessed with the following confirmatory factor analysis model:

$$\underline{x} = \Lambda \underline{f} + \underline{z} \quad (1)$$

$$\Sigma = \Lambda \phi \Lambda' + \psi \quad (2)$$

where $\underline{x}$ is 9 x 1 vector of observed scores based on various combinations of traits and assessment methods, $\underline{f}$ is a vector of hypothesized underlying constructs or factors, $\Lambda$ is a matrix of factor loadings relating $\underline{x}$ to $\underline{f}$, and $\underline{z}$ is a vector of unique scores (i.e., errors of measurement). $\Sigma$ and $\phi$ are variance-covariance matrices of observed variables and factors, respectively and $\psi$ is a diagonal matrix of error variances. Maximum likelihood estimation procedures are available either through Sörbom and Jöreskog's confirmatory factor analysis with model specifiations (1976) or their LISREL IV program (1978).

The sequence of hypothesis to be tested are dictated by the construct validity models presented in Figure 2. The first question has to do with whether or not the three trait factors are sufficient to reproduce the original variance-covariance matrix. The null hypothesis here is that the method variance is essentially zero. There are 45 observed variance-covariances and 21 unknowns to estimate in this particular constrained model leaving 45-21 or 24 df for the large sample maximum likelihood ratio statistical test. If the null hypothesis is not rejected, one would conclude that method variance is not a significant source of confounding variance when interpreting the relationship among the traits. If as expected, the null is rejected then the next step is to test a less constrained model that allows for trait and method factors.

Consistent with the diagram in Figure 1, the matrix equation shown in Model II Figure 2 specifies a model in which each measured variable has a non-zero loading on one trait and one method factor. In this particular test, the traits and methods are allowed to be intercorrelated among themselves but not with each other. If the goodness of fit doesn't significantly improve as shown by a decrease in the $\chi^2$/df ratio or alternately the root mean square residual, then alternative models can be specified depending on the interpretations of the size of specific residuals. Regardless of goodness of fit, one might wish to examine the inprovement, if any, from "freeing up" the trait-method factors cross correlations. It may well be that hardware based methods of assessment might have a positive relationship with performance on measured traits such as mechanical comprehension.

Assuming that Model II or Model III provide reasonably good fits to the data, one can proceed with the usual convergent and divergent criteria for the construct validity of the three traits and methods. However, what is of critical importance here is not so much verifying the discriminant validity of the traits but answering questions such as, is any particular method (e.g. computer adaptive testing) more valid for any one or all traits? This can be tested by constraining the factor loadings within traits to be equal, and inspecting the increment in the $\chi^2$. A second question has to do with whether there is comparatively more method variance for the computer adaptive testing procedure than the remaining two methods? When the solution is standardized the square of the respective factor loadings indicate the proportion of variaince in each measure which is due to trait variance, method variance, and error variance. One might expect based on theory that the computer based adaptive testing might well have the smallest error variance, but may also demonstrate non-trivial amounts of method variance (cf., Bejar & Weiss, 1978).

## Validity Generalization Across Subpopulations

The above precedure can be generalized to estimate and test whether the various relationships between trait-method variances, and covariances are invariant across populations. For example, theory suggests that for less well prepared subpopulations, e.g., high school dropouts, and/or extremely well prepared subjects the computer adaptive testing procedure should yield more accurate results. This could be tested by grouping subjects by amount of formal education and comparing the method loadings within traits from one population to another. Similarly, one could examine evidence on whether or not the traits and method variance relationships are invariant across populations such as race and sex groups. Support for this type of population invariant construct validity is gathered by constraining: (1) the method loadings within traits to be equal across population, and (2) the variance-covar iances among trait factors to be equal across populations and then inspecting the increment in the $\chi^2$. This kind of investigation has been formulated within the test bias or test fairness framework by Rock et al. (1981)

## External Validity

External validity relationships with both traits and method factors can be estimated by taking the "fitted" multi-trait multi-method factor solution and using factor extension procedures to regress the criterion on the factor solution.

Inspection of the resulting pattern of loadings on the trait factors will indicate the validity of the various traits unconfounded by method variance. For example, if the criterion is a paper and pencil test, then one might get an inflated estimate of the validity of the traits if they were only measured by paper and pencil procedures. In short, part of their validity is in a sense due to predictor-criterion contamination due to shared methods.

## SUMMARY

A construct validity paradigm based on maximum likelihood factor analysis is outlined for comparing the validity generalization of selected assessment methods across both traits and populations of interest. The "fitted" construct validity solution can then be extended on the criterion. This procedure allows one to not only arrive at estimates of the predictive validity of the traits free of possible method contamination but also the predictive validity of the methods independent of traits.

References


Alwin, D. F. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (Ed.) Sociological Methodology, 1974, San Francisco: Jossey-Bass, 79-105.

Bejar, I. I., and Weiss, D. J. A construct validation of adaptive achievement testing. (Research Report 78-4) Minneapolis: University of Minnesota, Department of Psychology, 1978.

Campbell, D. T., and Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Gialluca, K. A., and Weiss, D. J. Effects of immediate knowledge of results on achievement test performance and test dimensionality. (Research Report 80-1) Minneapolis: University of Minnesota, Department of Psychology, 1980.

" Joreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971a, 36, 109-133.

Kreitzberg, C. B., Stocking, M., and Swanson, L. Computerized adaptive testing. Computers and Education, 1978, 2, 319-329.

Lord, F. M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-242.

McBride, J. R. Research on adaptive testing, 1973-1976: A review of the literature. A paper submitted to the Department of Psychology in lieu of a special preliminary examination. Minneapolis, Minnesota: University of Minnesota, 1976

McKinley, R. L, and Reckase, M.D. A successful application of latent trait theory to tailored achievement. (Research Report No. 80-1) Columbia: Educational Psychology Department, University of Missouri, 1980.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Rock, D. A., Werts, C., and Grandy, J. Construct validity of the GRE across populations: An empirical confirmatory study. (GRE Report No. 78-1) Educational Testing Service, 1981.

" " Sorbom, D., and Joreskog, K. G. LISREL IV: analysis of linear structural relationships by the method of maximum likelihood. User's Guide, National Educational Resources, Inc., Chicago, Illinois, 1978.

" " Sorbom, D., and Joreskog, K. G. COFAMM: confirmatory factor analysis with model identification. User's guide. Chicago, Illinois: National Educational Resources, Inc., 1976.

Weiss, D. J. The stratified adaptive computerized ability test. (Research Report 73-3) Minneapolis: Department of Psychology, University of Minnesota, 1973.

Werts, C. E., Linn, R. L., and Joreskog, K. G. Estimating the parameters of path models involving unmeasured variables. In H. M. Blalock, Jr. (Ed.), Causal models in the social sciences, Chapter 23, Chicago, Illinois: Aldine-Atherton, 1971, 400-409.

Whitley, S. E., and Dawis, R. V. The influence of text context on item difficulty. Educational and Psychological Measurement, 1976, 36, 329-337.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

Yen, W. M. The extent, causes and importance of context effects on item parameters for the latent trait models. Journal of Educational Measurement, 1980, 17, 297-312.

AD P001398

# Are We Correcting for Guessing in the Wrong Direction?

Howard Wainer
Educational Testing Service
Princeton, N.J. 08541

## Introduction

The development of multiple choice tests had as a concomitant event the increased likelihood of correct guessing. Traditional corrections for guessing typically examine the incorrect responses, and assume that, for say a k choice item, for every k-1 incorrect responses observed there was one correct response that was arrived at through guessing. Thus a so-called "formula score" on a k possibility multiple choice test was the [number correct] MINUS [number incorrect/k-1]. There are many variants on this general scheme (e.g. De Finetti, 1965) but all assume that the aim is to obtain some measure that has incorrect responses as a viable carrier of information. They all, given their assumptions about respondents' behavior, yield an unbiassed estimate of number correct.

Use of these simple techniques for the correction for guessing is unsatisfactory for a variety of reasons. Lord and Novick (1968, p. 303) write,

> "...examinees who have partial information about an item do not respond at random, nor do examinees with misinformation about that item. In these situations, wrong answers cannot be equally attractive to the examinee. No simple correction formula is appropriate in these cases."

They later state (p. 309),

> " The simple knowledge or random guessing model is used extensively despite its several weaknesses. One such weakness is that it ignores the possible day-to-day variations in examinee performance ... which experience has taught us cannot be neglected. An equally serious weakness is that the model assumes that if the examinee is unable to pinpoint the correct response, then he is completely ignorant in this situation and has no basis for choosing among the possible responses. This second assumption can seldom be seriously entertained."

Lord and Novick present a convincing argument against the use of incorrect responses as the sole carrier of information about guessing. The basis of much of their argument is that simple guessing cannot be thought of as the cause of much of what is observed in a test response vector. Since most of these correction methods were working

from a background of true score theory, there was little else that could be done.

With the development of latent trait models (what is now being called "Item Response Theory" or IRT) a new aspect of the test response vector became available to aid in the estimation of the testee's ability--the pattern of the response vector. Through the use of IRT the items could be ordered by their likelihood of being answered correctly, and unusual responses could be noted. Thus an incorrect response to a very difficult item could be seen as a predictable event, and tended to confirm the fit of the model. A correct response to an easy item was also anticipated, and further corroborated the model. Unusual events were difficult items answered correctly (evidence of guessing) or easy items answered incorrectly (evidence of "sleeping") [1]. In practice guessing occurred much more frequently than sleeping. The strategy to correct for guessing was to either eliminate items with large residuals from the IRT model, or to change their response to conform to expectation. The former was the more frequent choice, but both have the same outcome; the reduction of the ability estimate.

Note that the introduction of a non-zero lower asymptote to the item characteristic curve implies that a person can get an item correct without knowing the answer. This has as a consequence a reduction in ability estimates. Further, this method for dealing with guessing causes very substantial increases in the standard errors of the estimates of item difficulty and should be avoided if possible (see Thissen and Wainer, 1981).

The principal difference between the traditional corrections for guessing and the IRT correction is that the former focuses on the number of incorrect responses, the latter principally on the unlikely correct ones. In either case the result is in general the same -- the lowering of the score of the testee. It is clear that guessing inflates the observed score of the testee, but is lowering the score the best way to go about correcting it?

Obviously if we are sure that the unusual response pattern was generated by guessing the correct fix is to reduce the ability estimate appropriately. The problem is that we can never be sure that it was guessing that yielded the observed unusual response pattern. Alternative explanations are possible, and sometimes even likely. One possible alternative explanation is that the trait being measured, while unidimensional for most of the tested population, is multidimensional for some small but identifiable sub-population. For example, in a test of drug knowledge, one student indicated that he knew very little about drugs, but he got three rather difficult questions correct. The first impulse was to ascribe these correct

responses to guessing. Upon a closer look it was discovered that the questions answered correctly dealt with mescaline and peyote, and the student turned out to be a native Indian for whom these kinds of drugs were commonly used in religious ceremonies. Clearly if we reduced this student's ability estimate based upon the low probability of his knowing the correct answers to these questions (the probabilities coming from our test model) we would have seriously underestimated his drug knowledge. It is the problems associated with correcting for guessing when the responses observed were not generated by the sort of Bernoulli trials posited by most correction schemes that is the subject of this paper.

## The Start of an Alternative

Before we discuss alternative schemes for dealing with unlikely responses let us restate what we want out of a test. Usually what is desired is the best estimate of the testee's ability we can muster and an honest measure of that estimate's accuracy. This is usually concretized by an ability parameter and the standard error of that parameter. True score theory yields standard errors through various sorts of reliability coefficients (Lord and Novick, 1968, p. 154ff). Typically the shortening of the effective test length that is the result of guessing is not fully represented in the error estimate. Using the three parameter latent trait model (see for example Birnbaum, p. 453ff, in Lord & Novick, 1968) the standard error of the ability parameter is affected by the guessing parameter (see Jones, 1981, for details of how to measure the size of this effect). Furthermore, there have been very few applications in which the guessing parameter of the item was successfully estimated. This is partially due to the problem of trying to estimate a parameter in a region where there is little or no information.

Consider the following scenarios in which a test, scored from, say, zero to 100 is administered and the mean ability turned out to be 50 with a standard deviation of 10. Suppose further that we are forced to provide an estimate of each individual's ability:

1) Student A does not show up for the test.

2) Student B shows up but hands in a blank answer sheet.

3) Student C hands in an answer sheet on which he got 20% of the items correct, and these were all among the 25% easiest items.

4) Student D also got 20% of the items correct, and they were scattered randomly throughout the exam.

5) Student E got the 20% most difficult items correct and all others wrong.

How should these various situations be treated? Student A, most would agree, should be recorded as 'missing data', and if we must assign him a score we should somehow impute one based upon the rest of the sample. Certainly we should not assign a zero ability estimate, for his absence from the test could have been caused by a wide variety of circumstances unrelated to his ability (i.e. illness, transportation failure, family circumstance, etc..). Thus a sensible way to treat such an individual (aside from saying "did not show up --- no score assigned") would be to assign the mean, and to use the standard deviation of the tested sample as the standard error of this estimate. This is in keeping with such correction schemes as Kelley's equation

$$\tau = \rho\ x + (1-\rho)\ \mu$$

(1947, p. 409) in which estimated true score ($\tau$) is obtained by regressing observed score (x) toward the mean ($\mu$); the extent of the regression effect is the reliability ($\rho$) of the test. In the current situation the "test" is very unreliable and so assigning a score of 50 with a s.e. of 10 would not be untoward.

The second scenario is not very different from the first. Once again the answer sheet is blank, the only difference being that the testee showed up at the testing center. Nevertheless, we can imagine many plausible reasons why the test was handed in blank which are unrelated to ability. Once again we are probably underestimating the testee if we assign a zero score, and some sort of estimate based upon an imputation scheme would surely serve better.

Note that in both of these situations the correction that is being used <u>boosted</u> rather than lowered the examinee's observed score. This contrasts sharply with traditional corrections.

Moving on to the third scenario we find that the response pattern fits our expectations of how a person ought to take the test, and so our confidence in the traditionally estimated ability estimate is well described by the estimated standard error. Note that in this situation we have a pattern that is quite likely, and so even though the total score is about what would be expected by chance the pattern of responses argues against it.

The fourth situation shows a pattern that might have been generated by guessing, but it might also have been generated by a pattern of knowledge different from what we anticipated. Thus we might lower the ability estimate following traditional practice, or we might instead conclude

that we have not measured this person very well (his pattern does not fit the test) and so we regress his ability estimate in the direction of the mean and expand the standard error. When the standard error gets to 10 we can conclude that we know as much about this person after viewing his test response vector as we would have, had he not taken the test. Of course experience has taught us that individuals who hand in answer sheets like this come from a population whose mean is considerably lower than that of the general population, and so we might end up regressing the score downward.

The last situation is the most difficult to explain. Most explanations that suggest themselves indicate a much higher ability. One is reminded of the test results of Galois in his two entrance examinations for the Ecole Polytechnique. Perhaps such a pattern was generated by a very bright student who was bored by the exam and only started to answer questions when they became interesting. Certainly we cannot conclude that this student was guessing, and we would be far better off concluding that we have not measured him very well and impute a more middle score and a large standard error.

The preceding five scenarios represent in extreme what all test patterns show to some extent or other; a pattern that does not conform exactly to what we, the test maker, expected. It is my contention that we are more honest in scoring if, when we find a test pattern that is unlikely to have been generated by someone behaving as we hypothesized we should reflect our uncertainty in their ability by regressing their score toward the middle and by suitably increasing the standard error of our estimate. This will avoid the uncomfortable position of having to explain to an irate parent that the reason their child received a lowered score was because some of the items he got right were too hard for him. Rather we can say that his score is higher than the number right would indicate because he responded in such a haphazard way that we don't know what to make of it. The strategy of course is to use more heroic measures(e.g. retesting) to obtain ability estimates for those for whom the standard errors of estimate are unacceptably large.

One can think of this process in a Bayesian way, by having a prior ability distribution generate a posterior score; the observed test score with an associated fit statistic being the data that pull the prior toward the posterior. If the fit is poor the data have little effect. If the fit is good the data have a more profound effect. The fit statistic can also used to inflate the standard error of the estimate (see Finney, 1952). Further details on one way to accomplish this are in the next section.

## A More Specific Plan

The foregoing discussion was more epistemological than operational. We pointed out the problems associated with assuming that wrong answers or unusual response patterns were the result of guessing and not, say, idiosyncratic knowledge. The schemes proposed for correcting this problem were general, and few details were elucidated, although some may see some similarities between the use of Kelley's equation and the more general James-Stein estimators (Efron and Morris, 1975).

In this discussion we have purposely gone too far. Certainly cogent arguments can be made for regressing the ability estimates obtained from unusual response patterns inward from their observed values. We have tried to elucidate some of them here. Yet experience has taught us that guessing is an activity that occurs frequently, and accounts for much (though by no means all) of the anomalous responses observed. Therefore we would like to propose a middle way. It would seem that a proper scheme to correct for unusual response patterns would be in-between the traditional correction (that assumes that all anomalies are from guessing) and one that professes complete ignorance of the cause of such unusual response patterns. The method that we prefer regresses the 'corrected' ability estimates inward, although not usually as far as using an uncorrected raw score.

Recently Wainer and Wright (1980), proposed an amalgam of the Jackknife and the Sine M-Estimator (the so-called AMJACK estimation scheme) to robustly estimate ability in the Rasch Model. The aim of this procedure was to obtain an estimate of ability from each item responded to, and to robustly estimate the middle of this distribution. The notion was that items that were responded to in a non-stereotypic manner would look like outliers and would be largely discounted in their contribution toward the ability estimate. The result of this was that ability estimates tended to be regressed inward toward the center of the ability continuum from their position after the traditional correction. The extent of this regression was dependent upon the fit of the response vector to the Rasch model: the better the fit the less the regression effect. Using this scheme we found that the estimates obtained were better in many ways than the maximum likelihood estimators of ability. In particular, even when the data fit the model the AMJACK estimator was more efficient than the MLE for short to modest length tests (test lengths of less than 40 or so items); as tests grew longer the asymptotic optimality of MLE began to manifest itself. When the data became noisier the superiority of the AMJACK estimator over MLE grew larger.

In addition to the increase in efficiency and accuracy of the AMJACK estimator, we found that the standard error of the estimator, calculated directly from the Jackknifed Pseudovalues, increased as the deviation from the Rasch model increased. This once again is in line with the general guidelines we described above for a sensible test scoring scheme, although the actual estimates of standard error tended to be too conservative.

## An AMJACK Example

To better understand how the AMJACK estimator works let us consider a simple numerical example. Suppose we give a ten item test whose difficulties span the range from -2 to +2 logits uniformly, and we observe (among the many patterns of responses) several shown below:

Ability Estimates on a short test (with some noise) by five methods

| Patterns | MLE Uncorrected θ | MLE Uncorrected S.E. | MLE Corrected $\theta_{C_1}$ | MLE Corrected $\theta_{C_2}$ | AMJACK θ | AMJACK S.E. | BIWEIGHT θ | BIWEIGHT S.E. |
|---|---|---|---|---|---|---|---|---|
| (Some with guessing) | | | | | | | | |
| 1) 1110000000 | -1.15 | .79 | -1.15 | -1.15 | -1.05 | .64 | -1.07 | .78 |
| 2) 1110000001 | -.56 | .76 | -1.12 | -1.15 | -.86 | .79 | -.91 | .77 |
| 3) 1110000010 | -.56 | .76 | -1.11 | -1.15 | -.82 | .78 | -.81 | .77 |
| 4) 1110000100 | -.56 | .76 | -1.09 | -1.15 | -.70 | .76 | -.69 | .76 |
| 5) 1110001000 | -.56 | .76 | -1.07 | -1.15 | -.60 | .74 | -.61 | .76 |
| 6) 1110010000 | -.56 | .76 | -1.01 | -1.15 | -.54 | .70 | -.55 | .75 |
| 7) 1110100000 | -.56 | .76 | -.56 | -.56 | -.49 | .64 | -.52 | .75 |
| 8) 1111000000 | -.56 | .76 | -.56 | -.56 | -.52 | .56 | -.53 | .75 |
| 9) 1111110000 | .56 | .76 | .56 | .56 | .52 | .56 | .49 | .75 |
| 10) 1111110001 | 1.15 | .79 | .68 | .56 | .66 | .83 | .95 | .78 |
| 11) 1111110010 | 1.15 | .79 | .73 | .56 | .80 | .79 | 1.00 | .78 |
| 12) 1111110100 | 1.15 | .79 | 1.15 | 1.15 | 1.02 | .71 | 1.02 | .78 |
| 13) 1111111000 | 1.15 | .79 | 1.15 | 1.15 | 1.05 | .64 | 1.01 | .78 |
| (Some with sleeping) | | | | | | | | |
| 14) 0110000000 | -1.84 | .89 | -1.37 | -1.15 | -1.20 | 1.04 | -1.66 | .85 |
| 15) 0111000000 | -1.15 | .79 | -.68 | -.56 | .66 | .83 | -.96 | .78 |
| 16) 0111110000 | 0.00 | .74 | .51 | .56 | .34 | .76 | .27 | .75 |
| 17) 0111111000 | .56 | .76 | 1.12 | 1.15 | .86 | .79 | .90 | .77 |
| 18) 0000011111 | 0.00 | .74 | ? | ? | 0.00 | 1.36 | 0.00 | .74 |
| The difficulties are | -2.0 | -1.6 | -1.1 | -0.7 | -0.2 | 0.2 | 0.7 | 1.1 | 1.6 | 2.0 |

The estimators shown reflect different philosophies of correction. The first column of 'uncorrected $\theta$' is the standard maximum likelihood estimate of ability obtained from the Rasch model. As is obvious, the pattern is unrelated to either the ability estimate or its standard error. This reflects the well-known property of the Rasch model of raw score being a sufficient statistic for ability. Ordinarily one detects unusual patterns in this model by examining goodness-of-fit of the model to each person rather than the standard error.

The two columns labeled 'corrected' are two different methods of correcting for guessing that are representative of two classes of corrections. $\theta_{c1}$ examines the probability (under the model) of each response. If a response is too unlikely to have occurred under the model that item is omitted from consideration in the determination of the person's ability estimate. This reduces the test length and so increases the standard error. The gradual increase in ability that is evident in response patterns 1 through 7 is due to the average difficulty of the test getting progressively greater as less and less difficult items are omitted. The column marked $\theta_{c2}$ uses the same determination rule as $\theta_{c1}$ except instead of omitting the item it <u>changes</u> the response so that the pattern fits the model. Thus it makes response patterns 2-6 identical to pattern 1, and similarly changes pattern 14 to be like pattern 1.

The last two estimators are the AMJACK (mentioned previously) and Mislevy and Bock's (1982) Biweight estimator. This latter scheme is a robust measure that uses the adaptive biweight in the likelihood equation. Its behavior has not been thoroughly tested, but illustrative examples like that in table 1, indicate that it (or methods that develop from it) may be worthwhile to pursue.

A careful examination of Table 1 will provide for the reader a feeling for how these corrections work, and indicate that the two robust methods shown offer a middle ground. They are less extreme than either of the two corrections, yet not so forgiving as the uncorrected estimator. We believe that the AMJACK estimator reacts to unusual response patterns in a way that closely mirrors what a thoughtful test scorer would do [discount an unlikely event as bizarre and provide a more or less identical ability estimate (with a boosted standard error)]. The Biweight estimator seems to follow this behavior very closely as well, <u>except</u> that the standard error does not seem to expand apace. This latter effect is most visible in the last case (18) in which only AMJACK indicates the unusual nature of this response pattern (by giving an almost

doubled standard error).

We believe that the AMJACK and the Biweight estimators are pretty good first choices to correct for unusual response patterns, in that they move in the correct direction. We prefer the AMJACK because it expands the standard error in a way that mirrors our anticipation. Yet even the AMJACK is just a first step. It may be that a full-blown Bayesian approach that determines the amount of regression effect from more direct evidence of guessing and sleeping will be better still. Yet, such data are hard to find, and thus for many applications AMJACK may do as well as we can expect without extensive empirical prior distributions at our disposal.

Under what circumstances can this be used?

Whether or not a robustification scheme like AMJACK (or some improved Bayesian method) can be successfully applied in paper and pencil tests is uncertain. There are political reasons why having a complicated scoring method sitting behind a nonlinear test model may not be easy to implement widely at the current time. There is, however, one avenue of application that fairly screams for it -- Computerized Adaptive Testing.

To better understand the problem area, and how AMJACK-like techniques can be helpful let us consider the task in CAT. As an individual answers questions the computer algorithm must determine new items to ask. There are several item selection procedures that have been suggested (i.e. choose the available item that maximizes information, or one that reduces the posterior variance). All of these are (or at least should be) in practice, very similar to one another. The reason that this is the case has to do with the structure of a well-developed item pool. More specifically, the slopes of item characteristic curves of items which find themselves in the pool will be nearly equal. This is because items whose slopes are low or negative are poor items and will have been eliminated from consideration in initial pre-testing and item screening. Items whose slopes are very steep are very rare--indeed finding one almost always signals an artifact. Thus equal slopes tend to be the rule (note also that modest differences in slopes -- all other things being equal -- have virtually no effect on information or posterior variance).

Second, if the CAT is working properly we will observe data on inappropriately administered items even more rarely than is currently the case with paper and pencil tests. Thus the estimation of a non-zero lower asymptote will be nigh onto impossible (there is no 'c' in CAT). This implies that with information gathered in any reasonably efficient CAT system we will be unable to tell the difference between an

item with a non-zero lower asymptote from one with a zero asymptote.

These two characteristics lead us to the conclusion that the single driving characteristic of an item that will determine its selection within a CAT mode is its difficulty. Thus regardless of the item selection method used, the item difficulty will be the thing we are most interested in. To choose a reasonable next item we merely have to match, as closely as possible, the difficulty to the current estimate of the examinee's ability. It is here where these new ability estimators come into their own. After each item it gives a good estimate of ability and an honest estimate of its standard error. The former aids in item selection the latter in determining appropriate stopping.

AMJACK has only been tested (so far) within the context of the Rasch model. Certainly there are circumstances where this simple model will not fit. Yet within CAT it would appear that the chances of the Rasch model fitting are very good indeed. Further, the robust character of the ability estimation scheme gives us protection from even substantial departures from fit. Moreover it characterizes poor fit with big standard errors. Obviously AMJACK isn't the end, it merely moves a little in the right direction. More work is needed.

## Some Concluding Remarks

In this paper we have concerned ourselves with the problem of how to treat unusual responses in an individual test pattern. These are most often caused by what has been referred to previously as guessing. Traditional corrections for these unusual points yield reduced ability estimates. We have argued that there are enough alternative explanations for unusual responses to indicate that a useful strategy might be to allow that the unusual response pattern tells us that the person taking the test did not respond the way we expected, due either to his fault (he guessed) or to ours (the test was not unidimensional). It is certain that we do not know as much about him as we would have had he responded in a manner described by the test model. This increase in our ignorance is better admitted by regressing the ability estimate that has been corrected in a traditional way inward and increasing the standard error. We suggest that one way to use such scores would be to determine the accuracy required for the purpose at hand and to set aside all individuals who were not measured with sufficient accuracy (scan standard errors) and use the ability estimates of the rest. Those who were not measured with sufficient accuracy would then be subjected to different treatment (another test, a different kind of test, etc.). It is clear that we do far less damage to individuals

for whom the test is inappropriate by stating such than by reducing their score. This is particularly true now with the necessity of making public test vectors and scoring keys. The explanation that an ability estimate of a student was reduced because the items responded to correctly were difficult is an explanation that will cause more problems than it solves. The alternative offered is just a first approximation, but one that moves the correction for guessing in what we consider to be a more appropriate direction. It is, in prospect, a fruitful approach to use within the context of CAT.

## Afterthought

The problems of unusual responses discussed in this paper can be diminished in other, non-statistical, ways. Most obviously, we could not use the standard format multiple choice test. Gulliksen (personal communication) has often suggested that instead of having several choices next to each item one could just have a long list of choices at the beginning of each test section. With a long enough list the probability of choosing the correct response by chance is small enough not to yield serious difficulties. He also mentioned that this would make the task of item writing much easier, since one merely had to think up questions with their answers, and did not have to perform the very difficult task of thinking up reasonable wrong answers.

Another suggestion (Wright and Stone, 1979, p. 188) is not to present items to individuals that are too far away from their ability. In particular, if the probability of their responding correctly to an item is less than chance do not present that item. This does not cause much of a loss in the accuracy of ability estimation, since items that far from ability in their difficulty carry very little information anyway. The problem with implementing this suggestion in general conditions for wide range tests is that we do not know an examinee's ability a priori. With tailored testing this becomes more feasible. A problem with this approach is that one does not get a check on the fit of the model if no items are offered that would allow an unusual response. This strategy is reminiscent of the wicked witch's instructions in The Wiz, "Don't you bring me no bad news!" Of course a middle path is to ask enough items that are a little extreme so that we can estimate model fit, but not so many to compromise the efficiency and integrity of the test. The trade-off between efficiency/redundancy and the testing of model fit cannot be gotten away from entirely -- only minimized. The extent of redundancy used should reflect our faith/knowledge of match between the test we are using and the model that scores it.

Looked upon as several parts of a single solution it seems to me that each of these sorts of suggestions can be

implemented so as to better accomplish the specific aim of more precise estimation of ability, and a more honest measure of the accuracy of that estimate.

## References


De Finetti, B. (1965) Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical, 18, 87-123.

Efron, B. and Morris, C. (1975) Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association, 70, 311-319.

Finney, D. J. (1952) Probit Analysis. London: Cambridge University Press.

Jones, D. H. (1981) Precision of the ability estimator in the absence of exact item parameters. Unpublished manuscript, Princeton, N.J.

Kelley, T. L. (1947) Fundamentals of Statistics. Cambridge: Harvard University Press.

Lord, F. M. and Novick, M. L. (1968) Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley.

Mislevy, R. and Bock, R. D. (1982) Biweight estimates of latent ability Educational and Psychological Measurement (in press).

Thissen, D. and Wainer, H. (1981) Some standard errors in item response theory. Submitted for publication.

Wainer, H. and Wright, B. D. (1980) Robust Estimation of ability in the Rasch Model. Psychometrika, 45, 373-391.

Wright, B.D. and Stone, M. H. (1979) Best Test Design. Chicago: MESA Press.

## Footnote

[1] Since events that we are calling "guessing" or "sleeping" could have a wide variety of other causes -- such as unusual patterns of knowledge -- this may be a poor choice of terminology. Nevertheless we will continue to use these terms, as a convenient notation, to describe these two kinds of unusual events.



AD P001399

# LEGAL AND POLITICAL CONSIDERATIONS IN LARGE-SCALE ADAPTIVE TESTING [1/]

by

Brian K. Waters

Gus C. Lee

Human Resources Research Organization

As the term implies, adaptive testing is defined as a method of test construction wherein the items presented to a subject are selected iteratively dependent upon previous responses, thus "adapting" the test to the subject. Theoretically, such individual testing should provide more accurate measurement than group testing. Both simulated and live data studies have reported that the proportion of items required to reach a given level of reliability in a computer administered adaptive test (CAT) are about one-half to three-fourths of those required by a conventional, paper and pencil test. Such dramatic efficiencies occur because after each item response, the computer program selects the next test item from the item pool which will provide the maximum amount of information about the examinee. McBride (1979) provides an excellent review of the advantages and possible disadvantages of CAT. There is little doubt that the use of interactive computer testing will increase enormously in the coming decade.

CAT is a technology preparing to make the transition from the laboratory to an operational environment. The vast majority of research and development in CAT has been sponsored by the Military Services, particularly the Navy, since Lord's early work on item response theory and CAT during the 1960s. Today, the Department of Defense (DoD) is sponsoring a large-scale, multi-year, project to develop a CAT system for implementation in Armed Forces Examining and Entrance stations across the country.

DoD assigned the Department of Navy the responsibility for CAT development, with the Marine Corps as the executive agent. The Naval Personnel Research and Development Center (NPRDC), is currently in the process of selecting contractors to design, develop, and try out a prototype CAT system for delivering the Armed Services Vocational

---

[1/] A paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D.C., October 28, 1981.

Aptitude Battery (ASVAB) adaptively to military applicants at nearly 1000 testing sites. Also, our host, ARI, has just released a "request-for-proposal" for a seven year Army selection and classification system. CAT would likely be a part of such a system. Outside of DoD, the Coast Guard is sponsoring field research into CAT, and the commercial testing industry is investigating the large-scale use of CAT. Although no commercial tests have actually begun using CAT delivery, the American College Testing Program (ACT) and Educational Testing Service (ETS) are working toward that goal. Carol Dwyer's paper given yesterday afternoon at this conference mentioned that ETS is currently considering development of two-level sequential tests for a major admissions program.

Thus, the time has arrived when we must start considering some of the "real world" issues which CAT will face as actual decisions about examinees are made using adaptive testing. One such set of issues involves the legal and political considerations which could arise as CAT becomes an operational reality. This paper will discuss the legal and political implications of CAT. The authors will pose questions, not answers--questions that need to be seriously considered as large-scale CAT approaches implementation.

To date, very little, if anything has been published on legal and political considerations of CAT. Warm (1978, p. 122) questioned the legal defensibility of having examinees take different numbers of items, or different sets of items. Also, perhaps the examinee who answered the higher percentage of items correctly received the lower test score. Wiskoff (1979) and Waters (1979) called for CAT researcher attention to such legal issues surrounding CAT.

CAT is a subset of testing in general. And, as I'm sure you all are well aware, testing has come under extreme pressure in the political and legal environments during the past decade. Strong political lobbies (Ralph Nader, 1979) have directed stinging criticism of commercial testing programs, and testing has become a frequent subject in litigation.

An annotated bibliography of court cases relevant to employment decisions (Cascio & Bernadin, 1981) was published by the Air Force Human Resources Laboratory (AFHRL). This very useful document, completed under AFHRL contract by McFann-Gray & Associates, reviews 232 court cases from January 1971 - January 1980 dealing with adverse impact, unequal opportunity or pay, and bias in personnel selection, classification and evaluation systems. Each annotation provides the case reference, case source, court decision, critical cases cited as a basis for the decision, evidence of adverse impact, evidence of job-relatedness or validity, type of selection procedure, factors impacting the decision, effects of expert testimony, and implications

for personnel policy. Not surprisingly, the authors of this paper found that at least 60 of these cases directly involved testing as a central issue. Overwhelmingly, the major focus of legal attention in these 60 cases was test validity, and the clear conclusion to be drawn was that job-related empirical validity was preferred by the courts. In fact, a hierarchy of "acceptability" of test validity methodologies was evident. Only empirical validity was accepted unless practical constraints made empirical validation studies unrealistic. When empirical validation was impossible for a given test application, then content validation based upon careful job/task analysis was credible. Third in the list of court-accepted validation methods was construct validity. Finally, face validity was given virtually no weight in the reviewed cases.

In a computerized adaptive testing mode, the validity issue portends possible problems in court. In CAT, it is not practical to validate the "test", only the item pool from which each test is drawn. This "validation" is completely different from empirical validations that the courts have previously accepted. Will this kind of predictive validity be satisfactory in a legal battle? Will the courts accept content validity when a very large number of items are said to measure a single trait? Rock & Bejar (1981) suggest that, construct validity is more defensible in CAT, but the courts have been hesitant to accept (or likely to understand) construct validation under conventional testing.

The latter point, that judges and juries may not understand many of the complicated technicalities involved with CAT, promises to be a major hurdle to CAT implementation. How do you explain latent trait theory, esoteric item selection strategies, "occult scoring methods" (Lord, 1978), and myriad other inexplicable jargon surrounding CAT to a court? Will a jurist accept the expert witness psychometrician who testifies that person A's theta is higher than person B's even though they have taken no items in common; that person B answered a higher percentage of the items correctly; and that person B had to take twice as many items as person A to estimate his theta as accurately? Clearly the CAT community has a major educational and public relations chore in the court room and through the media before CAT will be accepted as a valid measurement procedure.

One example of such a problem in the political arena occurred with the ASVAB, where test scores are reported to Congress. A serious calibration error occurred on ASVAB in 1976, and the psychometrician's credibility with the Congress suffered. It has been said that Congress feels that DoD psychometricians are some kind of amateur magicians out to perform statistical legerdomain. Imagine how they will react to latent trait theory?

The credibility of CAT test scores may also be weakened by calibration problems like Douglass (1981) discussed. Parameter estimation methods and item linking procedures (Dorans, 1981) similarly are statistically complex activities which would be very difficult to explain in court or in Congress. We need to translate the scientific jargon that has become part of the CAT vocabulary into clear, concise, and comprehendible language that will communicate to the layman in testing.

Another related issue is the recent truth-in-testing movement. After an initial strongly negative reaction by the testing industry to legislative directives on the release of test items to examinees, the major commercial test companies seem to have reluctantly accepted the concept. Computerized adaptive testing offers the prospect of making truth-in-testing more palatable to test developers. The release of items from a large pool to an examinee would not likely damage future administrations due to test compromise since theoretically every examinee takes an individually tailored test. Under current test development procedures, the new legislation will likely lead to expanded requirements for test items, additional validation studies, and increased test fees for examinees.

Theoretically, CAT and item response theory should reduce cultural test bias (Pine, et al, 1979). The Cascio and Bernadin (1981) review cited many court cases in which alleged cultural test bias was a major issue, and numerous symposia, addresses, and paper sessions were presented at the APA Convention two months ago on the same subject. Certainly it would be very beneficial to CAT's credibility in the courts if clear, unequivocal evidence should evolve which showed reduced cultural test bias under adaptive testing. At this time, this remains a research question.

This paper has strived to do no more than simply suggest questions that need to be considered as CAT nears operational usage. One thing we can be absolutely sure of is that once personnel selection and classification decisions begin to be made using CAT, there will be legal challenges to the validity of the measurement process. We need to anticipate these challenges, to conduct the research to answer the legal questions, and to understand enough about legal processes and judgments to "sell" the benefits of CAT to the courts and the public.

## REFERENCES


Cascio, W. F., and Bernardin, H. J. Court Cases Relevant to Employment Decisions: Annotated Bibliography, ARHRL-TR-80-44. Brooks AFB, TX: Air Force Human Resources Laboratory, February 1981.

Dorans, N. J. Why and How to Insure that Items in the Same Pool are Appropriately Credentialed: Data Collection Strategies for Item Linking. Paper presented at the 23rd Military Testing Association Conference, Arlington, VA: October 1981.

Douglass, J. B., and McColskey, W. "A Comparison of Item Response Theory Item Calibration Techniques." Paper presented at the 23rd Military Testing Association Conference, Arlington, VA: October 1981.

McBride, J. R. Adaptive Mental Testing: The State of the Art. Technical Report 423. Washington, D.C.: US Army Research Institute for the Behavioral and Social Science, November 1979.

Nader, R. The Reign of ETS: The Corporation that Makes Up Minds. P. O. Box 19312, Washington, D.C. 20036.

Pine, S. M., Church, A. T., Gialluca, K. A., and Weiss, D. J. Effects of Computerized Adaptive Testing on Black and White Students. Research Report 79-2, Minneapolis: University of Minnesota, March 1979.

Pine, S. M., and Weiss, D. J. Bias-Free Computerized Testing: Final Report. Minneapolis: University of Minnesota, March 1979.

Rock, D. A., and Bejar, I. . "Validity Considerations for Adaptive Testing Systems". Paper presented to the 23rd Annual Military Testing Association Conference, Arlington, VA: October 1981.

Warm, T. A. A Primer of Item Response Theory. Technical Report US-CG-941278, Oklahoma City, OK: U. S. Coast Guard Institute, 1978.

Waters, B. K. Discussant Remarks: Computerized Adaptive Testing Symposium. American Psychological Association Annual Meeting, New York: September 1979.

Wiskoff, M. Symposium Chair Remarks: Computerized Adaptive Testing Symposium. American Psychological Association Annual Meeting, New York: September 1979.

Panel: Skill Qualification Testing - An Evolving System

Chair: Stanley F. Bolin, US Army Research Institute, Alexandria

Robert Eastman, James Sova, US Army Training Support Center, Ft. Eustis, VA; Joan Harman, John Kessler, Douglas Macpherson, US Army Research Institute, Alexandria, VA.

Panel member presentations described research efforts directed toward specific issues in Skill Qualification Testing. Dr. Kessler documented the value of fast SQT feedback as perceived by troops and noted that troops tended to be increasingly passive about remedial training, expecting their units to do such training rather than taking personal initiative, when feedback was delayed beyond two weeks. Mr. Macpherson demonstrated that sergeants underestimate written test item difficulty in a way that could lead to unexpected SQT task failures. Dr. Harman showed that soldiers more often attributed high failure rates to unit duty relevance as contrasted with either training or testing factors. Dr. Eastman demonstrated that SQT scores for helicopter mechanics were valid against supervisory ratings of performance and that the current cut score was much too high to reflect this validity. Dr. Eastman also reported for Mr. Sova on recent study of ten SQTs; overall, this study showed that (1) SQT drives training in combat arms more often than it does in noncombat fields and that SQT scores increase with training, (2) SQT discriminates between good and poor performers without requiring more reading skill than soldiers possess, (3) SQT results are not correlated with troop motivation or attitudes toward the Army.

Mr. Ray Carroll of the General Accounting Office made an unprogrammed report on the GAO's recent evaluation of SQT. GAO recommended that the present be scrapped for poor return on the very large costs in development and maintenance of the system. GAO suggests that lower ranking soldiers be trained and tested by their sergeants using Soldier's Manuals and SQT test materials as necessary for unit training but without centralized management and upward reporting of score information. For the sergeants, GAO suggests written testing for promotion purposes. Mr. Carroll noted that these views and suggestions are from a draft GAO report which will go to the Army for formal comment in the near future.

Dr. Bolin concluded from panel discussion that the SQT system will likely evolve toward two systems using criterion referenced testing concepts for two different target populations, with training as the primary goal for the lower grades and promotion as the primary goal for the higher grades.

AD P001400

# SUPERVISOR RATINGS AS CRITERIA FOR SKILL QUALIFICATION TESTS

Robert F. Eastman

U.S. ARMY TRAINING SUPPORT CENTER

Supervisor ratings of overall job performance were correlated with the SQT scores of MOS 67N, utility helicopter mechanics. A high positive correlation ($r = .74$) was found for lower skill level (SL 1) soldiers. The correlation was significantly lower for SL 2 ($r = .64$) and SL 3 ($r = .35$) soldiers. The high positive correlation at SL 1 indicates that the SQT is a valid instrument for discriminating between MOS performers and non-performers in selected MOS. The valid supervisor ratings obtained were used as criteria to determine where the optimum cutscore for the SL 1 would fall if it was based on ratings of overall job performance. The cutscore which optimally discriminated between performers and non-performers, fell about 15 points lower than the current cutscore of 60. The findings confirmed that it is feasible to establish SQT cutscores empirically for MOS where valid supervisor ratings are obtained.

The SQT is designed to evaluat soldier performance at the task level and supervisor ratings are used during Skill Qualification Test (SQT) field validation tryouts to identify soldiers who are performers or non-performers on specific tasks. Test items are then screened and standards for task test units are established based on how well performers do and how well the test items discriminate between performers and nonperformers (TRADOC pamphlet, 351-2, 1980). However, because total SQT scores are used to make decisions about promotion and reenlistment eligibility, the extent to which they are related to overall job performance is an important issue. If SQT scores assess the overall job proficiency of soldiers, they should be positively correlated with supervisors' ratings. A related question is whether or not supervisors' ratings can and should be used to establish SQT cutscores. Currently, passing 60 percent of test units is the score required to "verify" that a soldier meets the minimum standards. At the task level, one can argue that subject matter experts should be able to agree what constitutes successful task performance. Nevertheless, the current policy provides for establishing test unit (task) standards from the actual performance of soldiers. It would seem much more difficult to set a specific cutscore on a test covering many tasks to be used to make personnel decisions about an individual's total job competence, yet that percentage (60 percent) is arbitrarily set. If supervisor ratings of overall job performance are correlated with SQT scores it may be feasible and desirable to use supervisor ratings as criteria for empirically establishing cutscores for specific SQT (Berk, 1980; Davis, 1980).

The purposes of this research are: to correlate supervisors' ratings of overall job performance with the SQT scores later obtained by helicopter mechanics; to determine the cutscore which optimally discriminates between performers and nonperformers and compare it with the 60 percent cutscore.

## METHOD

### Analysis Approach

1. Identify soldiers with an MOS scheduled for testing in about one month.

2. Obtain supervisor ratings of the soldiers in selected units.

3. Determine if the supervisor ratings are reliable.

4. Correlate supervisor ratings with soldiers' SQT scores.

5. Determine if the ratings are valid and if soldiers rated as performers score significantly higher on the SQT than nonperformers.

6. Determine the cutscores for the SQT which best agree with the ratings of supervisors in terms of correctly classifying performers vs. nonperformers.

7. Evaluate the impact on the difference between the empirically determined optimum cutscore and the current 60 percent standard.

## Subjects

Utility helicopter mechanics (MOS 67N) at Fort Benning, Georgia were selected as the research sample because it was feasible to collect supervisor rating data on them one month before they were scheduled for testing. Ratings were obtained on 80 soldiers in the 121st Aviation Company and the 498th Medical Company. Fifty-two were E2-E4, skill level one (SL 1); 17 were E5, skill level two (SL 2); and 11 were E6, skill level three (SL 3). Ratings were obtained from six second and third level supervisors (E6-03) in each of the two units.

## Date Collection Procedure

1. Rosters listing soldiers scheduled to take the SQT were obtained before visiting the two units. The rosters were obtained from the Training Standards Officer at Fort Benning, not from the units.

2. NCO supervisors and one unit commander were individually instructed about the purposes of the ratings and were asked to rate soldiers on the SQT roster in terms of overall job performance and soldier skills. They were explicitly instructed that they were not to base these ratings on any knowledge of the SQT scores these soldiers made last time they took the test. They were told that the purpose was to test the SQT and that it was critical for them to be as objective as possible. The instructions were handed to each rater and any questions they had were answered before they rated the soldiers. The instructions to the raters and the rating scale are shown in Appendix A.

3. SQT scores for the soldiers were obtained from the SQT Management Directorate, Fort Eustis, as soon as the tests were scored and available on the SQT file. Skill level one (SL 1) soldiers took 67N2180, SL 2 took 67N3180, and SL 3 took 67N4180. Supervisor ratings and all SQT scores available for individual soldiers were then matched for analysis.

## RESULTS

SQT scores were obtained for 66 of the soldiers who were rated by as many as six but no fewer than three supervisors. Data for 14 soldiers were not available or they were rated by fewer than three supervisors. The number of soldiers at each skill level, the average rating and the means and standard deviations for the SQT Total Score, Skill Component (SC), and Hands-On Component (HOC) are shown in Table 1. (No scores are reported for the Job Site Component because all soldier levels scored 100 percent on this component.)

As anticipated, skill level 2 and 3 soldiers were rated significantly higher than SL 1 soldiers ($t(64) = 19.25$ $p < .01$). The mean SQT total score for the 41 skill level 1 67N was significantly higher than that of the population ($Z = 2.58$; $p = .01$). To determine if this indicates that the ability of soldiers in these two units are above average in general ability, the mean AFQ (58) for the sample of 41 was compared with that (57) of a random sample of 100 67N. Since no difference in ability was found, the high SQT scores are probably attributable to the emphasis on individual training in these units.

Table 1

Means and Standard Deviations of Average Ratings
SQT Total Score (SQT), Skill Component (SC), and Hands-On Component (HOC)

| | | Ratings | | SQT | | SC | | HOC | |
|---|---|---|---|---|---|---|---|---|---|
| SL | n | $\bar{X}$ | s | $\bar{X}$ | s | $\bar{X}$ | s | $\bar{X}$ | s |
| 1 | 41 | 3.3 | .84 | 58.0 | 14.7 | 50.3 | 17.3 | 84.5 | 17.2 |
| 2 | 16 | 4.0 | .60 | 64.9 | 21.1 | 57.2 | 25.6 | 86.7 | 14.8 |
| 3 | 9 | 4.1 | .61 | 62.9 | 7.3 | 61.8 | 7.1 | -- | -- [1] |

For analysis purposes, the ratings of each supervisor were converted to numerical ratings: 5 = best of performers; 4 = performer; 3 = borderline; 2 = nonperformer; and 1 = worst of nonperformers. A mean rating for each soldier was calculated from the three to six individual ratings.

The interrater correlations, shown in Appendix B, ranged from .23 to .75 and indicate a moderate degree interrater reliability. All of the correlations in one unit (498th) were statistically significant ($p < .05$) and the correlations in the other unit (121st) were significant where they were based on an adequate number of pairs of ratings. Although the supervisors rated soldiers on overall job performance, the reliabilities were probably reduced because the ratings were based on different samples of behavior (e.g., technical inspector, floor supervisor, shop supervisor, platoon sergeant, quality control NCOIC, C.O.). Although reliability was lower, it is preferable to get higher validity from composite ratings based on a wider range of job performance observations.

The correlation of the average (mean) of the supervisor's ratings, $\bar{R}$, with soldiers' SQT scores are shown in Table 2. The correlation decreased from a substantial correlation for SL 1 to a moderate one for SL 3. However, the small numbers and the compression in the range of ratings for SL 2 and SL 3 soldiers suggest caution in the interpretation of these correlations. It is apparent that the SQT Total Score and SC are more highly correlated with the average ratings than is the HOC, particularly for SL 1.

[1] Skill level 3 had no HOC

Table 2

Correlations Between Average Ratings ($\bar{R}$) and SQT, SC and HOC Scores

| SL | n | $\bar{R}$·SQT | $\bar{R}$·SC | $\bar{R}$·HOC |
|---|---|---|---|---|
| 1 | 41 | .74 | .73 | .25 |
| 2 | 16 | .64 | .62 | .49 |
| 3 | 9 | .35 | .39 | -- |

For purposes of analysis, performers were defined as soldiers receiving a mean rating greater than 3.00. This definition was based on the rating scale and established a priori. Figure 1 shows the distribution of the SQT scores of performers and nonperformers for SL 1 67N.[2] A t-test comparing the mean scores confirms that the performers scored significantly higher than the nonperformers ($t(39) = 7.52$; $p < .01$).

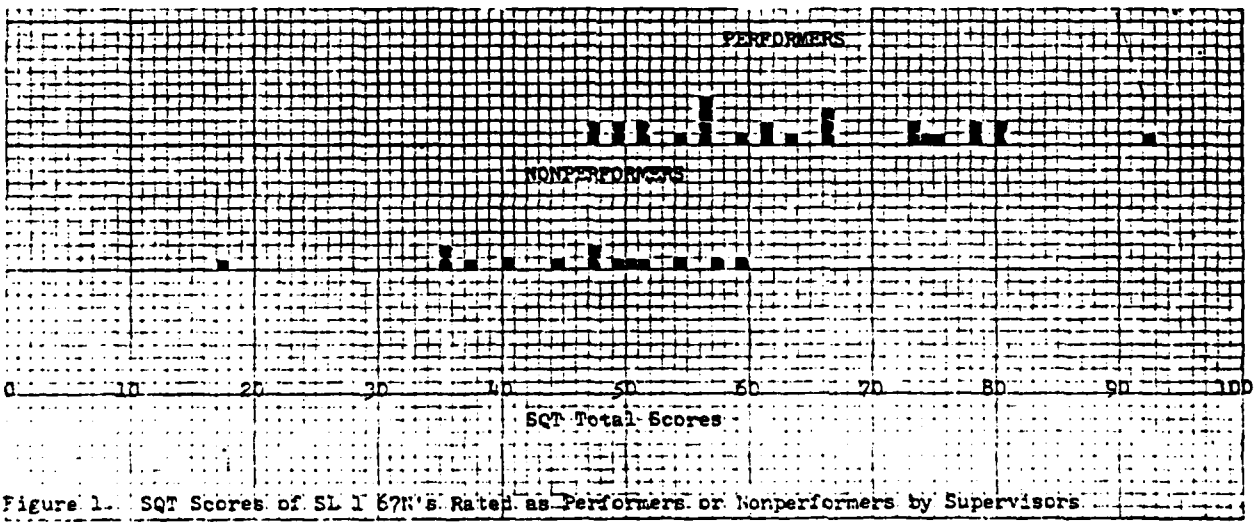



Figure 1. SQT Scores of SL 1 67N's Rated as Performers or Nonperformers by Supervisors

The substantial correlations of ratings with SQT scores and the moderate interrater correlations indicate that the supervisors' ratings

[2] No comparable analysis of SL 2/3 soldiers was possible because all but one were rated as performers.

are valid and reliable criteria for the 67N test. Consequently, the supervisors' ratings were used to identify the optimum cutting score on the SL 1 test. Table 3 shows the frequency and probabilities of correct classifications and misclassifications for different cutting scores. The best estimate from the data in Table 3 is that the optimum cutting score should be set between 45-51 because the probability of making a correct decision is maximized over this range. The true optimum cutting score for the population of SL 1 67N would be even lower than the 45-51 estimated from these data because the mean of the research sample was significantly higher (five points) than that of the population. For example, if 48 were adopted as the optimum cut score because it is the midpoint of range 45-51, an estimate of the optimum cutscore for the population, 43, can be obtained by subtracting five points from 48.

Table 3

Frequencies and Probabilities of Correct Decisions and Misclassification Errors for Different Cutting Scores

| Cutting Score(s) | # of Correct Decisions | p. of Correct Decisions | # of Misclassifications | p. of Misclassifications | |
|---|---|---|---|---|---|
| | | | | False Negative | False Positive |
| 60-61 | 29 | .71 | 12 | .29 | .00 |
| 58-59 | 29 | .71 | 12 | .27 | .02 |
| 57 | 28 | .68 | 13 | .27 | .05 |
| 55-56 | 32 | .78 | 9 | .18 | .05 |
| 52-54 | 32 | .78 | 9 | .15 | .07 |
| 51 | 33 | .80 | 8 | .10 | .10 |
| 50 | 32 | .78 | 9 | .10 | .12 |
| 48-49 | 33 | .80 | 8 | .05 | .15 |
| 45-47 | 33 | .80 | 8 | .00 | .20 |
| 44 | 32 | .78 | 9 | .00 | .22 |

Table 4 shows the impact that cutscores in the optimum range would have had on the 67N SL 1 pass rate. It is apparent that a cutscore in the optimum range would have doubled the pass rate and been more consistent with the perceptions of supervisory personnel.

Table 4

Pass Rate for Different Cutscores

| Cutscore | Pass Rate |
|---|---|
| 60 | 34% |
| 51 | 58% |
| 50 | 59% |
| 49 | 60% |
| 48* | 63% |
| 47 | 64% |
| 46 | 68% |
| 45 | 69% |
| 44 | 73% |
| 43** | 74% |

*Optimum cutscore based on sample data
**Estimate of optimum population cutscore

## DISCUSSION AND CONCLUSIONS

A substantial correlation, r = .74, was obtained between the SQT scores of SL 1 helicopter mechanics and supervisors' ratings. Why should this be? The validation procedures for SQT are conducted at the task level, and a correlation with overall job performance this high was unanticipated.

The 67N SQT was selected for study because of the timing of the administration of the SQT at Fort Benning and because it was one of 10 SQT being evaluated in a separate effort being conducted by the SQT Management Directorate (Brittain and Sova, 1980). Apparently, this was a fortuitous choice. In the 10 SQT analysis, the 67N SQT was found to correlate the highest of

the MOS evaluated with self (r = .51) and immediate supervisor's ratings (r = .36). Helicopter mechanics' SQT scores are highly correlated with ratings because: they work in their MOS; it is a highly technical MOS where work is carefully inspected; they train less (20 hours) for the SQT than soldiers in other MOS; they work on a single system; and most of the 67N SQT consisted of MOS related test units. Therefore, the supervisors' overall ratings for these soldiers are probably based on both MOS related performance and general ability. Because the 67N do not train a lot for the SQT, most of the variance in their written test scores is probably related to their daily work and general ability. In short, this MOS may be a best case for obtaining a high positive correlation between ratings and SQT scores.

The selection of second line supervisors as raters was intended to reduce some of the bias inherent in peer ratings which would be obtained with first line supervisors. The second line supervisors were also selected to insure that common soldier skills would be included as part of the ratings, because training on these tasks is often conducted by the best qualified NCO rather than first line supervisors. In fact, the second line supervisors for 67N tended to base their ratings on different aspects of MOS-related performances. The interrater agreement was only moderate (.23 - .75), although the concurrent validity of the average ratings was quite high. It appears that a valid sampling of a range of job-related behavior was obtained by using three to six second line supervisors. In the aforementioned 10 SQT evaluation, the correlations of a single supervisor's ratings with 67N scores was much lower (.40) than was obtained here. However, the ratings of first line supervisors were generally higher, and more restricted in range, than those of second line supervisors. It is apparent that with first line supervisors, a scale is needed which separates soldiers without requiring the supervisor to designate a soldier as a nonperformer.

The use of 67N and several second line supervisors seems to explain the high correlation obtained. However, it is reasonable to assume that similar results would be obtained by applying these procedures to other technical MOS where soldiers work in their MOS and do not work on a large number of diverse pieces of equipment. These results indicate that for selected MOS, the SQT is a valid indicator of overall MOS performance. Although the ratings were obtained before the soldiers took the SQT (to insure that the ratings were independent of SQT scores), they can be considered as concurrent criteria to assess the validity of the SQT. If the ratings show a moderate to high positive correlation with SQT scores ($>$ .40), then it is valid to use them as criteria for establishing SQT cutscores. These findings demonstrate that it is feasible to obtain the rating data needed to set cutscores using an empirical method. Further research is needed with combat MOS using more sophisticated methods that take into account factors such as the proportion of performers vs. nonperformers in the population and differential weighting of Type I vs Type II misclassifications (Davis, 1980; Steinheiser, et. al., 1978).

The data reported above suggest that essentially the same procedure used here can be applied operationally for selected MOS; identifying units scheduled for testing early in the test window, obtaining valid and reliable multiple supervisor ratings and empirically establishing a cutscore for specific SQT. Optimum cutscores established in this manner, whether lower or higher than the current arbitrary 60 percent mark, should have the effect of improving the actual and perceived fairness of SQT.

## REFERENCES


Berk, R. A. Determination of Optimal Cutting Scores in Criterion-Referenced Measurement, Journal of Experimental Education, 1976, 45, pp. 4-9.

Berk, R. A. Determination of Cutting Scores for the Skill Qualification Testing Program Final Report under Contract DAAG 29-76-0100 with Battelle Columbus Laboratories for Army Training Support Center, Fort Eustis, Virginia, October 1980.

Brittain, C. V. and Sova, J. L. Analysis of 10 SQT, Internal Report for SQT Management Directorate, Army Training Support Center, Fort Eustis, Virginia, October 1981.

Davis, B. Setting SQT Cutscores, Proceedings of the 22nd Annual Conference of the Military Testing Association, October 1980.

Steinheiser, F. H., Epstein, K. I., Mirabella, A. and Macready, G. B. Criterion-Referenced Testing: A Critical Analysis of Selected Models, U.S. Army Research Institute Technical Paper 306, August 1978.

TRADOC, Guidelines for Development of Skill Qualification Tests (draft) published by U.S. Army Training and Doctrine Command (TRADOC), Fort Monroe, Virginia 1980.

## APPENDIX A

## SOLDIER EVALUATION PROCEDURE

The purposes of this evaluation are (1) to identify which of the soldiers you supervise are competent in terms of their overall job performance and soldiers skills, and (2) to determine who are the "very best" and the "poorest" soldiers in terms of overall job performance and their soldier skills. This evaluation procedure will consist of two steps:

1. Rate the competence of the soldiers on the roster by placing a "Y" alongside the soldier's name if you know he (she) is competent in terms of overall job performance; place an "N" next to the soldier's name if you know he (she) is not competent in terms of overall job performance; if you are uncertain whether the individual is competent or incompetent, place a "?" alongside his (her) name. Draw a line through the name of any soldier you do not supervise.

2. Go back over the roster and identify with a "+" (plus) the soldiers you know to be the very best of those you supervise, individuals who would be real assets in a combat enviroment, the last soldiers you would want to lose. Identify with a "-" (minus) the soldiers you know are the poorest of those you supervise, individuals who would probaly not perform well in combat enviroment, the first soldiers you would have replaced if you had the opportunity.

Please complete this evaluation using any sources of information readily at hand, but without consulting with other supervisory personnel in your unit. If you feel that you can improve your evaluations by consulting with other NCOs your ratings can be changed later.

## APPENDIX B

### Interrater Correlations

498th Medical Company[1]

| RATER | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | .47(35)*** | .56(35)*** | .48(35)*** | .60(33)*** | .68(34)*** |
| 2 | | .55(36)*** | .42(36)** | .58(33)*** | .36(35)* |
| 3 | | | .42(36)** | .45(33)*** | .54(35)*** |
| 4 | | | | .59(33)*** | .56(35)*** |
| 5 | | | | | .73(32)*** |

121st Aviation Company[1]

| RATER | 2 | 3[2] | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | .75(15)*** | - | .56(29)*** | .57(29)** | .23(12) |
| 2 | | - | .75(20)*** | .36(20) | .31(10) |
| 3[2] | | | - | - | - |
| 4 | | | | .42(42)*** | .25(21) |
| 5 | | | | | .32(20) |

*p $<$.05; **p $<$.01; ***p $<$.005 (one tailed test)

1 - Entries are correlation coefficient, r, and (n) numbers of soldiers rated.

2 - Only rated six soldiers.

AD P001401

# Success and Failure in Skill Qualification Testing: Troop Views

Joan Harman, Ph.D.
Research Psychologist
US Army Research Institute
for the Behavioral and Social Sciences

A contribution to the ongoing development of the SQT system would be made by exploring reasons for success and failure in testing. One source of this information is enlisted soldiers who take the tests. Accordingly, enlisted soldiers drawn from combat arms, combat support and combat service support MOSs were presented with questions taken from 1980 written SQTs. The questions used were those on which substantial numbers of testees had succeeded or failed. The soldiers were asked to account for the widespread success or failure by other members of their MOS. Most attributed both success and failure to whether tasks being tested were performed as part of unit duties. The lowest number of attributions concerned personal efforts to prepare for written testing. This suggests a need to organize and standardize written test training and either to eliminate or clearly stipulate individual soldiers' responsibilities for test preparation.

# Success and Failure in Skill Qualification Testing: Troop Views

## Introduction

Since the first Skill Qualification Tests were fielded in 1977, the system has been undergoing continuing development and revision to make it responsive to Army needs. This demands increasing efforts to make the system a valid reflection of soldiers' skills and a clear indicator to managers and planners of individual readiness. Accordingly, all facets of the Army have been concerned and involved with Skill Qualification Testing.

In 1980 the Office of the Undersecretary of the Army expressed a special interest in Skill Qualification Testing, with particular emphasis on interpreting low test results. At least in part, this added interest was prompted by media reports about soldiers performing poorly on these tests. One approach to gathering this kind of information is the subject of the research reported in this paper.

Of the three components of SQT -- on-the-job testing, hands-on testing and written testing -- failures have occurred most commonly on the written component. In fact, soldiers have a history of high levels of success on the other two components. Therefore, this paper deals exclusively with performance on the SQT's written test, called the Skill Component (SC).

Of sources of information that would account for soldiers' poor test performance, the most obvious may be test developers, commanders and trainers. Test developers, however, tend to emphasize problems in training programs whereas training managers and trainers focus on problems in the tests. Both of these viewpoints must be carefully weighed, since it is unlikely that any program for training or testing has no margin at all for improvement. There is another source of information about test performance, however, and it is one that may reflect no bias in behalf of either training or testing. This source is the enlisted soldiers who take SQTs. Insofar as individual soldiers are able to identify causes for their own or their unit's failure to answer questions about critical tasks correctly, they could provide clues to identifying training and/or testing variables which would enhance development of the Skill Qualification Testing system. An additional consideration is that even though the main emphasis here is on unsuccessful performance, reasons for success in testing could yield valuable information about reasons for failure.

To explore this hypothesis, enlisted soldiers assigned to Combat Arms, Combat Support and Combat Service Support MOS who took the 1980 Skill Component were interviewed. They were asked to account for widespread success on some SQT questions and for widespread failure on others.

## METHOD

### Subjects

Nine y-four enlisted soldiers participated in this research. They were drawn from two Combat Arms, one Combat Support, and two Combat Service Support MOS and stationed at Ft. Bragg, Ft. Carson, and Ft. Hood.

### Procedure

Skill Component questions were selected from 1980 SQT Item Analyses supplied to ARI by the Army Training Support Center. Criteria for question selection were:

1 - the set of questions for each MOS included both common soldier and MOS specific tasks,

2 - each question had only one correct answer,

3 - more than 50% of the soldiers tested answered the questions correctly/incorrectly. (The percent GO range for successful questions is 70-84 with a median of 81 and for unsuccessful questions is 12-48 with a median of 29.)

For each MOS, one successful and five unsuccessful questions were typed on 5x8 index cards in the same format in which they appeared in the 1980 SQT. That is, each card displayed one task title and number, one general situation, one question and its accompanying answer choices. Appropriate cards were presented one at a time to participants. Each soldier was asked to account for the wide ranging success or failure of soldiers in the MOS on that question. After the soldier's response was recorded, the answer selected was either confirmed as correct or was corrected. This procedure was repeated in the same way for all six questions for each MOS.

## RESULTS

Table 1 shows percent of soldier's responses, ranked from highest to lowest frequencies, in four general categories: performance attributed to events or practices occurring in units, performance attributed to training, performance attributed to test content, and performance attributed to soldiers' personal characteristics. It's important to note that subcategories under each main heading are not intended to be mutually exclusive. For example, under the category Unit Factors, it seems reasonable to assume that a task with which a soldier has had no experience is one that is not job related. Also, a Specialty Task, one that is performed by only a limited number of MOS members, would also be unrelated to the jobs of other members. Although somewhat interrelated, these subcategories were established because every effort was made to reflect the actual responses soldiers made and to avoid interpreting them. These data are summarized by MOS type in Table 2.

TABLE 1

PERCENT RESPONSES BY 94 ENLISTED SOLDIERS ACCOUNTING FOR SUCCESS AND FAILURE ON SKILL COMPONENT QUESTIONS

LESS THAN 50% OF POPULATION ANSWERED QUESTIONS CORRECTLY
(470 TOTAL RESPONSES)

| | 11H | 13B | 12C | 71L | 76X | |
|---|---|---|---|---|---|---|
| UNIT FACTORS | | | | | | |
| PROCEDURE DIFFERENCES | 10 | 8 | 19 | <1 | 10 | |
| NO JOB RELEVANCE | 1 | 8 | 6 | 24 | 10 | |
| NO EXPERIENCE | 2 | 5 | 14 | 9 | | |
| SPECIALTY TASK | 4 | 8 | 10 | 11 | 5 | |
| LOW PRIORITY TASK | | 1 | 1 | | 5 | |
| WORK OUT OF MOS | | | 1 | | 5 | |
| TOTAL | | | | | | 34 |
| TRAINING FACTORS | | | | | | |
| NO TRAINING | 14 | 9 | 14 | 7 | 35 | |
| NONE SINCE BASIC | | 1 | 3 | | | |
| NONE SINCE AIT | 2 | | | 3 | | |
| NONE SINCE SQT | 4 | | 1 | | 10 | |
| INFREQUENT | 9 | 2 | | | | |
| INCOMPLETE | 5 | <1 | 1 | <1 | 5 | |
| TOTAL | | | | | | 19 |
| TEST FACTORS | | | | | | |
| CONFUSING SITUATION/ QUESTION/ANSWER | 8 | 11 | 1 | 9 | | |
| POOR QUESTION | 2 | 3 | | <1 | | |
| ANSWER WRONG | 6 | 7 | 1 | | 1 | |
| EQUIPMENT OBSOLETE | 4 | | 1 | | | |
| TOTAL | | | | | | 15 |
| SOLDIER FACTORS | | | | | | |
| UNLEARNED/FORGOTTEN DETAILS | 8 | 1 | 1 | 9 | 5 | |
| INCOMPLETE KNOWLEDGE | | | 3 | <1 | | |
| NO STUDY | 2 | 5 | 1 | | | |
| MISUNDERSTANDING/ MISREADING | 2 | 6 | | 5 | | |
| CALCULATION PROBLEMS | 9 | | 1 | | | |
| TOTAL | | | | | | 14 |

MORE THAN 50% OF POPULATION ANSWERED QUESTION CORRECTLY
(94 TOTAL RESPONSES)

| | | | | | | |
|---|---|---|---|---|---|---|
| UNIT FACTORS | | | | | | |
| PERFORMED FREQUENTLY | 15 | 26 | | 57 | 25 | |
| FUNDAMENTAL MOS TASK | 20 | 26 | | | | |
| TOTAL | | | | | | 41 |
| TRAINING FACTORS | | | | | | |
| TRAINED FREQUENTLY | 45 | 17 | 50 | 4 | | |
| TOTAL | | | | | | 24 |
| TEST FACTORS | | | | | | |
| ANSWER EASY/OBVIOUS | | 9 | 14 | 14 | 50 | |
| TOTAL | | | | | | 11 |
| SOLDIER FACTORS | | | | | | |
| TEST SOPHISTICATION | 5 | 9 | 7 | 9 | | |
| TOTAL | | | | | | 7 |

Table 2

Percent Responses by 94 Enlisted Soldiers Accounting for Success and Failure on Skill Component Questions

| | FAILURE | | | | SUCCESS | | | |
|---|---|---|---|---|---|---|---|---|
| | Combat Arms (275 resp) | Combat Support (70 resp) | Combat Service Support (125 resp) | $\bar{X}$ | Combat Arms (55 resp) | Combat Support (14 resp) | Combat Service Support (25 resp) | $\bar{X}$ |
| Unit Factors | 25 | 51 | 44 | 40 | 44 | 0 | 52 | 48 |
| Training Factors | 20 | 20 | 17 | 19 | 27 | 50 | 4 | 27 |
| Test Factors | 20 | 7 | 9 | 12 | 5 | 7 | 20 | 10 |
| Soldier Factors | 15 | 7 | 15 | 12 | 1 | 7 | 8 | 7 |

The prominent feature of both tables is that the same rankings occur for both successful and unsuccessful performance. Another feature is that enlisted soldiers' attributions diverged from the tendencies of trainers and test developers to point to test and training factors and, instead, emphasized the importance of performing tasks as part of unit duties in order to be prepared to answer questions about them on the SC. The popularity of this response is demonstrated in Table 2. For both kinds of questions, the sums of the mean values of the remaining categories approximate the single values for Unit Factors. A description of Table 1 subcategories follows in order to reflect in greater detail what soldiers may be thinking.

Under the category Unit Factors dealing with unsuccessful performance, the first subcategory, Procedure Differences, encompasses responses indicating that unit procedures don't always match procedures detailed in Soldier's Manuals on which test questions are based. For example, Bridge Crewmen (12C) were asked about the first thing they would do to construct an individual defensive position in an area designated by their squad leader. The Soldier's Manual answer is "put in sector-of-fire stakes." Most crewmen, however, chose the answer "dig a hasty hole" and insisted that they would follow this procedure regardless of the Soldier's Manual directives. The subcategory No Job Relevance means that soldiers claimed testees chose the wrong answer because the task being tested is not performed as part of unit duties. This was especially true of Administrative Clerks (71L) whose duties tend to reflect the

Under the category Unit Factors dealing with unsuccessful performance, the first subcategory, Procedure Differences, encompasses responses indicating that unit procedures don't always match procedures detailed in Soldier's Manuals on which test questions are based. For example, Bridge Crewmen (12C) were asked about the first thing they would do to construct an individual defensive position in an area designated by their squad leader. The Soldier's Manual answer is "put in sector-of-fire stakes." Most crewmen, however, chose the answer "dig a hasty hole" and insisted that they would follow this procedure regardless of the Soldier's Manual directives. The subcategory No Job Relevance means that soldiers claimed testees chose the wrong answer because the task being tested is not performed as part of unit duties. This was especially true of Administrative Clerks (71L) whose duties tend to reflect the needs of the administrative office to which they are assigned, and who have few or no opportunities to type, for example, non-military letters. No Experience covers responses that indicated soldiers never performed the task. Specialty Task means that task performance is restricted to designated unit members. For example, Administrative Clerks claimed that such a task is Typing Military Orders and Bridge Crewmen pointed to Classifying Vehicles for bridge crossings. Low Priority Tasks tend to be Common Soldier tasks such as first aid and map reading. Soldiers indicated that MOS tasks receive much greater unit emphasis than do Common Soldier tasks. Work Out of MOS is a relatively slender category. It is included, however, because it is a fact of Army life that presents special training and testing difficulties. Very few Subsistence Supply Specialists (76X) in Continental United States perform MOS tasks on a daily basis. Also, even though few Bridge Crewmen attributed SC failure to this problem, an informal interchange with the group prior to interviewing included the fact that substantial numbers of them work out of their MOS.

The subcategories under Training Factors are self explanatory.

Under Test Factors, the first two subcategories, Confusion about test materials and the opinion that test materials are Poor could be related. The difference between a confusing question and a poorly constructed question may be marginal. Nontheless, efforts to avoid subjective interpretation establishes these are separate categories. The Answer Wrong subcategory covers flat statements that the answer soldiers were informed was the correct one is, in fact, the wrong answer to the question. In some cases, soldiers declared that all response choices were wrong. Equipment Obsolete refers to responses in which it was claimed that SC questions dealt with outdated equipment. For example, Antiarmor Crewmen (11H) questioned about engaging targets with an M72A2 LAW reported that they don't use the LAW because it has been supplanted by the TOW.

Under Soldier Factors, the subcategory Unlearned/Forgotten Details most often included details involving numbers; for example, the ratio of chest compressions to lung inflations in cardiopulmonary resuscitation.

Attributions about successful SC performance are virtually a mirror image of those dealing with SC failure. This observation is supported by the fact that anlaysis of differences between mean values of both sets of questions (see Table 2) using the Randomization Test for Matched Pairs shows no significance ($p > .05$). The only subcategory that may not be self explanatory is Test Sophistication under Soldier Factors. It covers responses indicated that soldiers are testwise. For example, those who reported that they didn't know or weren't sure of the correct answer but were able to eliminate clearly incorrect choices or picked up clues to the correct answer in the General Situation or in the test question.

A final significant feature of the Tables is that in no case do percentages sum to 100. Missing values can all be categorized as "I don't know" responses.

## DISCUSSION

The opinions of enlisted soldiers asked to account for widespread failure of individuals in their MOS on SC questions stress the influence of unit activities on test results as opposed to training, test or personal characteristics. This can be interpreted to mean that soldiers don't expect to be prepared to answer test questions on critical tasks that are not part of unit duties. That is, they appear to discount the influence of training on SC success and they show little acceptance of personal responsibility for test preparation. This seems to indicate a need for clearer communication about expectations surrounding SC preparation.

One area of misunderstanding may be related to the way the other two components, the Job Site and Hands-on, are handled. Job Site tasks are tested on-the-job and Hands-on tasks are tested at special test sites at which both training and testing take place. Preparation for these two components tends to be well organized and standardized from unit to unit and requires little initiative on the part of individual soldiers. The same passive approach soldiers are able to take to Hands-on and Job Site preparation seems to carry over to Skill Component preparation. SC performance might be improved if units would either organize and standardize written test preparation in the same way as the other two components are managed, or insure that individual soldiers' responsibilities are clearly understood.

## CONCLUSIONS

Techniques for preparation for SC testing need to be re-examined. Guidelines for Hands-on and Job Site preparation follow naturally from explicit directions for testing and scoring set forth in SQT Notices. This is much less the case for written testing.

Written test preparation could be organized and standardized in the same way as are the other two components. One way to do this would be to administer written tests on SC tasks prior to official testing. If, however, soldiers are to assume individual responsibility for written test preparation, this requirement needs to be communicated clearly and training in individual study methods may have to be adopted.

AD P001402

# Skill Qualification Test Feedback: Timeliness Matters

John J. Kessler
US Army Research Institute for the Behavioral and Social Sciences

A field experiment was conducted in which the time period was varied between completion of the Skill Qualification Test and feedback of results. Data are presented which show the marked decline in soldiers' perceived utility of the feedback over time. In addition soldiers indicate a greater tendency to put off utilization of the feedback the longer the feedback is delayed.

Background and Purpose.

Between 15 July and 30 November 1980, the Army Training Board (ATB) locally scored the Skill Qualification Tests (SQT) taken by soldiers at Ft Bragg, NC. Normally, these tests would have been shipped to the central scoring facility at Ft Eustis, VA, and the results shipped back to Ft Bragg. The Local Scoring Evaluation (LSE) project was carried out to determine whether the advantages of local scoring are worth the additional costs and efforts such a system requires. This paper describes a portion of the research conducted by the Army Research Institute (ARI) in support of the LSE project.

A major objective of ARI's technical advisory service (TAS) to the Army Training Board during the Local Scoring Evaluation was to determine whether the speed of feedback of SQT results had any immediate effects on individual soldiers. The existing centralized scoring system had a goal of thirty days' turnaround time for SQT results. Although this seemed inordinately long, no evidence existed to indicate whether any particular number of days should be preferred. Given that immediate feedback was not possible, the question became whether there was any practical turnaround time that might make a difference to soldiers.

To answer this question, a field experiment was set up in conjunction with the LSE project. The basic idea was to vary the length of time between completion of the SQT and feedback of results, and to obtain the soldiers' reactions to the variation.

Design Plan

The design plan divided the 82d Airborne Division into three approximately similar groups and assigned each to one of three treatment conditions The treatment variable was speed of SQT results feedback. Its three levels were Fast, Medium, and Slow. "Fast" meant that a soldier in the group received the results of his SQT within 1 to 15 days; "Medium" 16 to 30 days; "Slow" 31 to 45 days. The Slow condition was believed to simulate the current central system's normal operation.

Composition of the groups by unit was the following:

| FAST | MEDIUM |
|---|---|
| 2d Bn (ABN), 508th IN | 2d Bn (ABN), 504th IN |
| 3d Bn (ABN), 325th IN | 2d Bn (ABN), 325th IN |
| 2d Bn (ABN), 505th IN | 1st Bn (ABN), 505th IN |
| 1st Bn (ABN), 319th FA | 1st Bn (ABN), 320th FA |
| 782 Maint Bn | 407th S&S Bn |
| 82d Finance Co | 82d AG Co |
| 1st Squadron, 17th Cav | 82d Aviation Bn |
| 82d Signal Bn | 4th Bn, 68th Armor |
| 82d MP Co | |

SLOW

1st Bn (ABN), 504th IN
1st Bn (ABN), 325th IN
1st Bn (ABN), 508th IN
2d Bn (ABN), 321st FA
307th Med Bn
21st Chem
3d Bn (ABN), 4th AD Arty
313th MI Bn (CEWI)
307th Eng Bn

The formation of the groups was made with the following rules in mind: (a) Each group should be about the same size, (b) The MOS involved should be locally scored, (c) The MOS distribution across groups should be balanced, (d) For a brigade composed of three battalions, one battalion should be assigned to each group. The goal of the design plan was to define three comparable groups from which samples could be obtained and analyzed. In a later section the samples will be described.

## Instruments

The form of the SQT results feedback was a computer printout called the Individual Soldier Report (ISR) which gave the soldier's SQT score, a detailed list of failed tasks, failed items within tasks, and references to paragraphs in Soldier's Manual. The ISRs were generated by ATB's local scoring facility at Ft Bragg.

The instrument used to obtain soldiers' reactions to the feedback of their SQT results was the ISR/SQT Feedback Evaluation Record. Questions on the Evaluation Record were aimed at determining the utility of the ISR, SQT score, certain SQT preparation variables, and some indications of what these SQT results meant to the soldier. Try-out of this instrument during its developmental phase showed that nearly all soldiers could complete it in less than twenty minutes.

## Procedures

The procedures followed in this research were the following:

1. An ARI researcher personally delivered or observed the delivery of an ISR to a soldier.

2. The soldier was given sufficient time to examine his ISR and digest its contents.

3. An ARI researcher administered the ISR/SQT Feedback Evaluation Record to the soldier.

This data collection was accomplished under real-time operational condition and was dependent on each unit's SQT scheduling choices. Thus it was sometimes necessary because of manpower and time constraints to decide which units to sample. High density MOS were preferred over low density

MOS, large units over small ones, and combat MOS over non-combat MOS. Adherence to these priorities resulted in a fairly large total sample with sufficient numbers of soldiers at the treatment levels to permitf statistical analysis.

Sample Description

During the duration of the LSE at Ft Bragg 3166 SQTs were scored and their corresponding ISRs produced. A total of 576 soldiers completed the ISR/SQT Feedback Evaluation Record and 558 of these were usable in this analysis. As might be expected from the procedural priorities, the combat MOS* provided more than their share of the sample. In the analyzed sample 513 of 558 or 91.9% were combat MOS, whereas 2232 of the 3166 SQTs scored or 70.5% were in combat MOS.

At this point a variety of reasons compelled the reduction of the sample to be analyzed to just the combat MOS. Table 1 shows one of the realities of coordinating a real-time experiment under operational conditions: No non-combat MOS soldiers were obtained under the "Slow" feedback condition.

Table 1. Number of soldiers at each treatment level

| | Fast (1-15 days) | Medium (16-30 days) | Slow (31-45 days) | N |
|---|---|---|---|---|
| Combat MOS | 255 | 111 | 147 | 513 |
| Non-combat MOS | 28 | 17 | 0 | 45 |
| Total | 283 | 128 | 147 | 558 |

Tables 2 and 3 show the percentage distributions of the combat and non-combat MOS on two background variables which might interact with reactions to SQT results feedback, time in service and time left to serve. The dissimilarities between combat and non-combat MOS soldiers appear to further justify removing the few non-combat MOS soldiers from the analysis.

Table 2. Time in service (months) percentage distribution

| | 12 or less | 13-24 | 25-36 | 37-48 | More than 48 |
|---|---|---|---|---|---|
| Combat MOS (N=513) | 4.1 | 50.5 | 19.1 | 13.3 | 13.1 |
| Non-combat MOS (N=45) | 0.0 | 46.7 | 26.7 | 0.0 | 26.7 |

*Combat MOS were CMF 11, 12, 13, 17 or 19; non-combat MOS were any others.

Table 3. Time left to serve (months) percentage distribution

| | 12 or less | 13-24 | 25-36 | 37-48 | More than 48 |
|---|---|---|---|---|---|
| Combat MOS (N=513) | 25.7 | 39.8 | 28.5 | 3.7 | 2.3 |
| Non-combat MOS (N=45) | 28.9 | 57.8 | 8.9 | 0.0 | 4.4 |

Thus a fairly large (N=513) and fairly homogeneous (all combat MOS) sample was constructed for determining whether increasing the speed of feedback of SQT results makes any difference to soldiers.

Results

Of primary concern to ATB was whether increasing the speed of SQT results feedback mattered or had any effect on soldiers.

A question on the ISR/SQT Feedback Evaluation Record addressed this directly, i.e., "How useful to you is this ISR to the improvement of your military skills?" Responses were obtained on a five point scale: 5=very useful, 4=useful, 3=slightly useful, 2=not very useful, and 1=in no way useful. The means of the combat MOS soldiers for the three treatment levels are shown in Table 4.

Table 4. Ratings of ISR usefulness by soldiers with combat MOS

| Speed of SQT Feedback | Mean | S.D. | N | |
|---|---|---|---|---|
| Fast (1 to 15 days) | 4.27 | .89 | 253 | |
| Medium (16 to 30 days) | 3.25 | 1.49 | 111 | F=60.24 |
| Slow (31 to 45 days) | 2.62 | 1.34 | 146 | $p < .001$ |
| | 3.57 | 1.38 | 510 | |

Table 4 indicates quite clearly that soldiers had a more favorable view of the feedback of SQT results the earlier it came. Another variable which might have affected a soldier's immediate reactions to feedback was whether it was positive or negative, did he pass or fail. Tables 5 and 6 which separate passers and failers show the same pattern as before with a slight decrease in overall mean usefulness rating by the failers.

Table 5. Mean ratings of ISR usefulness by soldiers with combat MOS who passed SQT (SQT score of 60 or more)

| Speed of SQT Feedback | Mean | S.D. | N | |
|---|---|---|---|---|
| Fast (0-15 days) | 4.28 | 0.87 | 240 | |
| Medium (16-30 days) | 3.22 | 1.50 | 92 | F=91.87 |
| Slow (31 to 45 days) | 2.62 | 1.36 | 134 | $p < .001$ |
| | 3.59 | 1.38 | 466 | |

Table 6. Mean ratings of ISR usefulness by soldiers with combat MOS who failed SQT (SQT score of less than 60)

| Speed of SQT Feedback | Mean | S.D. | N | |
|---|---|---|---|---|
| Fast (1-15 days) | 4.15 | 1.17 | 13 | |
| Medium (16-30 days) | 3.42 | 1.39 | 19 | F=4.22 |
| Slow (31-45 days) | 2.67 | 1.03 | 12 | $p < .05$ |
| | 3.43 | 1.36 | 44 | |

The gradient of feedback utility shown in these results is useful to people who must design the data processing and distribution system. For people concerned with training, however, there is the question of what action will take place as a consequence of providing timely and useful feedback. Some questions were asked on the Evaluation Record in order to obtain a notion of what the soldiers would do. Results of two questions are shown on Table 7. One question asks the soldier when he expects to study on his own the items he missed; the other asks when he believes his unit should provide training on what was missed. The a priori best answer, i.e., a five on a five point scale, is "in the next few days." The expectation might be that failers would have lower means than passers for both questions, and they did. What is interesting, however, is that both groups look to their unit to move more quickly on taking remedial action than they intend themselves.

Table 7. Mean response of combat MOS soldiers.
(5=in the next few days, 4=next month, 3=In the next six months 2=just before the next SQT, and 1=don't know)

| Question | Passers | | Failers | |
|---|---|---|---|---|
| | Mean | N | Mean | N |
| When do you expect to do some training/studying, on your own on the task items you missed? | 2.8 | 466 | 2.5 | 45 |
| When do you believe your unit should provide some training on the tasks most soldiers failed? | 3.1 | 466 | 2.8 | 45 |

These questions were also analyzed in terms of the speed of feedback variable. Shown in Table 8 are the percentages of soldiers responding "Don't Know" to the two questions. There are two important points to be made in regard to these results. One is that these are rather substantial percentages. Perhaps this is a true picture as far as the soldier's own need or opportunity to self-train. The other point is that the percentage of Don't Knows increases as the time since SQT increases. It could be suggested here that the longer the delay between testing and feedback of results, the more depressed is the motivation to review or take any remedial actions.

Table 8. Percentages of soldiers with combat MOS responding "Don't Know"

| | Speed of Feedback | | |
|---|---|---|---|
| Question | 1-15 days | 16-30 days | 31-45 days |
| When do you expect to do some training/studying, on your own, on the task items you missed? | 25.9% | 27.0% | 32.0% |
| When do you believe your unit should provide some training on the tasks most soldiers failed? | 14.1% | 18.0% | 18.4% |
| | N=255 | N=111 | N=147 |

Discussion

In learning psychology, feedback is viewed as a secondary reinforcer and as such has its greatest effect when it is immediate. With soldiers and their SQT performance it is currently not possible to provide immediate feedback. It is, however, possible to significantly increase the speed of feedback from a matter of several weeks to a few days. In this research it was found that even though the speed of feedback falls far short of being immediate, speed still matters. The perceived utility of the feedback (the ISR) is a function of time, remaining high during the 1-15 day period and then dropping steadily to a point where its mean rating is less than "slightly useful" somewhere between 31 and 45 days after SQT completion. This pattern was the same for both failers, where the feedback had a negative message, and passers where the feedback was relatively positive.

A high value placed on the utility of feedback does not, of course, mean that the information supplied is taken to heart and remedial studying or training is pursued forthwith. Results of this research indicate, however, that soldiers who received their ISRs more quickly intend to do some studying or training sooner than those whose ISRs were less timely. Speed of feedback was also associated with the desire to have sooner unit training on items missed on the SQT. It might be hypothesized that the speed with which SQT results were fed back to soldiers had a pacing effect, i.e., rapid feedback suggested that rapid corrective action was expected whereas slow feedback suggested there was no urgency about correcting mistakes.

The increase over time of "Don't Know" responses to the questions of when studying or training will or should be done could have several causes. What seems most likely is that interest or motivation had waned with the passage of time. SQT was a hot topic for a week or two and then became another event to be forgotten. It could also be that for many soldiers there became objectively less reason to be concerned about their ability to perform the SQT tasks. Whatever the reason, it appeared that the longer the feedback of results was delayed, the less likely soldiers would be to constructively use information contained in the feedback.

## Conclusion

SQT feedback in the form of the ISR has a rather short shelf life, perhaps two weeks. The data presented show a noticeable decline of the perceived utility of the ISR over time. There are also indications that the longer the delay in getting the feedback to the soldier, the weaker is his intention to act on it. The results further show that the soldier expects more initiative from his unit than from himself when it comes to training on SQT tasks which were failed.

AD P001403

# Predicting Skill Qualification Test Item Difficulty from Judgments

Douglas Macpherson
Research Psychologist
US Army Research Institute
for the Behavioral and Social Sciences

Judgments of item difficulty by small groups of three to six non-commissioned officers were compared with observed item difficulties among soldiers in three military occupational specialties representing infantry, engineer and administrative career fields. Linear correlations between average judgments and observed difficulties were on the order of .50, but the scatter plots were triangular in appearance because objectively easy items were rarely judged to be difficult while objectively difficult items yielded a wide range of judged difficulties. Hence sets of items showing wide and fairly flat distributions of difficulty had been judged to be skewed toward the easy end of the difficulty distribution. These analytic observations suggest that NCOs involved in test construction may be making tests more difficult than they believe and that NCOs as trainers preparing soldiers for their SQTs may be underestimating the need for training. If the triangular relationship between judged and observed difficulty is confirmed in larger samples of items, then a simple expectancy table method might be used to predict objective test difficulty and training need.

# Predicting Skill Qualification Test Item Difficulty from Judgments

## INTRODUCTION

Teams of experienced soldiers, oriented in criterion referenced test development workshops and assisted by civilian test psychologists, produce Skill Qualification Tests (SQT). The procedures and policy guidelines were recently reviewed by a well known authority on criterion referenced testing (Hambleton, 1981). His assessment was that the prescribed procedures were excellent. His only reservations concerned the lack of reliability information and the obvious possibility of failure in execution. However, Hambleton apparently accepted that standards can be set by testing proficient and non-proficient personnel on each task and then selecting cut points which discriminate between them without considering the total scores.

This study examined one possible form of test development bias. We wanted to determine if supervisors exhibited "item leniency," and judged the items to be easier for enlisted personnel than the items were found to be. Thus we asked supervisors of a combat MOS (11H, TOE Gunner), an engineering MOS (12C, Bridge Construction Crewman), and an administrative MOS (71L, Clerk) to estimate what percentage of their troops would pass each item of the Skill Component (SC) portion of the appropriate SQT.

## METHOD

Subjects: The subjects were 10E5 - E7 NCO supervisors and 2 E4 Acting Supervisors for MOS 11H, 12C, and 71L at Forts Bragg, Carson and Hood. They were distributed as follows: (a) six 11H, (b) three 12C, and (c) three 71L. All supervisors held the appropriate MOS or a closely related MOS as their primary or secondary MOS. In addition they had held the MOS for an average of 4 years (range 10 months - 15 years) and had supervised for an average of 4 1/2 years (range 8 months - 18 years).

Procedures: Supervisors at the three posts reviewed the SC which had been administered to their subordinates within the last year. They were instructed to judge what percentage of their soldiers could get each item correct and to write the estimate in the SQT. Because of the length of the tests each supervisor was asked to rate a specified set of subtests which usually comprised about one half of the test. Some supervisors, however, completed all the items. The rating task required about a half hour to complete.

Criterion data: Item analyses for each SQT were provided by SQT Management Directorate of TRADOC. These analyses provided the total number in of testees in each sample and the percentage of testees selecting each alternative for each item.

## RESULTS

We determined the reliability of the criterion data for one SQT for which we had two samples. The results for each SQT were examined for rating reliability and rating validity. The rating reliability for each SQT is presented in terms of the total number of raters used. The reliabilities are then recalculated to the common metric of the correlation to be expected between a pair of raters by using the Spearman Brown Prophesy formula. Similarly rating validities are presented as the correlation of the mean rater estimates with the criterion, and then corrected for the attenuation due to predictor unreliability. As a result of these adjustments, reliabilities and validities may be compared across the three SQT. Finally, the validity results for the three SQT were combined. The analyses are presented in the form of two way tables.

Criterion Reliability. The 11H SQT results were obtained as two samples of 789 and 922 soldiers respectively. These samples consisted of all 11H2180 scores which had been transcribed by SMD before 21 Dec 80 and between 21 Dec and 22 Mar 81. The product moment correlation for the proportions of soldiers in the two groups who passed the items was .99. Table 1 displays the high correlation obtained and demonstrates that the test contained a relatively uniform spread of difficulty levels. Table 1 also indicated that there was little evidence for change in scores over the course of the two administrations. This was supported by the regression equation: Second Sample = .05 + .924 First Sample. The mean and sigma for the two samples combined were 61 and 20 respectively. Multiple samples were not available for the other SQT therefore criterion reliability analyses were not possible for those SQT.

**Table 1. Reliability of Item Analysis Difficulty Levels for 61 Item Test 11H2180 SC**

| | | First Sample | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | 10 19 | 20 29 | 30 39 | 40 49 | 50 59 | 60 69 | 70 79 | 80 89 | 90 100 | n | % |
| Second Sample | 90-99 | | | | | | | | 1 | 4 | 5 | 8 |
| | 80-89 | | | | | | | 1 | 9 | | 10 | 16 |
| | 70-79 | | | | | | 2 | 6 | 2 | | 10 | 16 |
| | 60-69 | | | | | 1 | 3 | 3 | | | 7 | 11 |
| | 50-59 | | | | 1 | 8 | | | | | 9 | 15 |
| | 40-49 | | | 5 | 4 | 1 | | | | | 10 | 16 |
| | 30-39 | | 2 | 4 | 1 | | | | | | 7 | 11 |
| | 20-29 | 2 | 1 | | | | | | | | 3 | 5 |
| | 10-19 | | | | | | | | | | 0 | 0 |
| | n | 2 | 3 | 9 | 6 | 10 | 5 | 10 | 12 | 4 | 61 | |
| | % | 3 | 5 | 15 | 10 | 16 | 8 | 16 | 20 | 7 | | 100 |

11H2180 Reliability and Validity. Six supervisors of 11H troops rated all subtests. These raters were divided into two groups and the mean difficulty ratings for the two groups were compared. The split half correlation was .78 (expected pairwise correlations of .54). Table 2 confirms that the two groups generated comparable estimates.

Table 2. Split Half Reliability of Estimated Item Difficulty Levels for SQT 11H2180 (6 raters)

| | | Group A (3 raters) | | | | |
|---|---|---|---|---|---|---|
| | % Correct | 20-39 | 40-59 | 60-79 | 80-99 | n<br>% |
| Group B (3 raters) | 80-99 | | 3 | 3 | 19 | 25<br>42 |
| | 60-70 | 1 | 2 | 5 | 6 | 14<br>23 |
| | 40-59 | | 1 | 6 | | 7<br>12 |
| | 20-39 | 6 | 8 | | | 14<br>23 |
| | n<br>% | 7<br>12 | 14<br>23 | 14<br>23 | 25<br>42 | 60<br>100 |

The validity correlation for the mean item difficulty estimates was .49. When corrected for predictor attenuation the r became .55. Although the standard deviations were about equal, 23% vs 21%, the mean estimated percent correct was 68% vs 61% for the actual test. Thus the NCOs tended to agree with each other but vary from the criterion and underestimate the difficulty of the test items. Furthermore, the underestimates were not uniform. The supervisors correctly estimated the difficulty of the easy items, but not the difficult items. As a result the upper left quadrant was empty, whereas the lower right quadrant contained 1/5 of the observations and about the same density as the validity diagonal. Since the quadrant boundaries represent approximately a 1.5 SD difference from valid predictions, the significance of the effect is apparent. These effects appear in Table 3.

Table 3. Validity of Item Difficulty Estimates for Test 11H2180

| Observed % Correct (n = 1711) | Estimated % Correct (6 raters) | | | | | |
|---|---|---|---|---|---|---|
| | % | 20-39 | 40-59 | 60-79 | 80-100 | n<br>% |
| | 80-100 | | | 5 | 12 | 17<br>28 |
| | 60-79 | | 1 | 6 | 6 | 13<br>22 |
| | 40-59 | 7 | 5 | 2 | 4 | 18<br>30 |
| | 20-39 | 3 | 1 | 6 | 2 | 12<br>20 |
| | n<br>% | 10<br>17 | 7<br>12 | 19<br>32 | 24<br>40 | 60<br>100 |

12C2180 Reliability and Validity. Three NCOs for 12C soldiers rated all subtests. The consistency of these ratings was determined to be .40 with Cronbach's Alpha; the pairwise correlation was estimated to be .31. The correlation across 121 items of these NCOs with the observed performance of soldiers on SQT 12C2180 SC was .50 (.79 corrected for predictor attenuation). The SDs for the NCO and criterion data were both 20%, whereas the means were 69% and 64% for the raters and the criterion respectively. The dispersion of the ratings is displayed in Table 4. The row marginal totals indicated that the test item difficulties were uniformly distributed.

Table 4. Validity of Item Difficulty Estimates for Test 12C2180 (SC)

| Observed % Correct | Estimated % Correct (3 raters) | | | | | |
|---|---|---|---|---|---|---|
| | % | 20-39 | 40-59 | 60-79 | 80-100 | n<br>% |
| | 80-100 | 1 | 3 | 9 | 18 | 31<br>26 |
| | 60-79 | 2 | 7 | 10 | 13 | 32<br>26 |
| | 40-59 | 2 | 11 | 9 | 13 | 35<br>29 |
| | 20-39 | 5 | 8 | 7 | 1 | 21<br>17 |
| | 0-19 | 2 | | | | 2<br>2 |
| | n<br>% | 12<br>10 | 29<br>24 | 35<br>29 | 45<br>37 | 121<br>100 |

In contrast, the column marginal totals indicated that the raters tended to generate a negatively skewed distribution. Again there tended to be many fewer responses in the upper left than in the lower right quadrant.

71L2180 Reliability and Validity. The 71L2180 Skill Component introduced additional complexities. It utilized multiple correct responses to some items. Since the multiple correct answer format is no longer used in SQT construction we excluded these items from the analyses.

Only three raters were available for this SQT. Rater #3 completed the entire questionnaire whereas two other raters completed the first and second halves of the SQT respectively.

The reliability of the three raters was estimated by correlating the estimates of rater #3 with the combination of the other raters separately. The reliability correlation was found to be .43. We regarded this as a pairwise correlation and did not reduce it further.

The validity was estimated by correlating the means of the appropriate pairs of ratings with the criterion scores. The mean and SD for the difficulty estimates were 65% and 21% respectively. However, the test was actually more difficult than the others examined. Thus the criterion mean was 50%. However the criterion variance was a typical 19%.

The validity correlation for the three raters was a low .28. Adjusting this correlation for rater unreliability increased the validity correlation to .40.

The distribution of the items is presented in Table 5. Table 5 displays the usual pattern of underestimating the item difficulties. The raters tended to rate the easy items as easy. However, they failed to demonstrate such consistency in their ratings of the difficult items. Instead they tended to use all rating values in evaluating the difficult items. Again, if they identified an item as difficult it was difficult.

**Table 5. Validity of Item Difficulty Estimates For Test 71L2180 (SC) Single Correct Response Items**

| | Estimated % Correct (mean of 2 raters) | | | | | | |
|---|---|---|---|---|---|---|---|
| Observed % Correct | % | 0-19 | 20-39 | 40-59 | 60-79 | 80-100 | n<br>% |
| | 80-100 | | | | 1 | 2 | 3<br>4 |
| | 60-79 | 1 | 1 | 2 | 4 | 9 | 17<br>25 |
| | 40-59 | | 5 | 6 | 13 | 7 | 31<br>46 |
| | 20-39 | | 2 | 3 | 4 | 3 | 12<br>17 |
| | 0-19 | | 1 | 2 | 1 | 1 | 5 |
| | n<br>% | 1<br>1 | 9<br>13 | 13<br>19 | 23<br>34 | 22<br>32 | 68<br>100 |

Item Difficulty Estimates Summarized. The results for the three SQT are summarized at the item level in Table 6. The mean and standard deviation for the observed difficulty across the three SQT were 58% and 22% respectively. Similarly, the mean and SD for the difficulty estimates were 68% and 21% respectively. Table 6 is in expectancy table format. The number in each cell is the percent of items in the table that are in the cell.

Table 6. Validity of Item Difficulty Estimates for Test 71L2180 (SC) Multiple Correct Response Items

| Observed % Correct | Estimated % Correct | | | | | | |
|---|---|---|---|---|---|---|---|
| % | 0-19 | 20-39 | 40-59 | 60-79 | 80-100 | n % |
| 80-100 | | * | | 6 | 13 | 20 |
| 60-79 | * | 1 | 4 | 8 | 11 | 25 |
| 40-59 | | 6 | 9 | 10 | 10 | 34 |
| 20-39 | | 4 | 5 | 7 | 2 | 18 |
| 0-19 | | 1 | 1 | | * | 2 |
| % | * | 12 | 20 | 31 | 37 | 100 |

* Only one response

The marginal totals demonstrated that the item difficulties were relatively uniformly distributed through the range 20% - 100%. However, the negative skew of the estimates is clearly seen. Thus, the pattern of errors was non-random and non-linear. As expected, Table 6 exhibits a nearly triangular matrix. The upper left quadrant contains only two percent of the sample whereas the lower right quadrant contains 19% of the sample and the cell densities resemble the cell densities on the main diagonal. Across the three MOS, easy items were rated as easy whereas the difficult items received almost any rating. However, if a supervisor rated an item as difficult then the item was difficult

## DISCUSSION

This preliminary research demonstrated that NCOs can predict the item by item test performance of their troops to a moderate degree. However, they do tend to underestimate the difficulty of more than 40% of the items. If our conjecture that the test developers would respond like NCOs is correct, then this study does suggest that one reason for low SQT scores is that the tests are harder than intended from a normative point of view. Furthermore this is not detected during SQT pretesting for good reasons. The SOP describes a procedure in which each subtest is tested independently of the other subtests using personnel who are expert and non-expert on that subtest. As a result no estimates of test total scores are available prior to field use of the test.

There is a second way in which NCO underestimation of item difficulty can yield low SQT scores. The NCOs who performed the ratings were responsible for training troops to take the test. They were provided with the topics to be covered in the test and sample items for each topic. They may have trained their troops to the apparent difficulty level of the test. As a result they may tend to provide insufficient training for the SC portion of the test.

The combination of the two effects - tests more difficult than intended, and inadequate training for the test - could account for the disappointing scores observed on many SQT.

## REFERENCE NOTE

Hambleton, R. D. Psychometric methods for the Skill Qualification Test - Final Report. Amherst, Mass.: University of Massachusetts, 1981.

## REFERENCES

Department of the Army. Guidelines for Development of Skill Qualification Tests (SQT) Policy and Procedures. TRADOC Pam 351-2 (Draft) Ft Monroe, VA: Headquarters, US Army Training and Doctrine Command, Updated (Supercedes TRADOC Pam 351-2, 1 Dec 1977).

Department of the Army. Skill Qualification Test (SQT) Policy and Procedures. TRADOC Reg 351-2. Ft. Monroe, VA: Headquarters, US Army Training and Doctrine Command, 27 April 1980.

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. J. Education Measurement, 1978, 15, 277-290.

AD P001404

## Readability of Materials

Duncan, R. Eric, 1LT, USAF Occupational Measurement Center, Randolph Air Force Base, Texas (Chair); J. Peter Kincaid and Richard Braby, Training Analysis and Evaluation Group, Orlando, Florida and Wulfeck, Wallace H. II, Navy Personnel Research and Development Center, San Diego, California; Lydia R. Hooke, HumRRO, Alexandria, VA.; Thomas G. Sticht, HumRRO; Nancy A. Thompson, Douglas K. Cowan, and John A. Guerrieri, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

The discussion in this session centered around three areas: (1) the relationship of symbols and words, (2) whether lists of acronyms can be used to reduce the readability level of text, and (3) how writers produce materials (goals and objectives). The majority of the time was spent discussing symbols and words. The problem, as addressed by Drs Lydia Hooke, Peter Kincaid, and Tom Sticht, involves how information is processed and, thereby, how it is decoded and synthesized. While some believed that reading is the assimilation of words in logical sequence, others viewed the process as an inference of meaning. The loci of controversy was the estimation of readability. If meaning and content are deemed as the important issues, one school of thought would view the issue as one of measuring the components of a text passage (words and sentences) and infering levels of measuring from the difficulty of the component parts. The other arguement voiced was the impossibility of assuming and measuring content understanding separate from the author of the material. An example was raised where it is possible to write a passage free from jargon and acronyms and at a grammatically low level but which could not be understood. In other words, the formulas used by DOD and other groups and agencies to estimate readability are structurally oriented, have little theoretical base, and bypass meaning. While no solution to this problem will ever be fully described, many ideas for (1) incorporating meaning into computer programs for accessing readability and (2) dealing with symbols (acronyms) were shared.

The most important information to come from the panelists was reemphasis of the need for theory. Most research has dealt with the mechanics of readability rather than the theory behind the mechanics. The goals of readability experts should not soley revolve around text and test construction, but should also include a theoretical perspective on the origins and the processes of human receptors of reading materials.

# Measuring and Improving the Readability of Military Documents

Lydia R. Hooke

Human Resources Research Organization

A presentation given at the 23rd Annual Conference of the Miltary Testing Association
Thursday 29 October 1981

I first became involved with the related enterprises of measuring and improving the readability of military documents during the two years I worked for the Air Force Human Resources Laboratory. I became convinced then that the attempt to measure readability is, at best, an incomplete and misleading undertaking; that this is particularly true for military documents and that imposing standards requiring certain levels of measured readability for documents is not a good way to improve document readability and may even be counterproductive. What I'm going to do today is outline a case against the utility of readability measurement for the military and then go on to introduce a possible alternative for improving readability, which relies on computer technology already available.

Let's start with a brief description of how readability formulas are typically derived. Their generation usually has three aspects. Certain easily measurable and unambiguously quantifiable characteristics of texts, e.g., mean sentence length or number of one syllable words in a passage, alone or in combination (apect 1) are correlate with the reading grade level (aspect 2) of people who perform adequately (according to some arbitrary criterion) on a verbal comprehension test constructed from the passage (aspect 3) and this grade level is assigned to the text. There are serious problem with each of these three aspects of the procedure: the text features used, the identification of text difficulty with the reading grade level of a population and the verbal (paper and pencil) tests used to indicate readability.

The first set of problems has to do with the type of text features which must be used as readability indicators. To perform correlations, one is forced to limit oneself to characteristics to which a number can be assigned and for which this number will vary in value continuously for different texts. Exclusive concern with such features means that one is forced to ignore the relationship between the words and sentences of the text and its meaning or communicative intent. Thus to a readability formula, perfectly meaningless texts can receive good readability scores. One might argue that for normal texts features, such as mean word or sentence length, are typically good indicators of clear meaning, but this is true only statistically and need not be so in any given case. People trying to improve writing should be induced to direct as much of their attention as possible to the relationship between what is said and what is meant, yet readability formulas may divert attention from this relationship.. Readability formulas, because they ignore meaning, cannot tell a writer exactly how to improve the relation between words and meaning yet this is after all the ultimate goal of the enterprise.

A related issue is that, because of the nature of the statistical assumptions underlying readability measurement, only longish connected texts can be assessed by these formulas. Yet, many of the written documents most important in the military (e.g., procedural instructions or forms) are not in this form and thus cannot be assessed. When, in the Air Force, we attempted to study the effects

of readability standards on actual Air Force writing (Hooke et al, 1980) we were hard put to find, among scores of submitted documents, six with several passages long enough for formula measurement. It is either ironic or encouraging, depending on point of view, that suggestions for improving the usability of real world documents (e.g., by making them into flow charts or proceduralized job aids) involve putting them into a form where they are not amenable to readability measurement.

A third problem wtih this sort of readability measurement is that not only is much traditional wisdom about what constitutes good writing ignored, but new and/or newly empirically proven insights about what makes texts readable, coming from fields of experimental psychology and education cannot be incorporated into formula assessment of readability. Thus, I can show, with any number of college sophomores, that too many of a certain type of construction obscures text clarity, but if I can't determine how much exactly is too many, I can't incorporate this wisdom into my assessment.

The next set of objections I have to readability formulas is that even those formulas developed and validated for adults give readability assessments in terms of school reading grade level. To say a certain document is written at an RGL of six is to say that it is suitable for an adult reading at that level. This implies that this adult is quite similar cognitively, verbally etc. to the typical sixth grade child. Yet, Sticht and James (1980) have shown that adults and children reading at the smae grade level on some reading test are very different in many cognitive and verbal skills. In the absence of clear ideas as to what the skills of an adult reading at a certain level are, the identification of these skills with a school grade appears misleading. Furthermore, the assignment of an RGL to an adult varies widely depending on the reading test he or she is tested with. For low skilled adult readers in a military population, we (Sticht, Hooke & Caylor, 1981) have found differences of more than three grade levels among commonly used adult reading tests. Thus, a text measured at the sixth grade level, even if all other assumptions are met, may not "match" the skills of an adult sixth grade reader if different tests were used to test the individual and in validating the formula

The third aspect of the derivation of readability formulas I object to is that they assign readability scores to texts based on performance of readers on paper and pencils tests created from similar texts. However, in the real world, particularly in the military, documents are written for particular operational purposes, not just for "understanding", as measured by CLOZE or other paper and pencil test performance. Thus, if criteria for readability are to be used at all they should be based, ncton how well people are predicted to fill in blanks on CLOZE tests, but on how well they are predicted to fulfill the actual purpose of the document, e.g., follow procedural steps correctly. The only way to make these predictions would seem to be to use such perfo rmances as criteria when the formulas are being developed. This is especially important because evidence is beginning to accrue (e.g., Wright & Reid, 1974) that the text features most conducive to verbal understanding are not always the same as the ones most conducive to correct operational performance.

One issu e that complicates matters for military writing is the question of technical words. Many commonly written about objects, administrative entities, functions, documents etc. have official names which are fatally prone to being polysyllabic and not on Dale-Chall's list. There is a long series of these for each service and many specifc ones for each specific job. The need

to use such terms raises the readability scores for military documents above what such scores would be if military technical terms were shorter or more generally familiar. To compensate for such words and phrases and still attain a low readability score, a military writer may feel he has to write the rest of his text at virtually the level of an elementary school reader. When Air Force writers were required to attain low readability scores for their texts, they persistently requested to be allowed to count field specific technical terms as if they had the formula measured characteristics of much easier words, i.e., as if they were monosyllabic. Writers claimed the right to do this on the grounds that these technical words were highly familiar and thus very easy to even the poorer readers in relevant fields. We, at AFHRL, decided to test this claim that familiar technical terms are as easy as short, familiar general words. To do this, we asked the writers to submit lists of the technical words in their documents which they claimed were so familiar to relevant field members as to warrant changing their formula count. These same writers were also asked to submit definitions for the selected terms. We then took the words and their definitions, made them into vocabulary tests and administered these tests to members of the appropriate fields. We found that, in all but a few fields, people performed surprisingly poorly with these supposedly very familiar terms. We thus found no support for the writers' claims. However, this did not make the writers' task any easier.

Related to the issue of technical words in the measurement of the readability of military documents is the even thornier question of what to do with acronyms. At least technical words are words and can be treated as such. With acronyms the natural relationship between number of syllables and difficulty or familiarity is distorted. Yet if one counts acronyms as they are pronounced, in computing the value of a readability measurement, one is assuming that this relationship holds. Any decision that is made about how to count acronyms is arbitrary. Yet, as anyone who has tried to compute the readability score of military writing knows, scores of such arbitrary decisions have to be made for every passage.

Even if readability formulas were agreed to produce acceptably accurate estimates of the readability of texts, it would still be a separable question whether requiring texts produced in an organization to have some specified readability formula score is a useful and/or effective way to improve the readability of these texts. Yet military organizations have made and do make such requirements (Cf. Air Force Regulation 5 -1 and MIL-M-38784A). It should be noted that everyone who deals with readability formulas, including their proponents, agrees that they are not to be written to. However, if acceptability of documents is made contingent on even a leniently applied standard related to formula score, what else is the poor writer to do? Writing to formula, ignoring all else, as a way to improve document readability is analogous to trying to improve children's reading ability by stretching them, because one has noticed that taller children read better than smaller ones. Even more enlightened, substance-oriented rewriting, with formula parameters kept in mind may tend to distract a writier from his main task of clear and adequate expression and communication.

Our experience in the Air Force was that, completely aside from whether requiring documents to have certain formula determined readability levels improved their readability, writers simply did not comply with the requirement. Thus, the majority of writers who had to submit documents written below about the 10th grade level according to the FORCAST readability formula did not do so, although they certified that they had. It is difficult to see how such a situation could possibly have any desirable effect on document readability, but equally difficult to blame the writers.

The readability formula the Air Force writers we were working with were required to use was the FORCAST. This is the simplest of all formulas to use since its value depends only on proportion of one syllable words in a text sample and its only constants are 10 and 20. Virtually every other readability formula requires much more complex text feature counting and arithmetic computation. To take the burden of these cmputations off writers, editors and their stafrs, as well as to assure that they are done correctly, computer programs have been developed to calculate the values of readability formulas for texts. Since the features of texts used in readability assessment are necessarily quantifiable, computers , using simple software,are very good at identifying, counting and calculating values based on them. I believe that the already existing capacities of these programs, if properly used and adapted, can form the basis of interactive systems which might actually serve to improve document readability.

When computer readability measurement programs first came out, I suspect they were simply intended to relieve people of the burden of arithmetic calculation. If their use is limited to just this, clearly they are subject to the same criticisms for use in the military as are use of readability formulas in general. However, as anyone who has played with one of these programs can tell you, the same routines written to calculate readability formula values, i.e, those that count words per sentence, count syllables, match words to those on a word list, can with rather little effort, be made to perform other services for the writer. For example, a function which tells you that a word is not on the Dale Chall list will also pick up misspelled words for you. A routine which lists all words over a certain number of syllables in a text, can also be made to suggest synonyms for certain of these words (Kincaid et al, 1980).

When such functions begin to appear, the program moves from readability measurement into an area which can be called *writers'* aid and which may incorporate some wisdom about good writing found in writer's guidelines and not available from formulas. However, simple programs of this type, although very good at what they do(counting, identifying, matching) are limited and cannot really make the kind of context dependent judgment about appropriateness, clarity, meaningfulness etc. which are required if a poorly written document is to be improved in readability. Such decisions still appear to require human judgment, even creativity.

Another thing which may cause the kind of programs I am talking about to be limited in usefulness is that they typically deliver their advice "off line", after a text is already written and has been typed into the computer for the purpose of assessment, not delivered in the context of the writing task itself. If anything at all is known about teaching problem solving (and good writing is problem solving par excellence), it is that particular suggestions are most effective when given at exactly the moment they are perceived as needed. Otherwise, they may simply be ignored as of no relevance. This may well be why writer's guidelines, handbooks etc. don't appear to work very well.

I believe the way to get around both these shortcomings of computer readability programs is to create an interactive writer's aid program interfaced with a text editor or word processor and continuously available on demand to the writer as he develops his text. If this is done properly, I believe it has the potential for creating not only better documents, but perhaps even better writers. I think it should be clear how such a program could provide immediately relevant advice on improving the features of a particular text. It may be somewhat less clear how such programs could get around the inherent limitations of the mechanical functions a computer can perform when compared to the complex, meaning dependent and context sensitive nature of writing.

As I have said before, the most important criticism of readability metrics is that they sidestep the substantive semantic and communicative aspects of writing. In revising a document to improve its readability, a writer should constantly make judgments about such matters. These decisions are beyond the state of the art of even the most advanced computer programs. (The most advanced of the genre probably are those developed by Bell Labs.) However, in an interactive system the computer is only called upon to do some of the work. The decision making capacity of the writer can be enlisted for the other part. The computer program's capacities can be used to provide input to the writer's decisions. This input can serve, for example to remind him of the particular relevance of certain principles or guidelines, make text aspects clearer by presenting them in relief or check for consistency (e.g., if a word is replaced by the writer, the program can present all other instances of that word for possible replacement), freeing the writer's resources to attend to more substantive aspects of text revision.

Before this discussion gets too abstract, let me give two examples of the kind of thing I mean. I know these are programmable because a colleague and I have programmed them, although the system she created is now defunct. I hope these examples show how guidelines for good writing can be incorporated into such a computer program without the necessity for readability measurement and also how the computer's capacities and the writer's can be used to complement each other.

The first example uses the principle of the "text skeleton". A message was available which briefly explained that too many negative words or phrases in a text could frequently impede readability, although in any given case, how many was too many was a matter of judgement. The writer then had the option of seeing a skeleton of his text with all words but negatives replaced by blanks. This made obvious where in the text negatives were thickest and gave the writer the option of revising those passages. The writer would additionally be able to add or delete negatives from the list the computer was using, allowing the program to learn from the writer, as well as vice versa. When a writer changed or replaced a negative, he could have the option of storing that replacement if he thought it effective. Lists of such previous successful revisions could then be made available in the context of the "negative" routine. Similar routines were being developed for subordinate clauses and verbal and other derived nouns.

My second example relates to an even less measurable aspect of text, that of text organization and movement of ideas. Again the computer routine is simply a "trick" which is designed to get the writer to consider certain principles in relation to his particular text and to enlist his knowledge and judgment in deciding whether and how to revise this text. A statement was available to the writer which described how paragraphs aided topic progression and how the first sentence in each paragraph usually provided a statement of that paragraph's topic. The writer could then choose to see a list of his topic sentences either in narrative or outline form, so he could more easily ascertain whether the topics progressed in a coherent manner and presented his ideas effectively. I hope these two examples serve to suggest the possibilities computer programs have for improving as well as measuring document readability.

Hooke, L.R., De Leo, P.J. & Slaughter, S.L. Readability of Air Force Publications AFHRL-TR-79-21,Lowry AFB, CO: Air Force Human Resources Laboratory, Sept, 1979.

Kincaid, J.P., Aagard, J.A. & O'Hara, J.W. Development and Test of a Computer Readability Editing System (CRES),TAEG report No. 83, March, 1980, Orlando, FL.

Sticht, T.G. & James, J.H. Auding and Reading Skills in Children and Marginally Literate Men. In HumRRo Professional Paper-10-78, Alexandria VA: 1978

AD P001405

# COMPUTER AIDS FOR AUTHORING TESTS

J. Peter Kincaid and Richard Braby
Training Analysis and Evaluation Group
Orlando, Florida

and

Wallace H. Wulfeck, II
Navy Personnel Research and Development Center
San Diego, California

## ABSTRACT

Computer routines developed to help in authoring and editing textual training materials were modified to aid authors of tests. The Navy's Computer Readability Editing System (CRES) aids in producing comprehensible text by flagging uncommon words and awkward sentences, suggesting replacements for awkward words or phrases, and giving the readability grade level. Additional routines were developed, based on the Instructional Quality Inventory, specifically for multiple choice and true/false test questions. These new routines calculate readability grade level of test questions, and flag some kinds of awkward or incorrect test item construction. The CRES routines, including the new test item features, are intended to be used as part of a computer-based publishing system. Our initial effort to provide feedback to authors of tests has convinced us that the general approach is viable and many new useful features could be added.

Computer routines to aid authors in developing tests are now feasible. Some routines have already been developed. For example, Roid and Finn (1978) have demonstrated routines for generating multiple choice test items from text passages. Additional routines to aid authors in test items would complement the Roid and Finn routines, or could be used separately.

## COMPUTER READABILITY EDITING SYSTEM (CRES)

This paper describes a new version of the Computer Readability Editing System (CRES) modified to provide aid in authoring tests. Like the original version of CRES, it is designed to make written material easier to understand. The original version of CRES, developed specifically for text, is documented in Kincaid, Aagard, and O'Hara (1980); Kincaid, Cottrell, Aagard, and Riseley (1981); Kincaid, Aagard, O'Hara, and Cottrell (1981); and Braby and Kincaid (1981-82).

The basic configuration of the original CRES is shown in the flow chart contained in figure 1. The steps depicted for using the CRES as part of a computer-based publishing system are:

- Choose program options.
- Author or typist enters text (typically 500 to 2,000 words).
- Text is analyzed and printed out and shown on the display.
- Author revises text prompted by computer-generated suggestions (see figure 2).
- Revisions are entered and stored.
- Revised text is again analyzed by the computer to obtain readability grade level and check for keying errors.
- Editor approves text which is then stored for final camera-ready printout (for example, using a daisy wheel printer).

The basic features of the original system include those which:

- Flag uncommon words - those not on a carefully prepared list of 4,300 common words or a series of supplementary technical word lists, each about 100-200 words.
- Flag long or awkward sentences - those with passive voice or double negatives.
- Suggest replacements for awkward words and phrases.
- Provide the readability grade level according to the Department of Defense standard, the Flesch-Kincaid formula.

In addition, the system flags misspelled words.

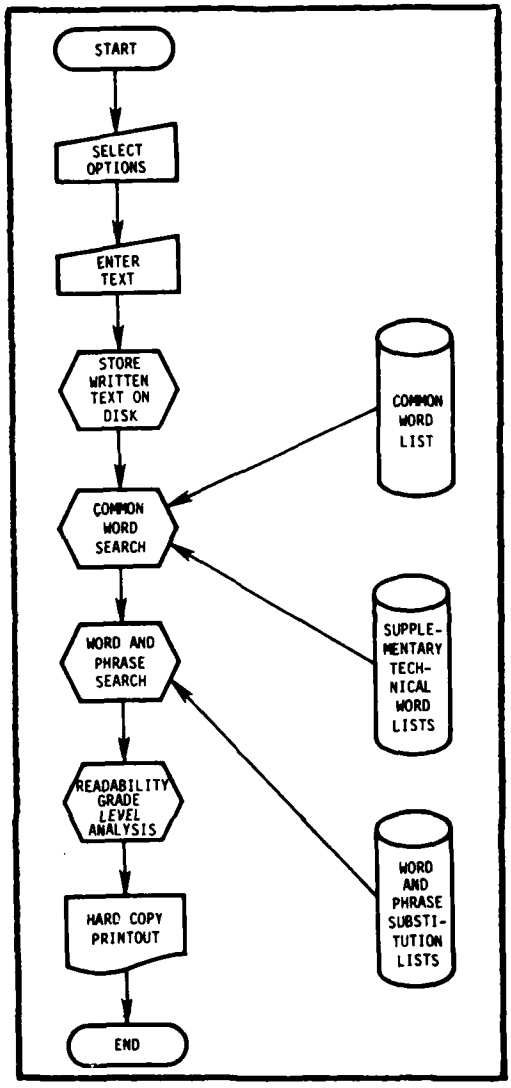



Figure 1. Flowchart Showing Phases of Operation of the Computer Readability Editing System (CRES)

## INSTRUCTIONAL QUALITY INVENTORY (IQI)

The Instructional Quality Inventory (IQI) is a set of procedures for quality control of instructional development, designed to parallel and supplement the Instructional System Development (ISD) process. The IQI procedures can also be used to evaluate existing instruction, and can be used as evaluation or acceptance tools for instructional programs obtained through contract. A series of Navy reports document the IQI (e.g., Ellis, Wulfeck, Merrill, Richards, Schmidt, and Wood, 1978).

Ellis, Wulfeck, and Fredericks (1979) describe a series of steps in using the IQI:

- Classification. The IQI procedures are based on a system for classifying objectives, test items, and instructional components. Classification is determined according to: (a) what the student is required to do with the information to be learned, and (b) what type of information the student is learning.

- Assuring Objective Adequacy. Since good instruction depends on careful specification of learning objectives, the first IQI procedure is to assure the adequacy of objectives. This is done by classifying each objective, and judging whether or not it accurately reflects the intended student performance after training.

- Constructive, Consistent, and Accurate Tests. The next IQI step is to make sure that tests accurately measure progress toward the objectives. This is done by assessing the consistency between each test item and its associated objective, and the adequacy of the item. Essentially, each test item must be classified in the same way as its objective and must be adequately constructed.

- Keeping Presentations Consistent. Instructional presentations contain various components, including statements of material to be learned, examples, and practice. For consistency, different combinations of presentation components are required depending on the classification of the objective.

- Applying Adequate Learning Principles. The final IQI step is to make sure that each required presentation component is adequate according to psychological principles of learning.

The IQI is, among other things, a checklist for preparing quality tests.

This paper describes the incorporation of a number of these items into the CRES, those which could easily be programmed.

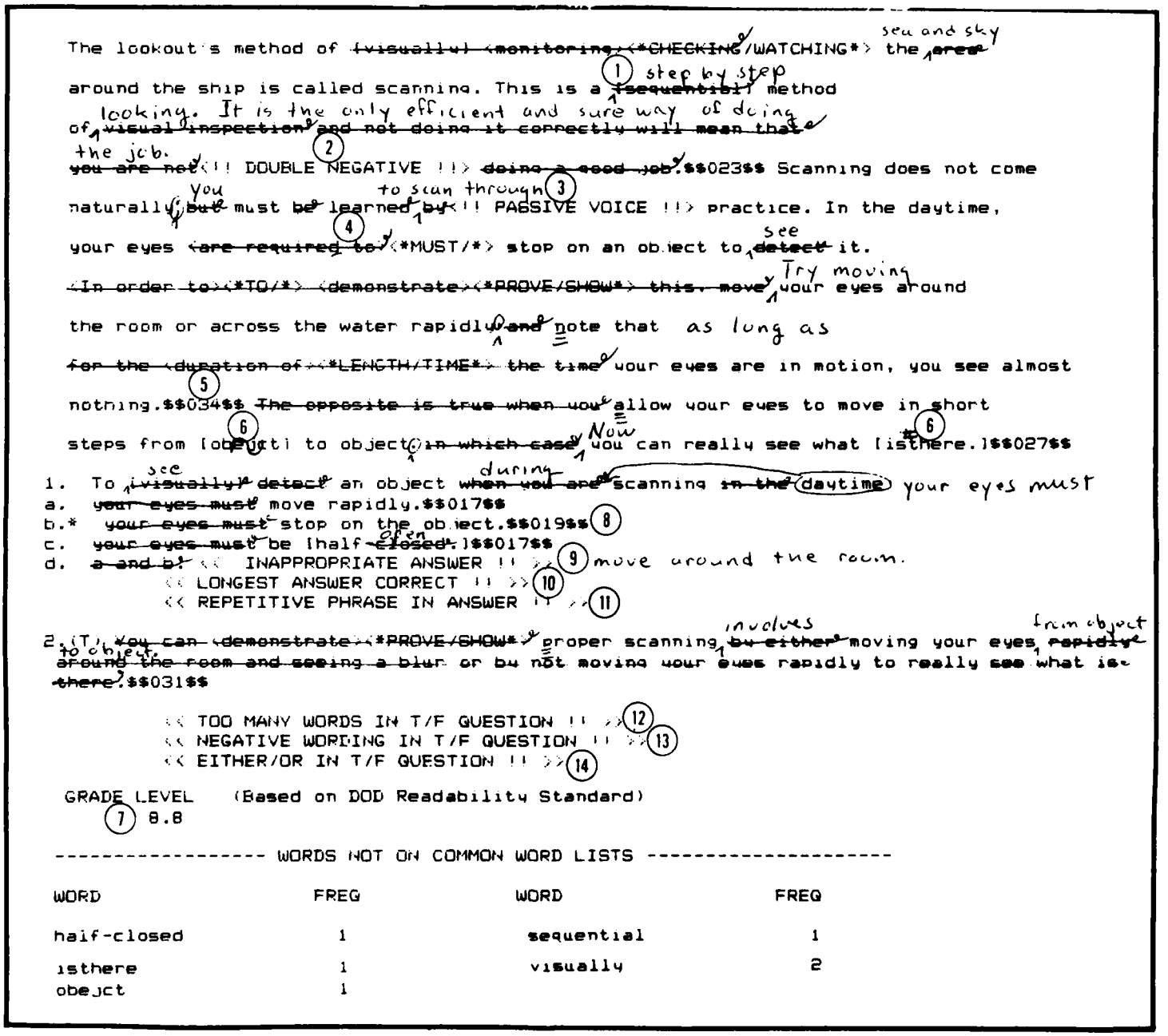
## EXPANDED CRES FOR TEST ITEM ANALYSIS

Figure 2 shows a CRES analysis and illustrates both the original features of the CRES (for analyzing text) and new features for analyzing test questions. It also shows handwritten editing changes suggested by the computer analysis. The passage and questions are intended for a Navy remedial reading workbook for sailors with no higher than sixth grade reading ability.

Features of the CRES designed for text are indicated by callouts 1-7.

1. Uncommon words are flagged: in this case "sequential."

2. Double negatives are flagged: in this case "not...not."



The lookout's method of ~~[visually] <monitoring> (*CHECKING~~/WATCHING*> the sea and sky ~~area~~ around the ship is called scanning. This is a (1) ~~[sequential]~~ step by step method of ~~visual inspection~~ looking. It is the only efficient and sure way of doing the job. ~~and not doing it correctly will mean that you are not~~ (2) <!! DOUBLE NEGATIVE !!> ~~doing a good job.~~$$023$$ Scanning does not come naturally, ~~but~~ you must ~~be~~ learned ~~by~~ (3) to scan through <!! PASSIVE VOICE !!> practice. In the daytime, your eyes ~~are required to~~ (4) (*MUST/*> stop on an object to ~~detect~~ see it. ~~<In order to> (*TO/*> <demonstrate> (*PROVE/SHOW*> this, move~~ Try moving your eyes around the room or across the water rapidly ~~and~~ note that as long as ~~for the <duration of> (*LENGTH/TIME*> the time~~ your eyes are in motion, you see almost nothing.$$034$$ (5) ~~The opposite is true when you~~ allow your eyes to move in short steps from [object] to object. ~~in which case~~ Now you can really see what [is there.]$$027$$ (6)

1. To ~~[visually] detect~~ see an object ~~when you are~~ during scanning ~~in the~~ (daytime) your eyes must
a. ~~your eyes must~~ move rapidly.$$017$$
b.* ~~your eyes must~~ stop on the object.$$019$$ (8)
c. ~~your eyes must~~ be [half-closed.]$$017$$ often
d. ~~a and b!~~ << INAPPROPRIATE ANSWER !! >> (9) move around the room.
<< LONGEST ANSWER CORRECT !! >> (10)
<< REPETITIVE PHRASE IN ANSWER !! >> (11)

2. (T) ~~You can <demonstrate> (*PROVE/SHOW*~~ proper scanning involves ~~by either~~ moving your eyes from object to object. ~~rapidly around the room and seeing a blur~~ or by not moving your eyes rapidly to really see what is ~~there~~?$$031$$

<< TOO MANY WORDS IN T/F QUESTION !! >> (12)
<< NEGATIVE WORDING IN T/F QUESTION !! >> (13)
<< EITHER/OR IN T/F QUESTION !! >> (14)

GRADE LEVEL (Based on DOD Readability Standard)
(7) 8.8

---------- WORDS NOT ON COMMON WORD LISTS ----------

| WORD | FREQ | WORD | FREQ |
|---|---|---|---|
| half-closed | 1 | sequential | 1 |
| isthere | 1 | visually | 2 |
| obejct | 1 | | |

Figure 2. Original Test Showing CRES Analysis (With Features Noted by Callouts, and Editor's Handwritten Changes)

3. Passive verbs are flagged: in this case "be learned." A passive verb is composed of a form of the auxiliary verb "to be" plus a past participle. Language experts agree that the active verb (in this case "learn") is generally easier to understand.

4. Replacements for awkward words and phrases are suggested: in this case "must" is suggested as a replacemnt for "are required to."

5. Long sentences are flagged and the number of words in the sentence is shown between dollar signs: in this case "$$34$$."

6. Keying errors and misspelled words are flagged: in this case "obejct" and "isthere." These are listed as WORDS NOT ON COMMON WORD LISTS.

7. The readability grade level calculated according to the Department of Defense readability standard, the Flesch-Kincaid formula, is shown: in this case "8.8." This grade level is for both the passage and the two test questions. The original version of CRES gives grade level only for text. Readability grade level for multiple choice questions is calculated by using the question stem as the beginning of the sentence for each alternative.

New features of the system, suggested by the IQI, specifically analyze test questions. They are indicated by callouts 8-14.

8. Long sentences in multiple choice items are flagged. The words in both the question stem and answer are counted as a single sentence and the number of words shown between dollar signs: in this case "$$19$$."

9. Inappropriate answers to a multiple choice question are flagged: in this case "a and b."

10. If the longest answer to a multiple choice question is the correct answer it is flagged. Test-wise students use this as a clue.

11. Repetitive words or phrases in answers to multiple choice items are flagged: in this case "your eyes must." If the same word or phrase starts each alternative in a multiple choice test item, it should be moved to the stem.

12. Long true/false test questions are flagged.

13. Negative wording in true/false test questions is flagged: in this case "not."

14. Certain complex true/false test questions are flagged: those containing "either...or," "neither...nor," and "or": in this case "either...or."

Figure 3 shows the revised test. It is improved in many ways compared with the original version. Not only is readability grade level reduced (from grade level 8.8 to 4.5), but the text is easier to read and several errors have been removed from the test questions.

The lookout's method of watching the sea and sky around the ship is called scanning. This is a step by step method of looking. It is the only efficient and sure way of doing the job. Scanning does not come naturally; you must learn to scan through practice. In the daytime your eyes must stop on an object to see it. Try moving your eyes around the room or across the water rapidly. Note that as long as your eyes are in motion, you see almost nothing. Allow your eyes to move in short steps from object to object. Now you can really see what is there.

1. To see an object during daytime scanning your eyes must
a. move rapidly.
b.* stop on the object.
c. be half open.
d. move around the room.

2.(T) Proper scanning involves moving your eyes from object to object.

Figure 3. Revised Test (Grade Level is 4.5)

## CONCLUSIONS

The extension of the CRES routines to aid in the development of test items appears useful. It should be noted, however, that the present effort was simply a demonstration. Several dozen automatic checks of IQI items could be added using current equipment and without a major difference in the type of computer algorithms already employed.

One entirely different kind of computer check could significantly increase the scope of computer-assisted authoring. Whereas the features described in this paper are automatic, a computerized IQI checklist could be added to the system as suggested by Spannaus (1980). For example, if the objective of a lesson is to learn nomenclature, the computer could ask the author, "Is a memory aid appropriate?" This kind of query is easier to program than the automatic checks described in this paper.

The routines described in this paper merely illustrate the value of the use of computers in the development of instructional material. We can expect many more such developments in the near future.

## REFERENCES


Braby, R. and Kincaid, J. P. "Computer Aided Authoring and Editing," Journal of Educational Technology Systems, Vol. 10 (2), 1981-82. (In press)

Ellis, J. A., Wulfeck, W. H. II, and Fredericks, P. S. The Quality Instructional Inventory II. User's Manual. NPRDC Special Report 79-24, 1979. Navy Personnel Research and Development Center, San Diego, CA 92152.

Ellis, J. A., Wulfeck, W. H. II, Merrill, M. D., Richards, R. E., Schmidt, R. V., and Wood, N. D. Interim Training Manual for the Instructional Quality Inventory. NPRDC Technical Note 78-5, 1978. Navy Personnel Research and Development Center, San Diego, CA 92512.

Fredericks, P. S. The Instructional Quality Inventory III. Training Workbook. NPRDC Special Report 80-25, 1980. Navy Personnel Research and Development Center, San Diego, CA 92152.

Kincaid, J. P., Aagard, J. A., and O'Hara, J. W. Development and Test of a Computer Readability Editing System (CRES). TAEG Report No. 83, 1980. Training Analysis and Evaluation Group, Orlando, FL 32813.

Kincaid, J. P., Cottrell, L. K., Aagard, J. A., and Risley, P. Implementing the Computer Readability Editing System (CRES). TAEG Report No. 98, 1981. Training Analysis and Evaluation Group, Orlando, FL 32813.

Kincaid, J. P., Aagard, J. A., O'Hara, J. W., and Cottrell, L. K. "Computer Readability Editing System," IEEE Transactions on Professional Communications, Vol. 24, pp. 38-41, March 1981.

Roid, G. and Finn, P. Algorithms for Developing Test Questions From Sentences in Instructional Materials. NPRDC Technical Report 78-23, 1978. Navy Personnel Research and Development Center, San Diego, CA 92152.

Spannaus, T. W. "Speculations on Computer Assisted Design of Instruction." Paper presented at the National Conference on Computer Based Education, Minneapolis, MN, October 1980. (ERIC #ED200215)

Paper presented at the 23rd Annual Conference of the Military Testing Association (MTA), October 20-28, 1981, Alexandria, Virginia.

# A Literacy Task Inventory for Identifying Literacy Skill Levels of Jobs

Thomas G. Sticht

Human Resources Research Organization

AD P001406

## Summary

Identification of the types of literacy tasks people perform on jobs permits the inclusion of samples of such tasks in various selection and classification instruments to more validly represent the cognitive skill demands of jobs. Also, the identification of job literacy tasks is needed to develop literacy training programs that more faithfully reflect the requirements for literacy in job training and at job sites. Specification of general skill levels needed for different jobs is useful for management purposes in establishing goals for selection and training, and for summative assessment of literacy programs. Research is reported to develop a literacy task inventory that provides both a specification of literacy tasks and an estimate of the general literacy difficulty level needed to perform these tasks. The method can also be used to establish the readability of textual, graphic and tabular materials, as well as combinations of these displays that appear in job literacy tasks.

## Introduction

Analysis of written language indicates that it differs from spoken language in three major features; the written language is more or less permanent, it uses the properties of light (color; contrast), and it can be arrayed in visual space. These features make possible the use of formats for communicating via graphic symbols that go beyond the words and sentences and connected discourse common to both speech and writing. Figures, charts, and tabular arrays of data are examples of such formats.

Traditional measures of readability that have been used to estimate literacy demands of jobs (Sticht and Zapf, 1976) use word difficulty and sentence length factors to estimate the reading difficulty level of job reading materials. Yet, even cursory examination of the types of reading materials (called "displays" in this report) used by people in work settings reveals that textual materials constitute only a small amount of what they must read. For this reason, the present research, conducted for the Navy, aimed to identify the types of reading tasks personnel performed in Navy jobs and to develop a methodology for scaling the reading difficulty of the types of tasks (not just materials) performed. A detailed discussion of the methodology, and a critique of several different methods for estimating reading requirements of jobs is presented elsewhere (Sticht, et.al., 1976). Here I will briefly summarize the outcomes of our efforts, and discuss certain methodological difficulties inherent in any attempt to define literacy demands of jobs, and hence in establishing levels of literacy for selection, classification and training decision making.

## Identifying Job Reading Tasks

In structured interviews with some 130 personnel in ten Navy ratings, we identified two major classes of reading tasks: reading to do something and reading to learn something. For job performance, reading to do tasks,

in which information needed in accomplishing some job task is looked-up, applied, and then forgotten constituted the bulk of the reading tasks, and hence our further analyses were performed only on reading to do tasks.

The next step in our procedure was to identify (1) the type of information sought in performing each task, and (2) the type of display in the reading materials, classified as either text: which would be written language; figures: including line drawings, photographs, schematic diagrams, etc.; tables: including both numerical and verbal tabulations; texts plus figures; texts plus tables; and tables plus figures.

This analysis revealed that the type of information sought was usually some type of factual data, or the person was trying to find out how to do something. Thus, categories of skills called fact finding and following directions were identified.

Analysis of materials by display types revealed that the combination of tables plus figures was only very rarely used, hence this type of display was not used in the subsequent research. Texts constituted the most frequently used type of display and made-up about one-third of the display types, with figures running a close second at somewhat less than 30% of the display types. Tables, texts plus figures, and texts plus tables fell in that order of frequency of occurence behind texts and figures.

There were differences among types of jobs in the relative frequency of uses of displays, with technical maintenance jobs using proportionately more figures and data oriented jobs using figures and tables to about the same extent.

## The Reading Task Inventory

By means of the classification system outlined above generic reading tasks were defined as the application of either fact finding or following directions skills to texts, figures, tables, text plus figures combined, and text plus tables combined. Tasks comprised of the two skills applied to the five display types were found in all ten jobs. They therefore represent, at an abstract level, the types of Navy reading tasks Navy personnel perform in the course of doing a job. In an abstract manner, this analysis answers the question: What are the reading tasks people have to perform in various jobs in the Navy? We can answer, they look up facts in texts, they look up directions in texts, they look up facts in figures, they look up directions in figures, etc.

Conceivably, we could develop an inventory by simply asking people whether they look up facts in texts, figures, tables, etc. In fact, in work for the Department of Manpower and Immigration in Saskatchewan, Smith (1975) and associates used a somewhat similar inventory approach in which they attempted to discover both what kinds of materials were read in a number of occupations (e.g., notes, memos, letters, directions, instructions, policy manuals, and the like) and what reading tasks are performed in those jobs (e.g., read to locate facts, to follow directions, to discover the main idea, etc.). To obtain this information, interviewers at times showed displays of the general type of material they were talking about. For instance, in determining if a given job required the reading of graphs, two graphs were shown as examplars and interviewees were asked to indicate whether they read similar graphs in performing their jobs.

A problem with the inventory approach in which people simply indicate whether they read some type of material is that it fails to distinguish among complexities of materials, and it provides no indication of the level of general reading skills required to perform the set of reading tasks in a given occupation.

To overcome these difficulties in the Navy research, a reading task inventory was constructed which included three levels of complexity for each type of display, i.e., texts, figures, tables, etc. The displays were taken from the Navy's Bluejacket's Manual. This 617 page manual is the basic manual for new Navy recruits. Therefore it is meant to be read using only general reading skills and knowledges, and its content is familiar to all Navy personnel. These features are important because a primary type of information desired for occupational reading requirements is "Data on the level of reading skills required to have access to the occupations". (Miller, 1974). Since The Bluejacket's Manual is an entry level manual, it represents the type of material that one must be able to read to have access to all Navy job training and occupational fields.

To develop an inventory that we could use to identify the kinds of reading tasks people perform, and the general level of reading skills needed to perform those reading tasks, we searched The Bluejacket's Manual to locate three concrete instances of each of the five abstract categories of generic reading tasks identified above. Three examples of each generic reading task permited us to use three levels of complexity for each reading task. These levels were confirmed by two judges.

Figure 1 provides an example of the type of display included in the inventory. On the left hand side is a sample of text plus table material from The Bluejacket's Manual. On the right hand side are the inventory questions. This particular page from the inventory is for fact finding, so job incumbents are asked: In your job would you ever have to perform reading tasks using material like this to look up facts? If they said yes, then they were asked questions about the frequency of performance, and then questions about the consequences of making a reading error in performing this kind of reading task. These data are used to make decisions about the criticality of the reading task.

To identify the general level of literacy required to perform each reading task, we wrote fact finding and following directions questions for each of the display types in the inventory. These job reading tests were administered to a sample of Navy recruits. Additionally, they were administered a standardized reading test. With these two sets of data, we could then determine how well young adults of differing reading skills could perform the job reading task test items.

Figure 2 shows the results of asking a following directions question using the same material as shown in Figure 1 as a fact finding inventory item. This type of display shows the job reading material on the left side of the page, and presents the type of task, the form (in this case E for easy), the question, and the test results, i.e., the percentage of personnel at each general reading grade level who got the correct answer to the reading task test item. In the case of Figure 2, we see that 10 persons read at the 6th grade level, and 40% of them got the answer to the question correct, using the material on the left side of the page (in the actual text, the material was in The Bluejacket's Manual. Thus people had to locate the material in the 617 page manual. They were given page numbers. By using

the intact Bluejacket's Manual we hoped to obtain a greater fidelity to the actual job reading situation).

At the bottom of the right side of the page the results of the use of the material in the inventory format is presented. For our exploratory study, only four persons from four jobs tried out the inventory. The results show differences in the reported frequency of use by these four personnel for this type of material. Obviously, large numbers of personnel are needed to obtain a reliable, normative view of the performance of various reading tasks in different jobs.

To identify the reading demands of any Navy job using this inventory approach, one would first administer the inventory to job incumbents to determine frequency and criticality of performance of each reading task. Then, to determine the reading grade level of difficulty for each type of reading task in the inventory, the job analyst would consult expectancy tables which show how well people of differing reading grade levels perform the reading task. At this point, a management decision must be made about what percentage of people should be able to perform the reading task. If it is determined that only 40% of the people should be able to perform the task, then, using the example of Figure 2, a 6th grade level of reading skill would suffice, and the task would be assigned a value of 6th grade level of difficulty. However, if it were determined that 80% of the people should be able to perform the task, then in the example of figure 2, it is at the 8th grade level where 80% of the persons first get the item correct (it is assumed that with larger numbers of persons taking the test, the fluctuations in the percentages correct as a function of reading skill level would be greatly reduced). This would then cause the item to be assigned an 8th grade level of difficulty, in effect, its readability score.

To determine the reading difficulty for a job, the reading grade level of each reading task indicated as being performed on the job in the inventory of job reading tasks is weighted by its frequency and criticality. These weighted figures are summed and the average weighted reading difficulty level is computed. The resulting average reading grade level is the level of general reading skill that is needed, on the average, to perform the reading tasks of a given job.

## Critique of the Reading Inventory Approach

In any approach to the development of an assessment instrument to evaluate the skill/knowledge levels required for successful performance in a domain, a variety of methodological and procedural problems are encountered. The experimental development of the Navy job reading task inventory is no exception. A discussion of some of these problems may be instructive for others who would set out to identify reading demands or "minimal competencies" for vocational literacy.

Key requirements of the inventory are that job incumbents respond to the generic aspects of a task display rather than to the specific content; that they respond to the levels of complexity; and that they respond to the distinctions between fact finding and following directions. In the research described, however, no good basis was established for assuring that the levels of complexity or types of uses (fact finding; following directions) actually entered into the interviewee's responses. There was evidence that indicated that, of the four people who tried out the inventory, two did not respond only to the generic aspects of the displays, but rather they responded in part to the specific content. In the work by Smith (1975)

this was not reported to be a problem. But neither was it detectable because of the methodology used in that study.

Concerning the job reading task test, aside from the technical problems involved with some items which could be remedied by careful redesign of questions, a major problem is in knowing how close the reading task questions approximate the real job reading tasks. It may be that the reading test imposes unrepresentative information processing demands which are not involved in the real life execution of job reading tasks. Indeed, the most difficult question to answer is that of the validity of the reading inventory/test as a measure of the reading demands of jobs. Is there any way to be certain that this entire procedure presents a valid estimate of the reading demands of jobs? This raises the question, how would we know? It may be easier to know that an approach is not valid than to know that one is. For instance, the approach in which a readability formula is applied to a sample of text materials is an empirical method for determining the average reading difficulty level of materials. However, the question of validity seems clear with the readability approach: it in no way involves figures and tables in the estimate; and, as we found in the Navy research, only some 30% of the reported reading tasks involved texts only; two-thirds used figures or figures plus texts combined.

In a review of some seven different approaches for estimating the reading demands of jobs, Sticht & McFann (1975) show that all seven approaches provide different estimates. Indeed, the very definition of a reading task differs from one to the other approach. From the present discussion, and the analysis of Sticht & McFann, it should be apparent that there is no such empirical "condition" or "event" or "thing" known as "the reading demands of a job". Hence, there is no one "true" way to establish "minimal" competency levels of literacy for vocational preparedness. Reading demands of jobs can only be estimated by procedures which are more or less systematic and performed according to more or less specifiable rules. The question of the validity of any estimate can only be answered in respect to a model or theory of job-related reading which would define systematic procedures for obtaining estimates of reading demands of jobs for various purposes permitted by the theoretical constructs involved.

It should be noted, however, that the foregoing problems of validity are not specific to the determination of job reading requirements. Indeed, they permeate all aspects of job and task analysis, and all psychometric approaches to the evaluation of skills and knowledges in any domain of activities. Within these limits, and they are imposing and formidable limits, which ought to be conducive of humility amongst psychometricians, job and task analysts, and educators, I believe that the inventory approach can, with refinements, offer useful information about the reading demands of jobs.

## References


Miller, G. (Ed.) Linguistic Communication: Perspectives for Research. Newark, Del.: International Reading Association, 1974.

Smith, A. Generic Skills for Occupational Training. Prince Albert, Saskatchewan: Training Research and Development Station, 1975.

Sticht, T. and McFann, H. Reading requirements for career entry. In: D. Nielsen and H. Hjelm (Eds.) Reading and Career Education. Newark, Del.: International Reading Association, 1975.

Sticht, T. and Zapf, D. (Eds.) Reading and Readability Research in the Armed Services. HumRRO FR-SE-CA-76-4. Alexandria, Va.: Human Resources Research Organization, September 1976.

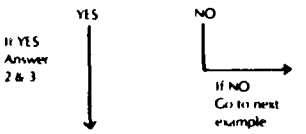
Sticht, T., Fox, L., Hauke, R., and Zapf, D. Reading in the Navy. HumRRO FR-WD-CA-76-14. Alexandria, Va.: Human Resources Research Organization, May 1976.

All ships are assigned designations – a group of letters which indicate their type and general use – and *hull numbers*, which are usually assigned in sequence to ships of a type as they are built. These identifying designations are used in correspondence, records, plans, communications, and sometimes on ships' boats, because letter and number designations are shorter than the ship's name – *Mission Capistrano* (AC 162) – and help to avoid confusion between such similar names as *Home* (DLG 30) and *Hornet* (CVS 12) or *Phoebe* (MSC 199) and *Phoebus* (YF 294).

The first letter in a designator is a general classification: *D* for destroyers, *S* for submarines, *L* for amphibious vessels, *M* for minewarfare vessels, *A* for auxiliaries, *W* for Coast Guard vessels, *T* for Military Sealift Command ships, and *Y* for service and yard craft. In combatant designations, the letter *N* means nuclear powered and *G* means the ship is equipped to fire guided missiles. A listing of most ship designations follows; minor yard craft and service craft have been omitted.

| | |
|---|---|
| AD | Destroyer Tender |
| ADG | Degaussing Ship |
| AE | Ammunition Ship |
| AF | Store Ship |
| AFS | Combat Store Ship |
| AG | Miscellaneous |
| AGDE | Escort Research Ship |
| AGEH | Hydrofoil Research Ship |
| AGER | Environmental Research Ship |
| AFG | Miscellaneous Command Ship |
| AGM | Missile Range Instrumentation Ship |
| AGMR | Major Communications Relay Ship |
| AGOR | Oceanographic Research Ship |
| AGP | Patrol Craft Tender |
| AGS | Surveying Ship |
| AGSS | Auxiliary Submarine |
| AGTR | Technical Research Ship |
| AH | Hospital Ship |
| AK | Cargo Ship |
| AKD | Cargo Ship Dock |
| AKL | Light Cargo Ship |
| AKR | Vehicle Cargo Ship |
| ANL | Stores Issue Ship |
| AO | Net Laying Ship |
| AOE | Oiler |
| AOG | Fast Combat Support Ship |
| AOR | Gasoline Tanker |
| AP | Replenishment Oiler |
| AR | Transport |
| ARS | Repair Ship, Salvage Ship |
| AS | Submarine Tender |
| ASPB | Assault Support Patrol Boat |
| ASR | Submarine Rescue Ship |
| ATA | Auxiliary Ocean Tug |
| ATC | Armored Troop Carrier |
| ATF | Fleet Ocean Tug |
| ATS | Salvage and Rescue Ship |
| AV | Seaplane Tender |
| AVM | Guided Missile Ship |
| CA | Heavy Cruiser |
| CC | Command Ship |
| CCB | Command and Control Boat |
| CG, CGN | Guided Missile Cruiser |
| CL | Light Cruiser |
| CLG | Cruiser |

(1) In your job would you ever have to perform reading tasks using material like this to look up facts?




(2) How *frequently* do you perform a reading task similar to this?

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 1 to 3 times a year | 1 time each month | 2 to 3 times a month | 1 or more times each week | Daily |

(3) What might be the consequence of a reading error with this type of reading task?

1 No consequence
2 I would be disciplined and some time would be wasted
3 The job would have to be done over again and some time would be wasted
4 The job would have to be done over and some material would be wasted
5 The job would have to be done over and some people would be inconvenienced
6 Equipment would be damaged or lost
7 I might be injured or other personnel might be injured

**Figure 1 Sample Page from the Navy Reading Task Inventory**

All ships are assigned designations – a group of letters which indicate their type and general use – and *hull numbers*, which are usually assigned in sequence to ships of a type as they are built. These identifying designations are used in correspondence, records, plans, communications, and sometimes on ships' boats, because letter and number designations are shorter than the ship's name – *Mission Capistrano*, (AC 162) – and help to avoid confusion between such similar names as *Home* (DLG 30) and *Hornet* (CVS 12) or *Phoebe* (MSC 199) and *Phoebus* (YF 294).

The first letter in a designator is a general classification: *D* for destroyers, *S* for submarines, *L* for amphibious vessels, *M* for minewarfare vessels, *A* for auxiliaries, *W* for Coast Guard vessels, *T* for Military Sealift Command ships, and *Y* for service and yard craft. In combatant designations, the letter *N* means nuclear powered and *G* means the ship is equipped to fire guided missiles. A listing of most ship designations follows; minor yard craft and service craft have been omitted.

| | |
|---|---|
| AD | Destroyer Tender |
| ADG | Degaussing Ship |
| AE | Ammunition Ship |
| AF | Store Ship |
| AFS | Combat Store Ship |
| AG | Miscellaneous |
| AGDE | Escort Research Ship |
| AGEH | Hydrofoil Research Ship |
| AGER | Environmental Research Ship |
| AFG | Miscellaneous Command Ship |
| AGM | Missile Range Instrumentation Ship |
| AGMR | Major Communications Relay Ship |
| AGOR | Oceanographic Research Ship |
| AGP | Patrol Craft Tender |
| AGS | Surveying Ship |
| AGSS | Auxiliary Submarine |
| AGTR | Technical Research Ship |
| AH | Hospital Ship |
| AK | Cargo Ship |
| AKD | Cargo Ship Dock |
| AKL | Light Cargo Ship |
| AKR | Vehicle Cargo Ship |
| ANL | Stores Issue Ship |
| AO | Net Laying Ship |
| AOE | Oiler |
| AOG | Fast Combat Support Ship |
| AOR | Gasoline Tanker |
| AP | Replenishment Oiler |
| AR | Transport |
| ARS | Repair Ship, Salvage Ship |
| AS | Submarine Tender |
| ASPB | Assault Support Patrol Boat |
| ASR | Submarine Rescue Ship |
| ATA | Auxiliary Ocean Tug |
| ATC | Armored Troop Carrier |
| ATF | Fleet Ocean Tug |
| ATS | Salvage and Rescue Ship |
| AV | Seaplane Tender |
| AVM | Guided Missile Ship |
| CA | Heavy Cruiser |
| CC | Command Ship |
| CCB | Command and Control Boat |
| CG, CGN | Guided Missile Cruiser |
| CL | Light Cruiser |
| CLG | Cruiser |

**NAVY READING TASK TEST/INVENTORY—RESULTS OF EXPLORATORY STUDY**

Type of Task: Following Directions Using Texts and Tables (Form E, Item 16)

Question: *Situation* — You are on watch and have been told to report all ships that you see. When reporting, you have been told to first give the ship's general classification and then the designation. You have sighted a light cargo ship.

What do you report?

Test Results: Percentage of personnel at each reading grade level who got the correct answer.

READING GRADE LEVEL

| | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|
| N | 10 | 8 | 7 | 8 | 11 | 8 | 7 | 6 | 17 | 82 |
| % | 40 | 50 | 86 | 62 | 55 | 62 | 71 | 83 | 82 | 66 |

Inventory Results: Frequency with which this type of task is performed

| | 1 1 to 3 times a year | 2 1 time each month | 3 2 to 3 times a month | 4 1 or more times each week | 5 Daily | Not Used |
|---|---|---|---|---|---|---|
| Electronics Technician | | x | | | | |
| Electrician's Mate | | | | x | | |
| Gunner's Mate | | | | | | x |
| Boatswain's Mate | | | | | | x |

**Figure 2 Performance of Personnel at Various Reading Grade-Levels on a Test of Following Directions Using Text and Tables**

# The Air Force Job Performance Appraisal System

Nancy A. Thompson
Douglas K. Cowan
John A. Guerrieri

AD P001407

Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

## Introduction

### Civil Service Reform Act

In October 1978, Public Law 95-454 was passed. This law is more commonly referred to as the Civil Service Reform Act of 1978. Among other mandates, the law stated that each agency was responsible for developing one or more performance appraisal systems which:

"(1) provide for periodic appraisals of job performance of employees;

(2) encourage employee participation in establishing performance standards; and

(3) use the results of performance appraisals as a basis for training, rewarding, reassigning, promoting, reducing in grade, retaining, and removing employees." (2USC4302)

The Air Force responded to the law by developing three systems: (1) the Senior Executive Appraisal System (SEAS), (2) the General Manager Appraisal System (GMAS), and (3) the Job Performance Appraisal System (JPAS). This paper summarizes the JPAS, describing the work plan, ratings, and implementation.

### Air Force Appraisal Systems

The Senior Executive Appraisal System was the first system developed and implemented in October 1979. The two major components are the performance appraisal and the Performance Review Boards (PRBs). These boards distinguish the Senior Executive Appraisal System from the other appraisal systems. Through several iterations, the boards eventually select the top performers in the Air Force (approximately 20% of the executives) and recommend these individuals to the Secretary of the Air Force for bonuses (Guerrieri, 1981).

The General Manager Appraisal System was implemented in October 1980 (Cowan, Thompson, Guerrieri, & Vitola, 1981). This system responds to Title V of the Civil Service Reform Act, requiring that supervisors or managers who are in General Schedule grades 13, 14, or 15 be recognized and rewarded according to performance by varying merit pay adjustments. Persons in this category have been designated as General Managers by the Air Force. Each General Manager is a member of a specific merit pay unit ($n \geq 35$) composed of designated units or homogeneous clusters of job types. Performance is



PREVIOUS PAGE IS BLANK

appraised based on the job performance elements and job standards recorded in the General Manager's work plan. If the General Manager receives an acceptable rating, merit pay is awarded according to a dollars per share computation as determined by the individual's rating and grade.

The Job Performance Appraisal System was implemented on 1 October 1981 and applies to all other employees (non-GM and non-SES). Included in this category are all General Schedule employees (GS 1-18) and all Federal Wage System employees (1-19, consisting of Wage Grade, Wage Leader, and Wage Supervisor). The supervisor, with employee encouraged participation, writes a work plan tailored to the uniqueness of that job. Of great importance in the work plan writing process is the establishment of channels of open communication between the supervisor and employee that guide employee performance throughout the year. The second part of JPAS is the actual performance appraisal. At the end of the rating period, the employee is rated based on the elements and standards in the work plan. Specifications for the JPAS can be found in Air Force Regulation 40-452.

Job Performance Appraisal System

## Appraisal System Design

An initial JPAS model was developed prior to CSRA. This conceptual model was modified to meet the specifications of the law. Two field tests were conducted during the developmental stage and, after several revisions, the final appraisal system was written. The first field test reached a large sample population in diverse job families at McClellan AFB, California. A second field test was conducted at Norton AFB, California, with one organization of predominantly General Schedule employees and one of predominantly Federal Wage Scale employees.

Both field tests resulted in similar findings. The field tested system (1) was viable, (2) did not discriminate against women and minorities, (3) effectively measured job performance, (4) was perceived as fair, (5) improved job related communication, and (6) had minimum ratee inflation. The field tests motivated refinements in the work plan procedure as well as the rating procedure. The final system is one that is simplistic in approach and achieves maximum objectivity.

## Work Plan - Elements

The work plan consists of two parts: 1) job performance elements and 2) performance standards. Job performance elements describe the actual tasks performed on the job by the employee. The job tasks are determined from the position description, the supervisor's and employee's knowledge of the job, and any other documents pertinent to the job. The elements should not be too detailed nor should they be so general that they cannot be easily measured. Elements specify actual work tasks rather than knowledge, skills, or abilities necessary to accomplish the job. There are two methods for entering elements on the work plan: (1) line entry method and (2) functional category method.

Line Entry Method. The line entry method of writing elements identifies

each job requirement using a 1- or 2-line statement. Examples of line elements are shown in Table 1. Since there are some jobs in the Air Force that have a limited number of relatively simple job tasks, the line entry method was proposed primarily for these jobs.

Table 1. Job Performance Elements - Line Entry Method

| PART I – WORK PLAN JOB PERFORMANCE ELEMENTS | | | | | | | |
|---|---|---|---|---|---|---|---|
| NUMBER EACH JOB PERFORMANCE ELEMENT AND SUBELEMENT. CHECK CRITICAL OR NON-CRITICAL BOX FOR EACH ELEMENT. ENTER RELATIVE IMPORTANCE POINTS FOR EACH ELEMENT | CRITICAL | NON-CRITICAL | RELATIVE IMPORTANCE POINTS | EVALUATION: DID NOT MEET | EVALUATION: MET | EVALUATION: EXCEEDED |
| 1. Posts stock list changes | | | | | | |
| 2. Documents technical activities. | | | | | | |
| 3. Reviews manuscripts for correct spelling and grammar. | | | | | | |
| | | | | | | |

## Functional Category Method

Functional category elements provide a method for clustering similar tasks under a single heading. A functional category element has a number of subelements which are specified in the same manner as line entry elements. At appraisal time, each element is evaluated. The advantage of the functional category method is that the element evaluation is based on several subelements rather than a single line entry. An individual may perform a job task that does not fit into a functonal category. Therefore, elements may be entered on the work plan using a combination of both methods. Table 2 shows examples of functional category elements.

**Table 2. Job Performance Elements - Functional Category Method**

| PART I – WORK PLAN JOB PERFORMANCE ELEMENTS | | | | | | |
|---|---|---|---|---|---|---|
| NUMBER EACH JOB PERFORMANCE ELEMENT AND SUBELEMENT. CHECK CRITICAL OR NON-CRITICAL BOX FOR EACH ELEMENT. ENTER RELATIVE IMPORTANCE POINTS FOR EACH ELEMENT. | CRITICAL | NON-CRITICAL | RELATIVE IMPORTANCE POINTS | EVALUATION: DID NOT MEET | EVALUATION: MET | EVALUATION: EXCEEDED |
| 1. COORDINATION: (1) Provides technical guidance to other DoD and government agencies. (2) Attends technical conferences. (3) Coordinates projects with other groups within the division. | | | | | | |
| | | | | | | |
| 2. EDITING: (1) Reviews rough draft manuscripts. (2) Ensures proper security warnings and markings are on reports. (3) Ensures proper format in final copy. | | | | | | |
| | | | | | | |

## Criticality

Each element must be designated by the supervisor as either critical or noncritical, depending upon the importance of the element to the entire job. A critical element is essential to the nature of the position, and failure to meet the requirements of the element warrants management action. The elements are also assigned importance points which must total 100. At least 51 of the relative importance points must be assigned to critical elements. Therefore, at least one element, and in most cases more than one element, must be designated as critical. Although critical elements must be assigned a minimum of 51 relative importance points, in some cases they may be assigned 100 relative importance points if the entire work plan is designated critical.

## Work Plan - Performance Standards

Performance standards are written to specify the level of achievement necessary for satisfactory performance of job elements. A minimum of one standard must be written for each element or subelement; but more may be written, if applicable.

Standards are written to reflect a level of performance considered satisfactory by the average employee and are measured in terms of timeliness, quality of work, or quantity of work. In many instances, a range of values rather than a discrete amount is used to specify the standard. Additionally, standards should be realistic, practical to observe, obtainable, and exceedable, whenever possible. Table 3 contains examples of standards.

Table 3. Performance Standards

| PART II – WORK PLAN PERFORMANCE STANDARDS |
| --- |
| NUMBER EACH PERFORMANCE STANDARD TO CORRESPOND TO THE JOB PERFORMANCE ELEMENTS AND SUBELEMENTS LISTED IN PART I. |
| 1. Stock list changes are performed at a rate of 65 to 70 changes per hour.<br>2. Documents data analysis in report form 2 months after end of the technical effort.<br>3. Manuscripts are reviewed and corrected 8 to 10 working days after receipt. |
| 1. COORDINATION: (1) Guidance is initiated 5 working days after request. One valid complaint is allowed per quarter regarding interaction with other DoD and government personnel. (2) Results of technical conferences are documented 4 to 5 working days after conference. (3) Weekly verbal coordination is made concerning each project. |

## Periodic Performance Reviews

The essence of the new performance appraisal system is establishing open channels of communication and supervisor/employee agreement of what is expected of that employee throughout the year. Periodic performance reviews are a very positive aspect of the system. Supervisors and employees are to meet several times during the appraisal period. Modifications may be made at review time to reflect changing job performance requirements. This is an opportune time for the supervisor to give encouragement for satisfactory performance and to motivate the employee to even better performance. It is also the time to encourage and assist employees who are falling below acceptable levels of performance to strive for improvement.

## Job Performance Appraisal

At the end of the appraisal period, the employee's accomplishment of each job performance element is evaluated in light of the standards established for the elements. Based on these evaluations, an overall performance rating is given. Certain personnel actions may then be taken pursuant to the overall rating.

### Element Evaluation

Every line entry or functional category element is evaluated as MET, EXCEEDED, or DID NOT MEET. Substantiation must be provided on the rating form for elements rated EXCEEDED or DID NOT MEET. An evaluation of EXCEEDED means that performance was better than the satisfactory level which was indicated by the standard. An evaluation of DID NOT MEET reflects performance that was unsatisfactory. Table 4 shows a portion of a job performance appraisal form filled in with elements, criticality, relative importance points, and element evaluations.

Table 4. Completed Job Performance Element Sheet

| PART I – WORK PLAN JOB PERFORMANCE ELEMENTS | | | | | | |
|---|---|---|---|---|---|---|
| NUMBER EACH JOB PERFORMANCE ELEMENT AND SUBELEMENT. CHECK CRITICAL OR NON-CRITICAL BOX FOR EACH ELEMENT. ENTER RELATIVE IMPORTANCE POINTS FOR EACH ELEMENT. | CRITICAL | NON-CRITICAL | RELATIVE IMPORTANCE POINTS | EVALUATION | | |
| | | | | DID NOT MEET | MET | EXCEEDED |
| 1. Monitors obligation and expenditure authority funds. | X | | 50 | | | X |
| 2. Maintains balanced financial records. | X | | 30 | | X | |
| 3. Prepares correspondence. | | X | 20 | | X | |
| | | | | | | |

## Overall Performance Rating

Once the element evaluations have been made, the overall peformance rating must be made according to the definitions of the ratings (see Table 5). The

employee must at least meet the requirements of every element to be rated Fully Successful. An excellent rating is given when the employee exceeds the job performance requirements of the elements which represent at least 50% of the relative importance points and meets the requirements for the remaining elements. Superior is given when the employee exceeds the requirements of every element. Minimally Acceptable is the rating given when the employee did not meet one or more noncritical elements, even if the employee met or exceeded the rest of the elements. Unacceptable is the mandatory rating when the employee did not meet one or more critical elements of the work plan.

Table 5

Overall Performance Rating Scale

---

SUPERIOR: Employee exceeds the performance requirements of all the job performance elements of the work plan.

EXCELLENT: Employee meets the performance requirements of all the job performance elements of the work plan and exceeds the performance requirements of the job performance elements which represent at least 50% of the relative weight in importance of the work plan.

FULLY SUCCESSFUL: Employee meets the performance requirements of all the job performance elements of the work plan.

MINIMALLY ACCEPTABLE: Employee meets the performance requirements of all critical job performance elements of the work plan, but does not meet the performance requirements of one or more noncritical job performance elements.

UNACCEPTABLE: Employee does not meet the requirements of one or more critical job performance elements of the work plan.

---

Ramifications

A rating of Fully Successful or higher is necessary for an employee to receive a within grade increase. Excellent or Superior ratings qualify an employee for performance awards. These awards include Sustained Superior Performance, Quality Step Increase, and other monetary and functional awards. Excellent and Superior ratings also add years of creditable service to an employee's service date for reducton in force purposes: Excellent adds two years and Superior adds four years to creditable service for the next year. A rating of Fully Successful or higher also makes the employee eligible for basic merit promotion.

Ratings of Minimally Acceptable or Unacceptable require some action from the supervisor. The supervisor must work with the employee rated minimally acceptable to help improve performance. It is appropriate to recommend special training to improve the employee's performance. An Unacceptable rating also requires career counseling. If performance continues at an unacceptable level, the employee may be reassigned, demoted or removed. Table 6 summarizes the ramifications of the overall performance ratings.

Table 6
Ramifications of Overall Ratings

| Ratings | Eligible for Within Grade Increase | Eligible for Merit Promotion | Eligible for Creditable Service | Eligible for Official Honorary or Monetary Rewards | Requires Employee Counseling |
|---|---|---|---|---|---|
| Superior | Yes | Yes | Yes-4 years | Yes | No |
| Excellent | Yes | Yes | Yes-2 years | Yes | No |
| Fully Successful | Yes | Yes | No | No | No |
| Minimally Acceptable | No | No | No | No | Yes |
| Unacceptable | No | No | No | No | Yes |

## Implementation

The Job Performance Appraisal System became effective 1 October 1981. As of this date, a work plan should have been written and approved for every General Schedule and Federal Wage Scale employee in the Air Force. Appraisals in the JPAS do not all occur at the same time. During the first year, no appraisals will be made prior to 1 February 1982. Thereafter, appraisals will be completed at least 60 days prior to the anniversary date of the last within grade increase or promotion. During the first year of employment, Federal Wage Scale employees will have a work plan for the first 26 weeks of employment and a rating will be made within 2 weeks of that first 6 month period. The work plan will be reaccomplished for the second half of the first year. Thereafter, FWS employees will be on the same time frames as the GS employees, i.e. rated on the anniversary date of their last within grade increase or promotion.

Extensive training for all phases of the new civilian appraisal system is provided in the Air Force. Two training courses were developed specifically for the JPAS. An 8-hour training course primarily designed for supervisors provides the information and practice necessary to write effective work plans. Employees may also attend the 8-hour course, either voluntarily or on the supervisor's recommendation. However, most nonsupervisory employees attend the 4-hour training course that presents an overview of the system enhanced by an audio-visual presentation and limited practice writing elements and standards.

Based on field test data, the attitude toward the new job performance appraisal system has generally been one of acceptance. If approached honestly, the new system provides a much more objective approach to evaluation than the system it is replacing. There are some objections, chief among them is that people frequently assert that standards just can't be written for their jobs. It is not easy to write standards and some jobs are more difficult to define in objective terms than others. It was found, through

various field tests, that standards can be written for every job. It requires an honest look at the job elements and the performance level required to satisfactorily perform the job.

Other objections are that the system is expensive and time consuming. These are both valid objections; but the potential for increased productivity and fairness resulting from the system far outweighs these operational problems. It requires the education of trainers and time off the job for training of every employee and supervisor under the JPAS. It also requires some quality time and effort on the part of the participants outside of training to write work plans, review them throughout the year, and render objective ratings. The long range benefits of an efficient and productive work force justify the difficulties that will be encountered.

In the final analysis the success of the system lies with the participants. It can only be as good as the integrity of the people who use it.

## References


Civil Service Reform Act of 1978, Public Law 95-454.

Cowan, D. K., Thompson, N. A., Guerrieri, J. A., & Vitola, B. M. Appraisal and Merit Pay Systems for Air Force Civilian General Managers. AFHRL-SR-81-36. Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Department of the Air Force. Performance Appraisal Program. (Air Force Regulation 40-452) Headquarters U.S. Air Force, Washington, D.C., 1 October 1980.

Guerrieri, J. A. Air Force Senior Executive Appraisal System. AFHRL-SR-81-11. Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Literacy: Assessment and Remediation

AD P001408

Duncan, R. Eric, 1LT, USAF Occupational Measurement Center, Randolph Air Force Base, Texas (Chair); Claudy, John J., American Institutes for Research, Palo Alto, California; Fischl, M. A., Kern, Richard P., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia; Wisher, Robert A., Navy Personnel Research and Development Center, San Diego, California; and Payne, Sandra A., Personnel Research and Development Center, Office of Personnel Mgt., Washington, D.C.

The discussion in this area involved (1) whether or not there really is a reading problem, and (2) the development of reading (literacy) skills in general and those specific to job situations. While Dr. M. A. Fischl played devil's advocate, it was generally accepted that the military services and other Government agencies need to teach people to read. The primary discussion focused on the breadth of literacy training. Before literacy instruction is given, an accurate assessment of reading skills is necessary. The Government's view of literacy assessment is job-related; hence, assessment measures being developed and used assess literacy using specific job or job-type materials. These materials incorporate not only job-specific text but also tables and figures. The advantage of Government literacy assessment is its specificity which is also its primary disadvantage. After identifying the literacy components of job proficiency, developing assessment tools, and instructing poor performers, will this investment pay off in the long run? While these steps may lead an individual to basic job literacy, the long-term transferability and skill retainability have yet to be sufficiently demonstrated. The panelists expressed a wary optimism for the future of job-specific literacy training but asked more questions than they answered. One question raised was whether we should give up trying to promote generalized literacy skills in the minimally competent or should we be content with minimal job-related reading skills? If we are content with the latter, will we need to generate literacy instruction for every job specialty where the minimally competent are admitted? While it is acceptable practice to train to reading proficiency before we train the job skills, the Government may be hesitant to lose personnel resources for extended literacy training. So far, DOD has survived with illiteracy in its midst. Maybe we really don't have a real literacy problem. Even with the great advances we have made in literacy assessment and instruction, many of the problems and questions remain unsolved and unanswered.

# Development of the Job Reading Test (JRT)

John G. Claudy, Ph.D.
American Institutes for Research
Palo Alto, California

This presentation describes the development, validation and standardization of three parallel forms of a short, multiple-choice test of job-related reading skills. The test was built specifically to measure skill in performing Army job reading tasks, i.e., skill at obtaining the information which a soldier needs to perform actual job tasks by reading the Army printed material which is furnished to provide that job information. Each form of the test contains materials representing the four types of Army job reading tasks identified in prior research (Locating Job Information in an Index, in Tables and Charts, and in Narrative Descriptions; and Following Procedural Directions), and the reading passages were drawn equally from Army publications used in six major, high-density MOS clusters (Clerk, Combat, Communications, Cook, Mechanic, and Medical). A simplified test format was developed to reduce administrative problems which might hinder some soldiers in taking the test, and testing time was held to forty minutes. Norm tables were developed to convert raw scores to their percentile equivalents in the standard mobilization base population.

The project I will be telling you about was in a sense a continuation of activities carried out as part of HumRRO work unit FLIT (Functional LITeracy) which was directed by Tom Sticht who spoke to us this morning. Therefore some background information is necessary to set the stage for my presentation. As a part of work unit FLIT the HumRRO Field Unit based at the Presidio of Monterey and Fort Ord in California developed three parallel forms of a Job Reading Task Test, or, as it is usually called, the JRTT. The content and coverage of the JRTT was determined on the basis of an extensive analysis of the types of printed materials Army soldiers actually use during AIT training and later on the job, and of how they use or deal with the materials. That is, what types of printed materials they use and how they use them. On the basis of this analysis it was determined that the enlisted personnel dealt with printed materials in four major ways:

1. They located information through the use of indices,
2. They extracted information from tables and charts,
3. They extracted information from narrative prose, and
4. They followed procedural directions.

Each form of the JRTT was constructed to contain an equal number of test items testing the examinees' ability to carry out each of these activities or skills using printed materials.

Since the JRTT was developed to provide an indication of the examinees' functional literacy with regard to Army jobs rather than their general literacy, all of the stimulus passages used in the test were drawn from Army training and technical manuals and materials. Insomuch as the majority of Army soldiers, especially soldiers with reading problems, are assigned to MOSs in one of six major MOS clusters, the stimulus passages used in the JRTT were drawn only from printed materials used by soldiers in these six MOS clusters. The six MOS clusters were: clerk, combat, communications, cook, mechanic, and medic.

Each form of the JRTT consisted of four sections, each section testing one of the four major ways printed materials are used by soldiers with the stimulus passages drawn from the manuals and materials used in one of the six MOS clusters. Directions for the section and the questions were presented on a right hand page and the printed stimulus material containing the information

necessary to answer the questions was presented on the facing left hand page. The answers to the questions were written by the examinees, in a brief, free-response format, on a separate answer sheet. A basic characteristic of the JRTT was that the answers to the questions were contained in the stimulus passage provided. Questions such as "Which of the following is the best title for this passage?", or "Which of the following words means the same thing as the word incorporate?" were not included. Thus the JRTT is a test of information use or extraction rather than of reading comprehension.

The project undertaken by AIR was to develop three new parallel, equally difficult forms of the JRTT in a multiple choice format, and to develop norms for these new forms based on the current mobilization population. To distinguish these new forms from the JRTT, they were termed the Job Reading Test, or JRT. As with the JRTT, all four types of reading tasks or skills were to be included in each of the three new JRT forms; moreover, it was also specified that stimulus passages from all six of the MOS clusters were to be included in each of the JRT forms. Thus six sections per form were required, rather than four, with two sections of each form testing the same skills applied to two different content areas. An equal number of test items for each skill were to be included in each form, so the six sections would have to have different numbers of items per section, rather than the same number as was the case with the JRTT.

As the initial step in test development, the 120 open-ended items of the existing JRTT forms were converted to five-option, multiple-choice format. These were supplemented by further test items based on the same stimulus passages. To insure sufficient items to permit each type of reading task and each MOS cluster to be represented in the final test forms, 16 additional reading passages were selected from Army publications and test items pertaining to these passages were prepared. After editing, twenty-four stimulus passages with an average of 12 items per passage comprised the final item pool.

As an initial check on the "goodness' of this item pool, six tryout sets of items were assembled and administered in a small scale field trial to soldiers entering on their initial duty assignment in the 7th Infantry Division at Fort Ord. In these trials examinees were required to attempt all items within a two hour time limit and to note any questions which they felt were unclear or unfair. In addition, to obtain testing time estimates, the examiner noted the time spent on each section of the test. In a one-hour debriefing session for each group,

the soldiers reported that they had encountered no procedural difficulties, liked the Army content of the test, and would accept the test as a fair measure of their ability to read Army materials. On the basis of examinee performance and comments, some items were revised, some were dropped, and the entire testing package was very carefully reviewed.

In the initial JRT tryouts, each tryout set of the test materials consisted of one booklet containing reading stimulus passages, a second booklet presenting the multiple-choice items, and a separate answer sheet. It was apparent that this format was unwieldly. In order to simplify the task of recording the responses and to minimize the amount of work surface required, a revised test format was developed for the item validation and maintained unchanged into the final forms. In this format the stimulus passages remain in a separate booklet. However, both the question booklet and a pad of answer sheets are attached to an 11" x 17" cardboard backing sheet. As each page of the question booklet is turned, a new set of items is uncovered in the question booklet and is automatically aligned with a new column of answer spaces which is exposed on the answer sheet. The five response options for each item are listed under the item in the question booklet, and each is followed by a horizontal dotted line that leads directly to the space on the answer sheet corresponding to that response option. Upon completing the test, the examinee's answer sheet is removed, thus exposing a clean answer sheet for the next examinee.

For purposes of item validation, four provisional forms of the JRT were assembled from the item pool and administered to enlistees at the Reception Station at Fort Dix, New Jersey, in October, 1978. Each provisional test form was approximately 75 items in length, contained an approximately equal number of items fom each skill area, and was comprised of items from each of the six MOS clusters. To determine the validity of the individual items, each provisional test form was administered to groups of Army enlistees of high and low reading ability. Since no direct measure of reading ability is routinely available in Army records, and because it was not feasible to administer an additional reading test in this setting, it was decided to use the Field Artillery (FA) Aptitude Area score of the ASVAB as the index of general reading ability. This decision was based on a known correlation of approximately .80 between the FA acore and scores on the Metropolitan Reading Achievement Test.

The provisional forms of the JRT were administered to groups of enlistees with high reading ability (FA $>$ 104) and with low reading ability (FA $<$ 96).

FA scores are expressed on the Army Standard Score Scale which has a mean of 100 and a standard deviation of 20, so these two groups were selected to be at least one quarter of a standard deviation above and below the mean of the standardization population respectively. In fact, their medians were more than one half of a standard deviation away from the population mean.

Item analyses were performed to yield for each item an index of validity or discrimination, defined in terms of the Phi Coefficient, and of difficulty in terms of the proportion of examinees answering the item correctly. For the discrimination index, membership in the upper or lower reading ability group, as defined by FA acores, was the criterion variable. It was decided in advance that only items with a corrected Phi Coefficient of at least 0.30 would be retained for possible inclusion in the final test forms. Of the 290 items administered in the item validation trial, a total of 235, or 81 percent had a corrected Phi Coefficient of at last .30, and of these 235, 66 percent had a corrected Phi of at least .50. For the 235 retained items, the difficulty range in the validation sample was 10 percent to 97 percent answering correctly with a mean of 66 percent of the examinees answering correctly. Two hundred ten, or 89 percent, of the items had difficulties in the 40 percent to 89 percent range.

An additional constraint on the test development effort was that the final forms of the JRT require no more than 40 minutes for administration. The number of items to be included in each of the final forms would have to be determined in such a way as to meet this constraint while at the same time permitting most of the examinees to at least attempt every item. After discussions with the ARI staff, it was determined that a 75 percent completion rate for examinees would be the target. Thus there arose the question of how many items can we reasonably expect at least 75 percent of the examinees to complete in 40 minutes. To answer this question we used timing information collected during the small scale tryout conducted at Fort Ord. Since the time each examinee had spent on each of the test sections had been carefully recorded, it was possible to calculate the times it required for 75 percent of the examinees to complete each type of item. These 75th percentile times were:

1.5 minutes per index item,
.9 minutes per table/chart item,
.9 minutes per narrative item, and
1.1 minutes per procedural directions item.

Based on these times, nine items of each type could be administered in 40 minutes with an expected completion rate of 75 percent. Accordingly, a final test length of 36 items was selected.

The task of assembling the three final parallel forms of the JRT from the pool of validated items was, in substantial part, a matter of inspection and judgment. Each form consisted of six parts structured as follows:

| | | |
|---|---|---|
| Part I | 5 items | Narrative |
| Part II | 9 items | Procedural Directions |
| Part III | 5 items | Tables/Charts |
| Part IV | 4 items | Narrative |
| Part V | 9 items | Index |
| Part VI | 4 items | Tables/Charts |

Of the twenty-four sets of items administered in the item validation, only two sets failed to yield a sufficient number of validated items to be used in the above test outline. Form A was constructed first on the basis of item difficulties to yield a somewhat easy test which maximized discriminations in the lower range of possible test scores. Where items of equal difficulty were available, item choice was based on the item discrimination index. Forms B and C were constructed to parallel as closely as possible the difficulty distribution of Form A. The three forms had mean difficulty indices of 62.2 percent, 62.3 percent, and 62.3 percent.

These three final forms of the JRT were then administered in the Reception Stations at Fort Dix, New Jersey, and Fort Leonard Wood, Missouri, in February of 1979 to obtain norming data. Following an orientation and briefing by AIR staff members, the tests were administered by Reception Station staff to all non-prior-service personnel being processed through the Reception Station. Approximately 650 examinees were tested in each of the two Reception Stations. Useable data were collected from a total of 1231 individuals divided approximately equally across the three forms.

The specifications for the project called for the development of norms to reflect the underlying mobilization population. The characteristics of the mobilization population are such that ten percent of the population should fall into each of the AFQT deciles. However, examination of the AFQT scores for the 1231 individuals who had taken the JRT forms clearly indicated that the available sample did not represent the mobilization population.

For example, only one-tenth of one percent of the cases fell in decile one, the lowest decile, and only about 5 percent of the cases fell in decile ten, the highest decile. However, approximately 22 percent of the cases fell into both the fourth and fifth deciles. In order to correct for this lack of representativeness on the part of the available cases, a statistical adjustment procedure was used to estimate what the mobilization population norms would be if a representative sample of the mobilization population were available.

In order to properly calculate the norms for the JRT forms, it was necessary to divide the AFQT distribution for the available cases into a discrete number of intervals, the number of intervals being a function of the total number and distribution on available cases. After consideration of several alternatives, 20 intervals were decided upon. In the mobilization population 1/20th, or five percent, of the population would be expected to fall into each of these 5-point-wide AFQT intervals.

Having divided the AFQT distribution into 20 intervals the remaining steps of the procedure for calculating the JRT norms were as follow:

1. Calculate the JRT mean and standard deviation for the available norming subsample examinees in each of the AFQT intervals for each form.
2. Plot JRT means and standard deviations against AFQT intervals.
3. Smooth the curves plotted in Step 2.
4. For AFQT intervals for which no cases were available or for which mobilization population proportions were grossly underrepresented, those in Mental Categories IV and V, project the smoothed curves from Step 3 and obtain predicted means and standard deviations for these AFQT intervals.
5. Using a computer program, for each AFQT interval, generate a sample of 50 constructed JRT raw scores having a normal distribution with mean and standard deviation equal to the smoothed mean and standard deviation determined in Steps 3 or 4.
6. Combine the 1000 JRT scores (20 intervals x 50 scores per interval) generated in Step 5 into a single distribution. This distribution is the best estimate of the distribution which would result if the JRT could be administered to a large randon sample from the entire mobilization population.

7. Calculate the percentile corresponding to each raw score in the generated distribution. This is the raw-to-percentile conversion table for the JRT form in question. There is a separate table for each of the three JRT forms.

With the calculation of the raw-to-percentile conversion table and the preperation of an administration manual the development of the JRT was complete. I wish that I could describe to you how useful the JRT has been in Army literacy screening and remediation programs. However, I can't since such use has not yet come to pass. In all fairness I must admit that a problem related to the JRT arose soon after it was delivered to ARI at the end of the contract in mid 1979. About that time some basic restructuring of the AFQT norming procedure was undertaken by the Services. Since the AFQT scores used in norming the JRT were subsequently revised, our JRT norms, to some degree, fail to represent adequately the mobilization population.

The JRT exists. Those of us who developed it think it is a good test. We hope that someday it will be renormed and used.

AD P001409



# A Theory and Model of Item Readability[1]

R. Eric Duncan
United States Air Force

[1]The views expressed in this paper represent those of the author and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

## ABSTRACT

This paper reviews the literature relevant to readability prediction, critiques past and present models of readability, and presents a comprehensive model of test-item readability. There are four basic groupings of characteristics which impact on the predictability of item readability:

(1) personal cognitive characteristics--those cognitive characteristics which each person brings into every testing situation;

(2) other personal characteristics--those motivational and tool-skill characteristics each person brings into the testing situation;

(3) environmental characteristics--the purpose and use a test will serve, what type of test it is, and the conditions of equal employment opportunity surrounding the test's use; and

(4) item characteristics--those semantic and syntactic item characteristics which have been identified as predictive of item comprehensibility. A definition of the comprehension of item text and a general procedure for estimating an item's readability are also discussed.

## Introduction

Test-item readability is a psychometric concept which has not been adequately addressed. Improvements in assessing test-item readability should address the following three shortcomings of past practices:

(1) Failure of past models to treat test items differently from common prose

(2) Failure to minimize item difficulty by accounting for that unique variance and inadvertent readability effects, and

(3) Failure of past models to account for all sources of variance.

Regarding the first-mentioned failure, most attempts at establishing a test-item readability model have treated tests as conjunctive prose (e.g., the prose that appears in textual material) with concomitant readability estimation procedures. Those attempts which have treated tests as disjunctive prose (examination at the item level) have failed to separate item readability from item difficulty or have examined item readability from an atheoretical position. My model of item readability treats each item independently of all other items, estimates the psycholinguistic as well as the structural components of an item, attempts to remove the difficulty factor from item readability estimation, and attempts to incorporate human performance factors in an item's readability estimate.

The second shortcoming I mentioned was the failure of the readability formulas to minimize the effects of readability and item difficulty on test results. In the case of a job knowledge test, for example, if there is no specifically identified requirement for reading ability, yet test readability affects performance on the instrument, poor readers may not have their knowledge validly tested. A reliable test-item readability formula, which should result from my model, would enable test developers to avoid this pitfall to a greater extent. If the readabililty effects on test items are not minimized, valid equal employment opportunity (EEO) complaints are a very real possibility.

The final deficiency of past practices which I mentioned is the failure to recognize that an assessment of test-item readability should incorporate all important sources of variance. While examining all sources of variance is virtually impossible, a model of item readability should address the most relevant sources of variance. To accomplish such a task, my model incorporates an estimate of a person's reading ability which is weighted by certain personal cognitive characteristics (those cognitive characteristics which each person brings into every testing situation) and other personal characteristics (those motivational and tool-skill characteristics each person brings into the testing situation). Thus, the estimate of person reading ability, used as the criterion in the estimation of item readability, will be adjusted to control for sources of variance which make item readability estimation less reliable. All important sources of variance have been placed in one of the following categories:

(1) Item characteristics (the semantic and syntactic characteristics of an item),

(2) Personal cognitive characteristics,

(3) Other personal characteristics, or

(4) Environmental characteristics (the purpose and use of a test).

I believe that these categories include all sources of variance relevant to the prediction of an item's readability. The purpose of this paper is to (1) examine the relevant literature regarding the failures of past practices, (2) describe the specific components which contribute to item readability, and (3) briefly introduce a methodology which will provide a usable item readability formula.

## The Problem of Comprehension

Before introducing the model which describes my theory of item readability, I must make one digression to examine the issue of comprehension. To date there has been no universally accepted definition of what comprehension is (Klare, 1980). This presents a tremendous problem. If there is no common definition of what comprehension is, how can we even attempt to reliably predict the comprehensibility of prose or items, yet make such prediction understandable to various users and generalizeable in varied conditions? The answer to this question is not easy. Klare (1980) states that the problem does not involve prediction of comprehension but the "inability to agree on measures and levels of comprehension" (p. 16). In other words, the problem of disagreement is due to the varied criterion measures of comprehension. With these measures of comprehension come the concomitant operational definitions of comprehension. Skinner (1957), Chomsky (1972), Thorndike (1917), and many others have described comprehension, but from their theoretical perspectives, limiting the definitions' generalizeability. The problem may also be due to the absence of a universally acceptable theory of readability. Williams and Siegel (1974, p. 10) have stated that theoretically ungrounded reading comprehension prediction equations do not accurately measure the comprehensibility of a typical test: in other words, a test item. In developing and testing models of readability, we can come closer to an acceptable definition of comprehension. After describing my model of item readability, I will introduce a definition of comprehension related more to items than text.

## A Need for Item Readability Estimation

Valentine & Vitola (1970) have expressed a need for accurate estimation of the readability of test items, but did so in a theoretical vacuum. While describing prose readability, Dale & Chall (1949) have unwittingly described the theory of item readability that I will detail by stating that readability is "the sum total (including the interactions) of all those elements within a given piece of printed material that affects the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimum speed, and find it interesting." However, they proceeded to estimate readability in a somewhat atheoretical manner. They indicated that those components important in determining readability were the length of a sentence and vocabulary difficulty (Dale & Chall, 1948 and 1949).

## Two Models of Readability - A Review of the Literature

### Guilford's Structure-of-Intellect Model

Two relatively sound models of readability have been described by Guilford and by Klare. Guilford and his associates (1950, 1954, 1966, 1967), developed a theory which views human mental activity and, thereby, readability (my inference) by describing three separate but interdependent characteristics. Derived from factor analysis, these characteristics are: (1) "contents"-- a factor illustrative of the form in which information is encoded, represented, stored, and decoded; (2) "operations"-- a factor which describes

the actions taken on the "contents"; and (3) "products"-- a factor which describes the results of any "operation". Deignan and Duncan (1978) reported differential performance on an achievement test based upon how students processed the "content" of learned material. Data from this study indicated that tests which were verbal in nature were completed more successfully by those who processed information verbally than by those who processed information more figuratively. Deignan, et al., (1980) provides data which give credence to Guilford's "structure-of-intellect" theory.

The characteristics described by Guilford were determined to be orthogonal (after orthogonal factor matrix rotation) and can be represented in geometric form by a cube as seen in Figure 1. The figure presents 120 unique cells of which many have been validated against either performance or tests by Guilford and his associates.

Siegel and Bergman (1974) suggested that structure-of-intellect components which indicated a high level of mental ability when used in textual material would be less comprehensible for individuals of low and middle abilities than those components indicating lower levels of mental ability. More precisely, textual materials incorporating high levels of these basic components of the structure-of-intellect would be more difficult to read than text with lower levels of these variables. While quite elegant, this interactive theory has one primary problem -- not all the cells are testable. The cells of Guilford's model are intended to be independent, each cell describing some unique variance in the model because of some unique combination of the primary characteristics. The failure of the model lies in its undescribed or untested cells.

## Klare's Readability of Reading Tests Model

Klare (1976) has described a simpler but more inclusive model of readability. In Figure 2, Klare's model of the interacting factors of readability measures is reproduced. His model can be summarized by stating the reader's performance on a reading comprehension test as a function of the reader's reading ability and motivation, the content and readability level of the material, and the testing conditions. This model includes those qualities essential for a universal theory of item readability; personal characteristics, environmental characteristics, and test-item characteristics. Research designed to investigate Klare's model has seemingly led to partial model verification (Klare, 1980). Specifically, the reader motivation and type-of-comprehension measure portions of the model have revealed that (1) when motivation is low, easier-to-read material is recalled (comprehended) in greater quantities than is hard-to-read material, but when motivation is high, reading difficulty has no effect on comprehension, and (2) most reading comprehension measures are negatively affected when prior knowledge can be used to answer questions on reading selections (Entin & Klare, 1980). The second result was also found by Hanna & Oaster (1978). There is empirical evidence to support portions of Klare's interactive model. In the present context, the problems with this model are (1) it doesn't address many of the item characteristics which may have an effect on an item's readability level and (2) it fails to examine to what extent the factors weigh in predicting reader performance.

## A Model of Item Readability

The model that I propose incorporates and weights these portions of the varied components which are necessary to a verifiable theory of item readability. Instead of a flow-diagram which fails to acknowledge all possible interactions or a cube which forces cell identification that may neither be relevant nor attainable, I prefer a pyramid (see Figure 3). The pinnacle of the pyramid represents item readability (comprehensibility). It is a

point (rather than a sphere, plane, or line) because it represents a unique value for every item. It does not vary significantly across samples of people for whom the respective item is designed nor is it affected by the item's difficulty (ability required to answer the item correctly). The pyramid's base consists of four subdivisions of those basic components which must be included in any model of item readability (personal, environmental, and item characteristics).

## Environmental Characteristics

The details that should be described are those of the environmental characteristics surrounding the answering of an item. The readability value of an item will vary depending upon what testing medium is used, the type of test in which the item appears, the purpose for which a test is taken, and the conditions of the social climate which limit the acceptability of decisions made from test scores. For a readability value that is theoretically "invariant", there seem to be many environmental conditions which can create variability. Variability in the readability value of an item will occur only if value calibration is attempted in an environment in which these characteristics are NOT held constant. For example, readability value invariance (within the confines of the standard error of estimate) will be present if an item's readability value is calibrated for a paper and pencil achievement test taken by individuals for promotional competition. This value, however, is invariant only within the stated parameters (i.e., for each set of conditions there is another readability value).

Given that differential selection will be reduced by use of (1) an item readability estimation procedure and (2) iteratively modified matrices of the utilities of improper and proper selection decisions (Darlington, 1976) (i.e., there is high utility for minimizing item readability effects), the social climate characteristic can be removed as a source of invariance. The thought may have occurred to question whether any of these environmental characteristics have a significant impact on an item's readability value. Deignan, et al., (1980) indicate that identical items appearing on a computer terminal and on paper were answered differently. This may be due to the fact that there is rarely more than one item appearing on a computer terminal at any one time while all test items are present on a paper and pencil exam. By having items to compare against or provide clueing, paper and pencil exam items should (on the whole) be responded to differently than items that appear singularly (and must be answered before progressing) on a cathode-ray tube. This characteristic may be related to the personel characteristic of test-wiseness, showing that none of these characteristics operates independently. Depending on whether a test is an achievement, aptitude, or diagnostic test and depending on how the test results will be used, a person's motivation may vary. These components, while probably not exhaustive, do tap the essential construct of environmental characteristics. Without this knowledge, item readability value estimation would most probably be so variable as to be unusable.

## Personal Characteristics

Two separate corners of the triangle's base are needed to detail the contribution of person (or personal) characteristics within the item readability model. Much research has been done to examine the cognitive and affective regions of the human psyche and is too extensive to address here. Suffice it to say that the data seem to support the contention that humans do possess separate, but not independent, cognitive and affective domains. Therefore, the present model of item readability should address both domains as well as include biographical data which can be used to investigate whether items are differentially readable.

## Personal Cognitive Characteristics

"Personal cognitive characteristics" make up a much purer component than "other personal characteristics." A person's reading ability must be accurately measured to be useful in establishing an item's comprehensibility level. In order to incorporate all relevant factors which may affect a person's score on a reading comprehension test (e.g., the Nelson-Denny Reading Test), it is necessary to obtain a weighted score based on those factors' (all person characteristics) contribution to reading ability. Before describing this weighted score further, those person characteristics which affect the weighted reading comprehension score will be examined.

Cognitive skills can be defined as a person's ability to process information and as that cognitive style which a person uses to perform a cognitive task. The literature is rich with reading models which use both top-down and bottom-up information processing approaches. Massaro and Taylor (1980) present an information processing model of reading printed text (See Figure 4). This model is tremendously important in that it describes reading at the well-organized micro-level. However, when test-item reading is examined in light of this model, the questioning component of an item may not be fully explicated. On the other hand, it does provide a construct for memory load in the answering of test items and is directly related to the item characteristic of grammatical structure. Some measure of memory ability should be related to a person's reading ability and is integral to my model of item readability. Using Guilford's model, memory has been shown to be a good predictor of reading ability (Williams and Siegel, 1977). Cognitive styles research indicates differential test-item performance for differing cognitive styles as mentioned earlier. For example, the field independence-dependence cognitive style has been related to non-verbal IQ and is related to test-taking abilities (Deignan, et al., 1980). The scanning cognitive style can be directly related to how item alternatives are selected and to which item pairing strategy provides optimum item statistical properties (see Note #1).

Another component of the personal cognitive characteristic is the individual's knowledge of the subject matter of an item. This component involves not only content difficulty but also how the tested information was obtained. Separating an item's content difficulty from its reading difficulty has been a concern of the testing community for many years and has not been sufficiently addressed. This issue has been the downfall of many test (not item) readability formulas. While intricately related to an item's comprehensibility, it is primarily a methodological issue and, as such, will not be discussed here (see Note #2). Establishing whether performance on a test item is learned or is a function of prior knowledge is also a methodological issue. There are measurement approaches which provide a methodology to assess these differences (e.g., various item characteristic curve (latent-trait) theories). One of these approaches will be used when this model is tested.

The final component to the personal cognitive characteristic of item readability involves the constituent of general intelligence. While it is not the purpose of this section to rehash the nature-nurture issue, suffice it to say that much literature exists in this area which is directly related to differential assignment due to readability effects (see Jensen, 1969, 1973a, 1978; Herrnstein, 1971, 1973; Kamin, 1974; Crawford, 1979; Wilson, 1978; Mili, 1969). It can be seen that IQ is not only related to a person's reading ability (because most IQ tests are highly verbal) but also to (1) the environmental characteristic of test usage, (2) person motivation, and (3) those item characteristics which are most influenced by verbal IQ (syntactic and semantic structures). The unique and combined contributions of intelligence on the weighted reading ability score will be evaluated to

determine whether this component adds significant unique variance to the weighted reading score via regression analysis. The primary reason for its inclusion is to counter arguments that all that is being estimated in my item readability formula is intelligence and not the readability level of an item.

Other Personal Characteristics

"Other personal characteristics" is basically an affective component with test-taking skills and biographics included. Test-wiseness can have an impact not only on an individual's reading score but on performance on the test items whose readability is being assessed. It is necessary to estimate how test-wise a person is (there are instruments to evaluate the quantity of test-wiseness) in order to remove that variance which is neither attributable to subject-matter knowledge nor reading ability. Once this is accomplished, a more stable estimate of a person's reading ability can be obtained.

Physical and mental condition is related to all other components in the model. A healthy constitution and an alert and moderately anxious demeanor are necessary precursors for accurate reading ability estimates (Gage & Berliner, 1975, p. 345). Manifest anxiety scales are excellent tools for establishing the anxiety levels of testees and will be used in testing this portion of the model.

Motivation of examinees has been shown to be related to performance on reading tests by Fass and Schumacher (as cited in Klare, 1980). By extension, this same result should be present in achievement, aptitude, and diagnostic tests to a greater or lesser degree. The literature in this area indicates that, with all other things being equal, the higher the motivation the better the performance on a test with less of an effect for item or test readability. However, for those with very low reading abilities, no amount of motivation can overcome the negative effects of their basic problem. It is an important component and should be included in the determination of a weighted reading ability score.

Item Characteristics

After assigning each item being calibrated a value that is the minimum possible reading level necessary to answer it correctly, the item characteristics can be used to compute a formula to estimate an item's readability level. We are simply trying to estimate item readability and not directly trying to establish causality. However, if all relevant personal and environmental sources of comprehensibility invariance are controlled or accounted for and the item characteristics significantly estimate comprehensibility, a causal inference could be made (but done so with great trepidation).

The item characteristic components of semantic and syntactic structure are very similar to prose readability theory components. However, most of these subcomponents (1a-f, and 2a and b of Table 1) have been altered somewhat to fit item writing constraints. A measure of the readability level, as computed by the Flesch formula, of the text from which the item was written is a necessary predictor of an item's readability level. This necessity is intuitively obvious if one considers, for example, the possible readability differences between a history test item and a physics test item.

The "information requested" component of this corner of the triangle's base is directly related to the "cognitive skills", anxiety, motivation, and "knowledge of the subject-matter" of the examinee and virtually all of the environmental characteristics (see Table 1). Of all the item characteristics this component has probably been the least

attended to by test readability researchers. In this context, the full implications and supporting literature can not be reviewed. However, it is quite important and will be incorporated in this model.

## Discussion

Earlier, I promised to propose my version of comprehension. I have done so in almost every section of the model I have discussed. My definition of comprehension is that amount of information which can be free recalled upon demand after reading a section of prose given the interactive effects of the personal and environmental characteristics and the syntactic and semantic properties which impact on that individual at the time the prose is read. The model presented here is testable and makes intuitive, if not logical, sense. What remains is a great amount of work necessary to verify this model of item readability, but it is possible.

To discuss the details of the methodology to test this model in this context would not fairly outline the necessarily complicated procedures. Suffice it to say that a regression analysis will be used to establish an equation to predict an item's "readability". Each item will have the variance due to its difficulty separated from the variance due to its readability level. The dependent variable will be a reading comprehension score weighted to exclude variance common to Personal Cognitive Characteristics, Other Personal Characteristics, and Environmental Characteristics.

## References Notes


1. Duncan, R. E. The Psychometric Properties of Different Pairing Strategies and Their Effect on Item Readability on a Multiple-Choice Test. Research Proposal; USAFOMC, Randolph AFB, TX, January, 1981.

2. Duncan, R. E. Development of an Item Readability Formula: A New Method with Significant Implications. Research Proposal; Department of Educational Psychology, University of Texas at Austin, January, 1981.

## References


Chomsky, N. Language and Mind (Enlarged Edition). New York: Harcourt Brace Jovanevich, 1972, p. 107

Crawford, Charles. George Washington, Abraham Lincoln, and Arthur Jensen: Are They Compatible? American Psychologist, 1979, 34(8), 664-672.

Dale, E. & Chall, J. S. A formula for predicting readability. Education Research Bulletin, 1948, 27, 11-20 and 37-54.

Dale, E. & Chall, J. S The concept of readability. Elementary English, 1949, 26, 19-26.

Deignan, Gerard M. and Duncan, Robert E. CAI in Three Medical Training Courses: It was Effective. Behavior Research Methods and Instrumentation, 1978, 10(2), 228-230.

Deignan, Gerard M., Seager, Brent, A., Kimball, M., and Horowitz, H. Computer Assisted, Programmed Text, and Lecture Modes of Instruction in Three Medical Training Courses: Comparative Evaluation. AFHRL-TR-79-76, Lowry AFB, CO., June, 1980.

Entin, E. B. and Klare, C. B. Components of answers to multiple-choice questions on a published reading comprehension test: An application of the Hanna-Oaster approach. Reading Research Quarterly, 2, 1980.

Guilford, J. P., Comrey, A. L., Green, R. E., & Christensen, P. R. A Factor Analytic Study of Reasoning Abilities. I. Hypotheses and Description of Tests. Los Angeles: University of Southern California, 1950.

Guilford J. P., Geiger, R. M., & Christensen, P. R. A Factor Analytic Study of Planning. I. Hypotheses and Description of Tests. Los Angeles: University of Southern California, 1954.

Guilford J. P., & Hoepfner, R. Structure-of-Intellect Factors and Their Tests. Los Angeles: University of Southern California, 1966.

Guilford J. P. The Nature of Human Intelligence. New York: McGraw-Hill, 1967.

Hanna, Gerald S., & Oaster, Thomas R. Toward a unified theory of context dependence. Reading Research Quarterly, 1978-1979, 14(2), 226-243.

Herrnstein, R. I.Q. Atlantic Monthly, September, 1971, pp. 43-64.

Herrnstein, R. I.Q. in the Meritocracy. Boston: Little, Brown, 1973.

Jensen, A. R. How much can we boost IQ and scholastic achievement? Harvard Educational Review, 1969, 39, 1-123.

Jensen, A. R. Educability and Group Differences. New York: Harper & Row, 1973, (a).

Jensen, A. R. The current status of the I.Q. controversy. Australian Psychologist, 1978, 13, 7-27.

Kamin, L. J. The Science and Politics of I.Q. New York: Halsted Press, 1974.

Klare, G. R. A second look at the validity of readability formulas. Journal of Reading Behavior, 1976, 8, 129-152.

Klare, G. R. A possible framwork for the study of readability. Proceedings of the Tri-Service Literacy and Readability Workshop, NPRDC SR80-12: March, 1980.

Massaro W. and Taylor, G. A. Reading Ability and Utilization of Orthographic Structure in Reading. Journal of Educational Psychology, 1980, 72(6), 730-742.

Mill, J. S. Autobiography and Other Writings (J. Stillinger, Ed.), Boston: Houghton Mifflin, 1969 (originally published, 1873).

Siegel, A. I. and Bergman, B. A. Readability/comprehensibility as related to the structure of intellect model. In A. I. Siegel and J. R. Burkett (Eds.), Application of Structure-of-Intellect and Psycholinguistic Concepts to Reading Comprehensibility Measurement. AFHRL-TR-74-49, AD-A001 573, Lowry AFB, CO: September, 1974.

Skinner, B. F. Verbal Behavior. New York: Appleton-Century-Crofts, 1957, p. 277.

Thorndike, E. L. The understanding of sentences. Elementary School Journal, 1917, 18, 98-114.

Valentine, L. D., Jr., and Vitola, B. M. Comparison of Self-Motivated Air Force Enlistees with Draft-Motivated Enlistees: AFHRL-TR-70-26, AD-713 638. Lackland AFB, Texas: July, 1970.

Williams, A. R., Jr., Siegel, A. I, and Burkett, J. R. Readability of Textual Material--A Survey of the Literature. Lowry AFB, CO, AFHRL-TR-74-29: 1974.

Wilson, E. O. On Human Nature. Cambridge, Mass.: Harvard University Press, 1978.

Table 1

Components of the Base in the
Item Readability Theory Pyramid

**Item Characteristics**

1. Grammatical structure (syntax) or Surface Structure (Chomsky)
   a. Bloom's taxonomical level
   b. centerembededness
   c. left-branching
   d. word and sentence length
   e. Yngve depth
   f. alternative pairing strategy
2. Semantic Structure
   a. abbreviation or jargon
   b. vocabulary level and familiarity
   c. voice
3. Test reference material - the readability level of text used for question construction
4. Information requested
   a. type
      1. figural
      2. verbal
         a) gist
         b) rote recall
   b. amount of learning time possible
   c. type of source from which material was learned
   d. practiced material
   e. content validity

**Personal Cognitive Characteristics**

1. Reading ability
   a. vocabulary
   b. comprehension
2. Cognitive skills
   a. information processing
   b. cognitive styles
3. Knowledge of subject matter
   a. prior
   b. learned
4. IQ
   a. verbal
   b. non-verbal

Other Personal Characteristics

5. Tool skills
   a. test-wiseness
   b. eye-hand coordination
6. Physical & mental conditions
   a. health
   b. alertness
   c. anxiety
7. Motivation
8. Biographics
   a. sex
   b. race
   c. SES
   d. developmental stage

**Environmental Characteristics**

1. Testing medium
   a. paper & pencil
   b. CAI
2. Test type
   a. achievement
   b. aptitude
   c. diagnostic
3. Testing purpose
   a. mastery
   b. qualifications
   c. selection/ promotion
4. Social climate
   a. differential selection
   b. use of utility model of validity

Figure 1

Guilford's Structure-of-Intellect Model
(Guilford & Hoepfner, 1971)

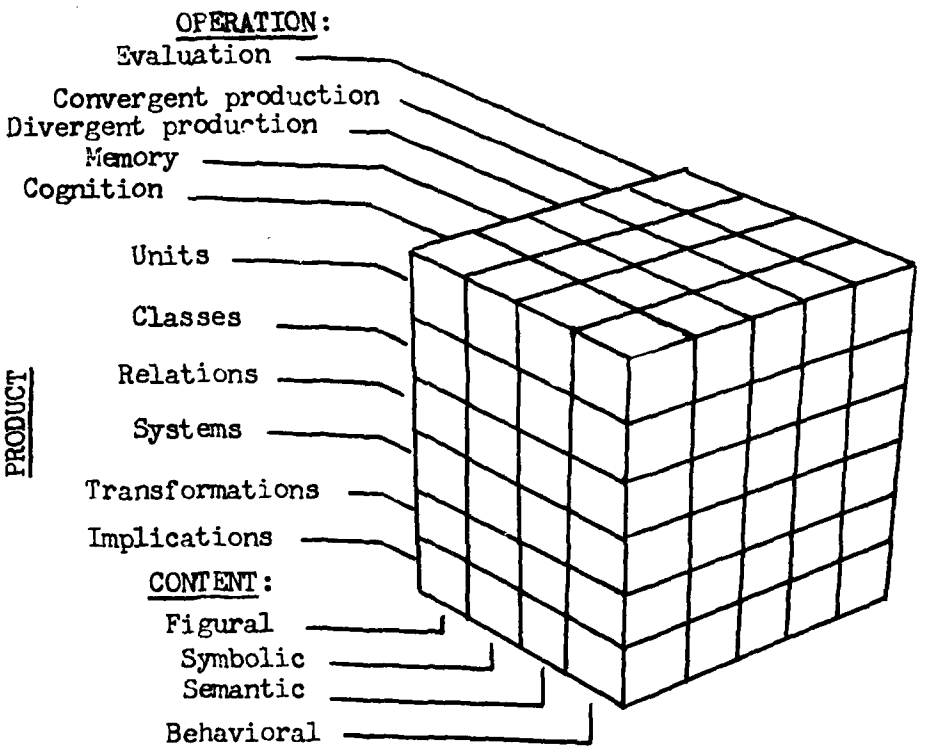


Figure 2
Some major factors interacting with readability measures in validity studies
(Flare, 1980)

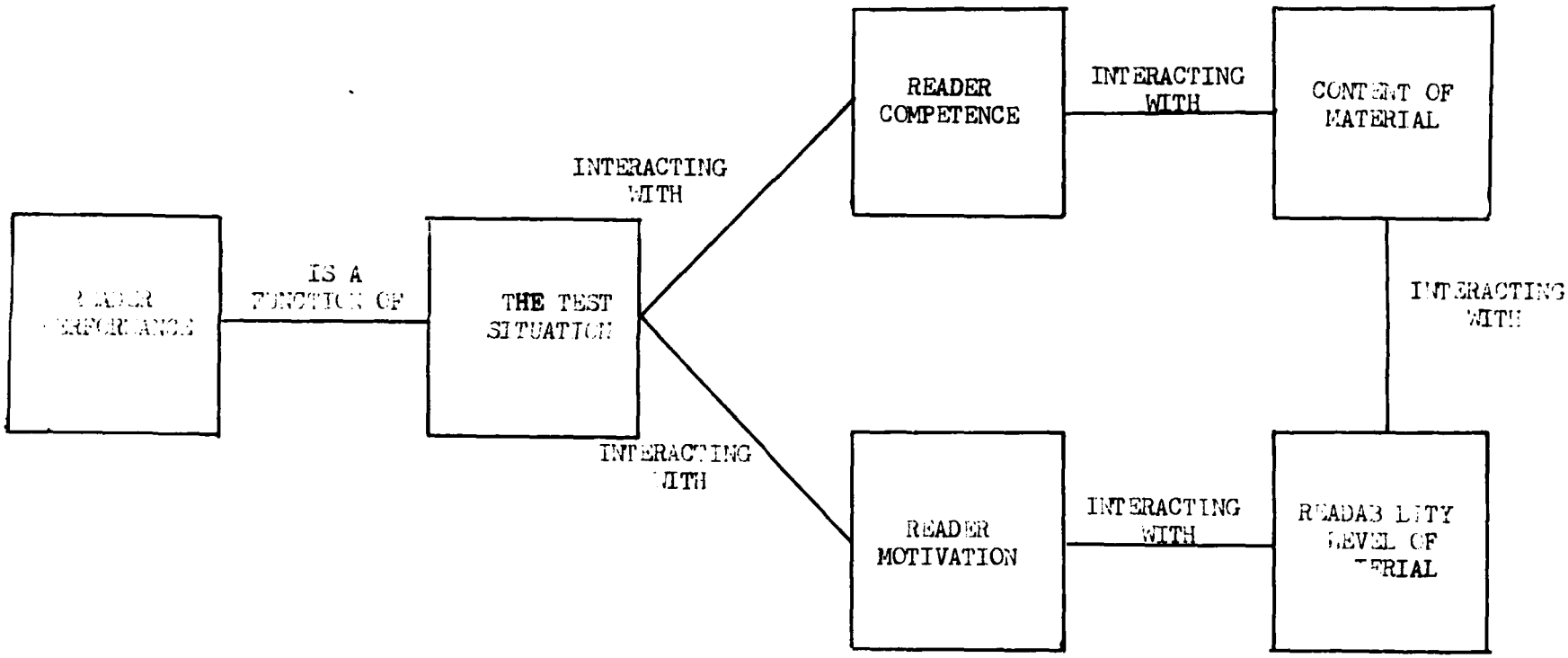


Figure 3

A Model of Item Comprehensibility

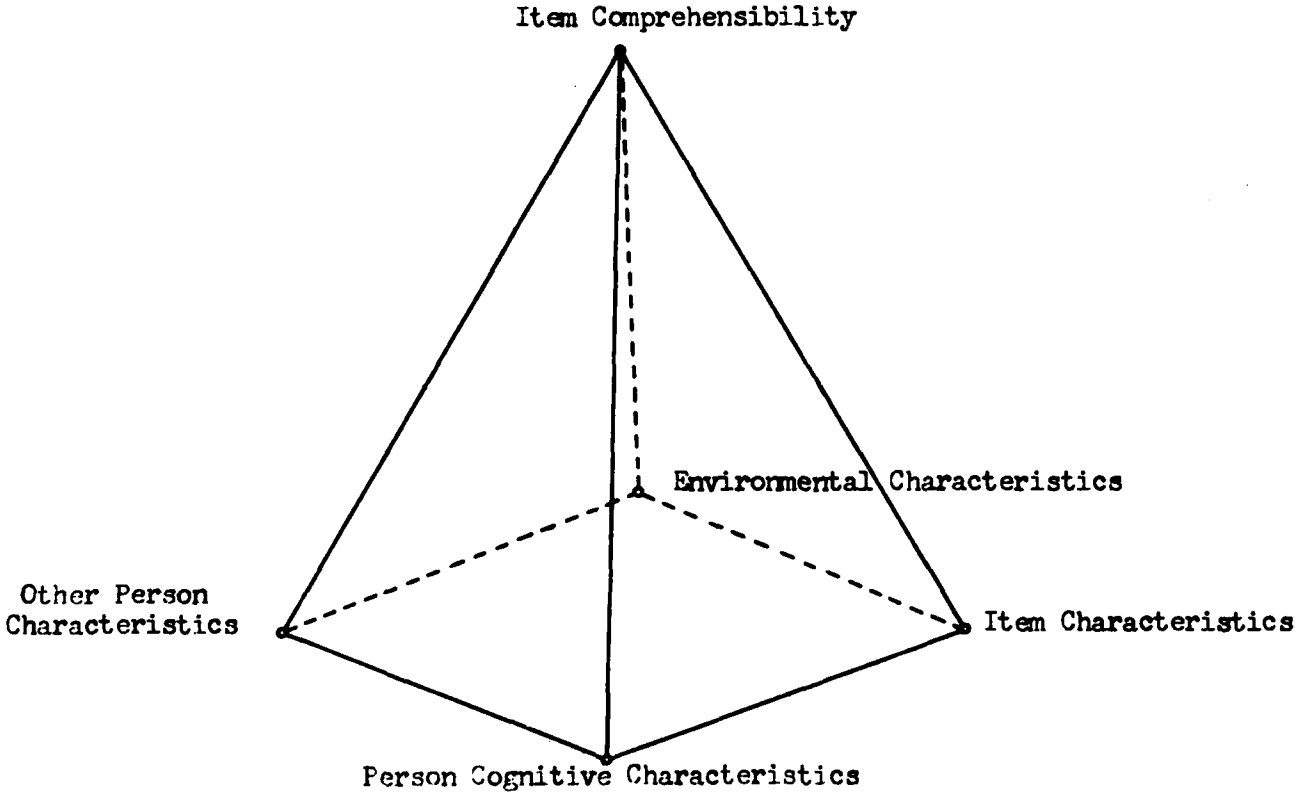
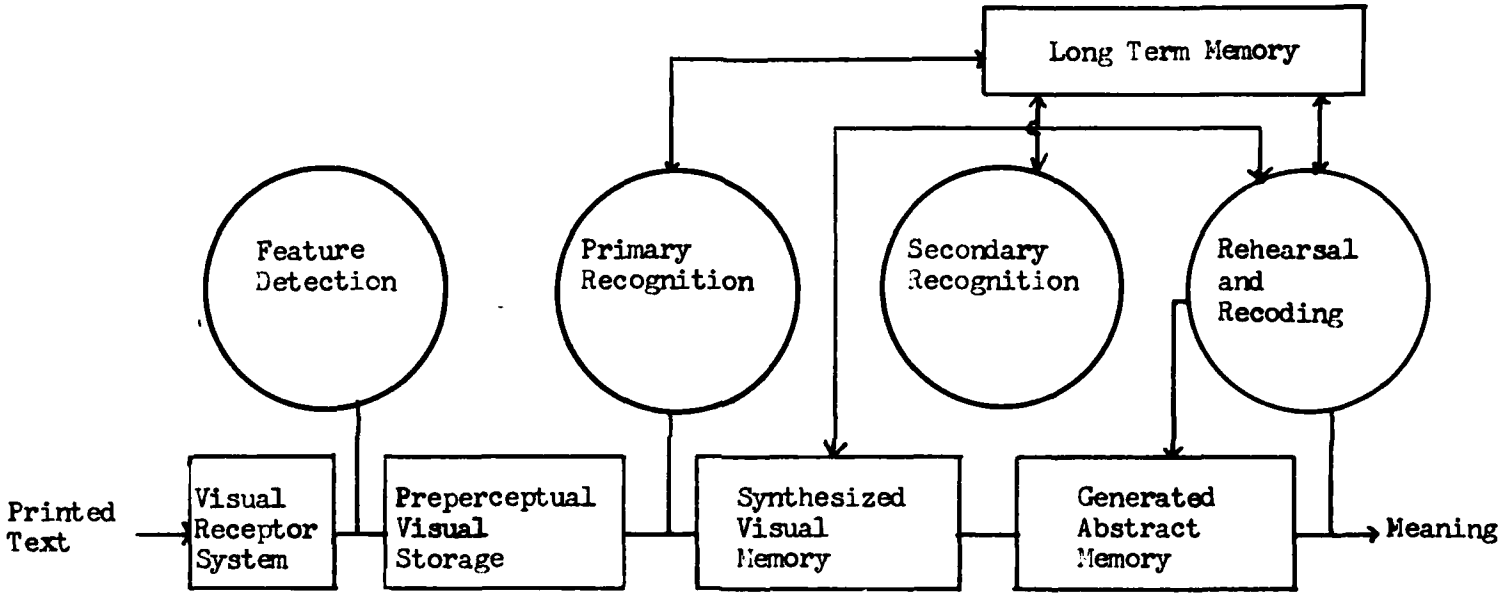


Figure 4
Information-Processing Model of Reading Printed Material
(Massaro & Taylor, 1980)



AD P001410

# SOLDIER READING ABILITY

## The Advocacy Point of View

M. A. Fischl
U.S. Army Research Institute

Presented in Panel Discussion
on
Literacy in the Department of Defense

Military Testing Association
Washington, D.C., 1981

# SOLDIER READING ABILITY

## The Advocacy Point of View

M. A. Fischl
U.S. Army Research Institute

The purpose of this presentation is to assume the role of devil's advocate, and provide information and data which may suggest a broader perspective for interpreting the incidence of so-called "literacy" problems in the military services.

By mid-1977 reports started to reach our office from the Army's Training and Doctrine Command and from its Forces Command, which has cognizance over individuals in operational units, that soldiers did not have adequate literacy skills. We addressed the issue in several ways: one was to consider development of a reading test to serve as a supplement to existing AFQT and ASVAB aptitude area screens; another was to consider modifying the AFQT to increment its reading ability demand; a third was to determine how good a reading screen could be devised directly from the existing ASVAB. This third approach produced some surprises.

In the Fall of 1977, 600 new recruits were administered the USAFI Reading Test, which is a form of the Metropolitan Reading Achievement Test, a fairly universally used and accepted test to evaluate literacy. Also obtained for these individuals were their scores on the 16 subtests of the then operational forms of the basic enlistment battery, ASVAB. Our statistical approach was to regress reading achievement scores on ASVAB scores in order to determine if some combination of ASVAB test scores could serve as a predictor of reading ability.

These correlation coefficients were exceedingly high; that is, the test on which enlistment applicants at that time were screened, correlated about as highly with the reading test as do alternate forms of the same instrument. The visual shows these coefficients.

How, then, might the services be experiencing so-called "literacy" problems when screening on a measure which is as good an indicator of reading ability as the reading test itself?

a. Reports also reached our office about breaches of ASVAB security. Although, as those r's demonstrate, it is virtually impossible to be illiterate and pass ASVAB fairly, test compromise--knowing some ASVAB answers ahead of time such as occurred in 1978--could have resulted in illiterates being admitted.

## ASVAB & LITERACY

Correlation with USAFI
Reading Test, Form D

N = 600

| ASVAB Test | r | Corrected for Reading Test Unreliability: If Reliability = 0.95 | Corrected for Reading Test Unreliability: If Reliability = 0.90 |
|---|---|---|---|
| AFQT | 0.78 | 0.80 | 0.82 |
| WK | 0.79 | 0.81 | 0.83 |
| WK + MK + SI | 0.84 | 0.87 | 0.88 |
| All ASVAB | 0.90 | 0.93 | 0.95 |

b. Later, reports reached our office of ASVAB norms yielding overestimates of qualifications. Army enlistment standards at the time utilized the total battery, setting cuts half a standard deviation below the mean (31st percentile) on 9 job related measures, buttressed by a 16th percentile on the general measure. Although, again, it is virtually impossible for someone who scores within half a sigma of the mean to be weakly literate, ability levels that might in fact be a full sigma or more below the mean would be a very different story.

The bottom line is that it is quite probable that most of the presumed "literacy" problems were not that at all, but were in reality test compromise problems or test norming problems.

---

Note should be taken that the forms of the ASVAB described in this presentation were subsequently replaced and have not been operational for almost two years.

Author: Richard P. Kern
For: Presentation, Literacy and DOD Reading Research Panel Discussion
23rd Annual Conference of Military Testing Association, 20-25 October 1981

ABSTRACT

AD P001411

# WHAT ARE THE LITERACY COMPONENTS OF JOB PROFICIENCY?
## An Objective, Opinionated Commentary

Military researchers in the areas of selection and assignment, training, and unit leadership and management have the common goal of improving job performance. Of necessity, each area of research approaches this goal from a different perspective. However, in the final analysis we are all dealing with the same nucleus, the individual enlistee, and the ability of these personnel to learn, perform and maintain their proficiency in performing their prescribed military duties.

Recently, the role of literacy in affecting military efforts to select, assign, train, lead and manage individuals and units has become a prominent issue. My objective in this paper is to identify what I perceive as serious pitfalls in the concept of literacy as presently reflected in applied military research as well as in operational practices. Support for this perspective will be offered based on research literature and findings from my own recently completed research. I will then propose a direction for future research. The goal of this research is to substitute a more diagnostic, functionally-related concept of the individual's job-related knowledge base for the present, relatively uninformative concept of general literacy.

# WHAT ARE THE LITERACY COMPONENTS OF JOB PROFICIENCY?
## An Objective, Opinionated Commentary

My paper addresses two questions: What is a job reading task? What is the purpose and potential value of tests built on this concept?

What is a job reading task?

The concept of a job reading task that I am going to present grew out of research conducted for the Army by Tom Sticht of HumRRO (1975). Major objectives of their research program were to develop ways of estimating minimal reading skill requirements of Army jobs and to develop alternative approaches the Army could use to effectively utilize low aptitude recruits. In addition to conducting research on the use of information provided by general reading tests and readability formulas, Sticht and associates interviewed job incumbents in their work settings to identify specific instances in which they had used printed materials to assist them in performing their work. For each incident cited, the interviewee was asked to describe the task or activity he had been engaged in and the information he had been seeking when he went to the printed material. In addition, he was asked to obtain the manual and show the interviewer the portions of the material he had used in attempting to obtain the information he had been seeking. Information obtained in this way was then used as the basis for identifying job reading tasks.

In the course of this research a model of a job reading task was developed. In this model a job reading task occurs when the worker uses visual information displays to resolve a question or otherwise obtain information to support performance of his task activities. A specific job reading task is defined in terms of both the question or nature of the information the person was seeking and the specific material used in attempting to obtain the information. Description of a job reading task as it occurs in the job setting contains at least the following

elements:

Condition: Worker performing task activity in the "natural" job setting

1. Recognizes need for information
2. Formulates question
3. Identifies a likely source
4. Locates source

   (if manual)

5. Searches manual for information responsive to question
6. Locates and processes print for information
7. Resumes task activity, applying (accepting/rejecting) information

Construction of job reading task tests.

As in dealing with most problems in life, there are two approaches to developing job reading tests based on job reading tasks, the ideal and the practical approach. Ideally, all one would have to do to construct a job reading task test would be to sample the job reading task incidents described in interviews such as those obtained by Sticht. Then using the model of a job reading task as a guide, you would simulate the full situation for each job reading task in performance test format. This, obviously, is not a feasible approach.

A practical approach is to obtain job reading task incidents and then classify the questions asked and the materials used. Assuming that questions asked remain cross-indexed to the materials used during this classification process you can then sample on the basis of either question or material classifications to form subtests for a job reading task test. In using this approach you are providing the "question" in the test format and a replica or copy of the actual job materials. As a result you have restricted the test to only those elements of the job reading task (described above) that occur after the worker has located the source of information. The test is now degraded as a test of actual job reading behaviors if for no other

reason than the fact that a "teacher" is not generally available on the job to tell the worker what question he should be asking and to provide him with the specific material needed to answer the question. Be that as is may, the degraded form still permits one to examine how well personnel can perform different types or sub-sets of question-material combinations which can be expected to be encountered in the job setting.

At this point I am going to put other methodological problems on "hold" in order to focus on a more important issue for advancing development of tests based on job reading tasks. Put simply, this issue is what is the purpose of these tests?

Why develop job reading task tests?

Research in this area in the military was funded under the pressures of Project 100,000 and the concerns over effectively utilizing recruits scoring in the lower range of aptitude scores. It has been my impression that when the words literacy or reading skills pop up commanders and administrators automatically tune to visions of low aptitude, functionally illiterate, personnel and remedial reading programs to teach phonics, word recognition and vocabulary. The targeting of remedial reading programs for those at only the extreme low end of tested reading ability appears to have "chicken-and-egged" (I don't know which came first) two perceptions of adult literacy or reading "problems". One view is that remedial reading training has not and can not demonstrate gains in reading ability that have any practical affect on improving learning and performance of military jobs. The second perception is that adult reading competency is a distinctly bimodal phenomenon. Under this perception an adult is either competent in meeting reading requirements of training and job performance or is functionally illiterate in all areas.

It is my contention that the continued reliance on reading tests designed to yield only a global screening or school grade placement score is responsible for

reinforcing and maintaining the pessimistic and restrictive perceptions of adult reading competencies I have cited above. As a result of our continued use of these tests we have very little data on how adults perform reading tasks that can be related to the use of reading to deal with the requirements of training, job performance, and general military life events. Aside from the work performed by Sticht, there has been little research in the military devoted to this objective. However, even his job reading task tests have been designed to yield only a global screening score.

With the above considerations (and opinions) in mind, I am proposing the following applied objectives for use in designing tests based on job reading tasks. I am offering these objectives to solicit this panel's consideration and discussion of their merits for moving the topic of job-related reading out of the remedial reading school and into the broader area of cognitive psychology and its possible contributions to improving training and job proficiency.

Applied Objective Proposed for Use in Designing Reading Tests Based on Job Reading Tasks:

1. Predict how well individuals can handle different types of job-specific reading requirements prior to job training.

2. Assist training developers identify and use reading task requirements in establishing training objectives and standards for course development.

3. Assess how well trainees can handle the different types of job reading requirements at end of training; Predict how well they can handle these types of job reading requirements when on the job.

4. Assist developers of manuals or other visual information displays to identify anticipated levels reading competencies for different types of job reading tasks.

Reference:

Sticht, Thomas G. (Ed.) Reading for Working, HumRRO, Alexandria, VA 1975.

AD P001412

# Job-Related Measurement of Reading Ability

Sandra S. Payne
Personnel Research Psychologist
Personnel Research and Development Center
Office of Personnel Management

Reading Comprehension is a requirement for satisfactory performance in a variety of jobs, and consequently a test of reading ability is frequently used in job selection. Because of recent increased emphasis on the job-relatedness of selection devices, it is desirable to have the reading ability test be as closely linked to actual job reading requirements as possible. This paper describes a set of procedures for developing a clearly job-related test of reading comprehension. The procedures include steps for sampling and measuring the reading level of materials used on the job in order to define the specific level of reading ability required. The test is then designed to match the required job reading level. The procedures are flexible, allowing for design of a test for use as a minimum competency measure or as a ranking device. Test questions can be written directly from the sampled job materials or from general materials. The procedures are presently being used for a variety of Federal test projects and a discussion of problems, improvements, and results from some of these projects are discussed.

Reading comprehension is a requirement for satisfactory performance in a variety of jobs, and consequently a test of reading ability is frequently used in job selection. Because of increased emphasis on the job-relatedness of selection devices, it is desirable that the reading ability test be as closely linked to actual job reading requirements as possible.

To meet this goal, we have developed a set of specific procedures for our use in the Federal examining program in the Personnel Research and Development Center of the Office of Personnel Management. The procedures include steps for sampling and measuring the reading level of materials used on the job in order to define the specific level of reading ability required. The test is then designed to match the specific job reading level. These procedures draw on the experience of both myself and other researchers in developing reading comprehension tests and in determining the reading level of written materials. The procedures were first applied in a test development project I conducted in 1976, and after some refinement, are being used in several different Federal test development projects now in progress. As new problems are encountered, the procedures are being modified and hopefully improved where necessary.

The procedures are flexible, allowing for design of a test for use as a minimum competency measure or as a ranking device. Test questions can be written directly from job materials or from general subject matter materials.

## Planning for Test Development

An initial assumption at the start of these procedures is that the reading comprehension test is only one factor in a complete selection process, and furthermore, it is assumed that the complete selection process has been carefully planned and is based on a study of the job requirements. Consequently, these procedures are not intended to provide documentation of the inclusion of a test of reading ability in the overall selection process. These procedures do, however, document the job-relatedness of reading comprehension test.

There are three major steps in the test development process: (1) identifying the reading materials used on the job; (2) determining the reading level of those materials; and (3) preparing and assembling the test questions.

Carrying out these steps requires assistance from three sources. The first source is a qualified test development specialist whose role is to guide the entire process and be responsible for the technical aspects of test development, assembly, and documentation. The second source is job incumbents, trainers, and supervisors of the job for which the test is being developed. These people, called "subject-matter experts," and hereafter referred to as "SMEs," are responsible for identifying the reading content of the job, determining the reading level, and preparing the test questions. The third source is the higher-level management of the agency (or agencies) in which the jobs exist. Through discussion and cooperation with the test specialist, management is responsible for determining the job(s) to be covered by the reading test, for selecting the SMEs, and for overall support of the test development effort. Strong management support is critical to the successful completion of the project, as considerable SME time away from regular duties is required.

The bulk of the work is done by the SMEs. We have developed our procedures to rely on the SMEs for two major reasons. First, to identify the required reading materials obviously requires job knowledge experts. It would be highly inappropriate to conduct this step from the personnel office, regardless of the depth of available job information. Second, the use of SMEs to determine the reading level and to write the test questions has an even more pragmatic base. Most personnel offices do not have the available staff to assign a large group of test technicians to a single test project. The only other available source of workers is the job incumbents themselves.

The selection of qualified SMEs is essential to the successful conduct of the test project. The SMEs who carry out the first step, identifying the required written job materials, should all be job incumbents or first-line supervisors of the job. Job training specialists may also be included when there is extensive formal training involved during the first months on the job.

If the reading test is to cover more than one job, or if there are several distinct specialties within the job, the SMEs should be chosen to

represent these different jobs or specialties. The SMEs should be experienced, so that they have sufficient knowledge of job activities that may occur less frequently than others.

The SMEs should also be chosen to reflect the ethnic, racial, and sex composition of the relevant work force. If there are no members of large racial or sex categories in the job at present, then special effort should be made to have the work done on the project reviewed or augmented by knowledgeable personnel specialists or other persons sensitive to the concerns of these groups.

The same group of SMEs can carry out the second and third steps of the test development procedure, or different groups of SMEs can be assembled. The SMEs who do the second and third steps do not have to be as carefully representative of the various job specialties as did the SMEs doing the first step. Training specialists, central office personnel, or staff members with specific question writing experience in their backgrounds are examples of the type of SMEs who could be selected. However, they should still be familiar with the written materials used on the job, the terminology associated with the job, and other aspects of the job.

The number of SMEs to be assembled to carry out each of the three steps will vary according to the amount of time allotted for completing the work, and the number of steps assigned to the group to do. In any case, there should probably be at least four group members, and there should probably not be more than ten members. The size which best allows for both optimal group interaction and variety of input is probably five to seven members.

Regardless of the number of SMEs assembled and which step is to be carried out, the SMEs should be chosen for their verbal abilities and skills. Also, it is important that one SME be assigned as the group leader. The leader is responsible for maintaining the SMEs' cooperation, and encouraging their best efforts at all times. The leader also works with the test specialist to keep the project on track.

## Logistics and Test Security

Attention to details concerning work arrangements, scheduling, and security of test materials will make the test development process move smoothly. To this end, a detailed project processing schedule should be drawn up and adequate workspace and typing support should be arranged before the group of SMEs meet to begin their work. Arrangements for security must be made, and the security of the test materials should be emphasized at the first meeting of the SMEs. Security is important in all three steps of the procedure, not just when test questions are being written.

## Special Problems of Non-Assembled SMEs

We prefer that the SMEs assemble in one central location and complete all work at that time. Occasionally this is impossible, and work must be done individually. The procedures can be carried out under these circum-

stances, but the time to complete the project is greatly lengthened, and security becomes more difficult to control. Some suggestions for dealing with these problems will be given when appropriate as the steps of the project are described.

Step 1: Identification of Job Reading Content

The first step in the procedure is to define the reading content of the job. The definition consists of a comprehensive list of all written materials which are used in the performance of the job. Only those materials which are actually necessary for work performance are included. Written materials which provide extra information, or nice-to-know information, or are used only by supervisors are not included.

To prepare this complete list, the SMEs first develop a preliminary list of possible job materials. This is effectively done in a brainstorming session, with each SME suggesting any written job materials that come to their minds. During this initial stage the list includes materials that may not be required by the job incumbents. If there is formal job analysis information available, it can be reviewed as a source of suggestions of materials used to perform different parts of the job.

Typical materials to be named would included rulebooks or manuals describing work procedures, reference materials like price catalogues, job training manuals, and incoming materials like letters, claims forms, order forms, requests for assistance, etc.

After the preliminary list is developed, the group of SMEs should individually rate each material on the list for its relevance to the job. The criteria shown in Figure 1 are used for this purpose. Majority rule prevails, and the result will be a list of materials which the SMEs agree are required to perform the job for which the test is being developed.

The reading content of the job can be identified if absolutely necessary without assembling the SMEs together. The best approach is for the test specialist, in conjunction with knowledgeable agency personnel, to prepare a preliminary list of possible job materials. This list is then mailed to the SMEs, along with instructions to rate the relevance of each of the materials, and also to add any materials which are missing. The ratings are then mailed to the test specialist, who compiles the ratings and prepares the final list of required written job materials. If there are major discrepancies between the ratings, or if many different materials are added to the lists, then a second preliminary list must be prepared and distributed for ratings. In addition, telephone calls must be made to solve major differences in ratings on the same materials. This process usually requires at least a month to complete.

Step 2: Measurement of Reading Level

The next step is to determine the reading level of the written materials used on the job. There are many formulas that have been devised to estimate

The term READING MATERIALS includes handbooks, pamphlets, forms, letters, manuals, catalogs, etc.

The term ENTRY-LEVEL refers to people who have been on the job one year or less.

ESSENTIAL READING MATERIALS:

| | |
|---|---|
| ARE | actually necessary to the performance of the entry-level job |
| ARE | required for successful completion of any required training program |
| ARE | any materials the position description for the entry-level job mentions as necessary |
| ARE | available to all entry-level job incumbents |
| ARE | materials the entry-level job incumbent cannot do the job without reading |
| ARE NOT | materials used to help prepare the job incumbent for promotion |
| ARE NOT | materials some job incumbents may use to help them understand other materials |
| ARE NOT | materials which are supplementary to the reading required in any required training program |
| ARE NOT | materials that would be good for the entry-level job incumbent to read if he or she had the time |

Figure 1. Criteria for identifying essential job reading materials.

or judge the reading level of written material. The formula used in our procedures is the Flesch Reading Ease Index. This formula is based on sentence length and number of syllables per hundred words in samples from prose passages (Flesch, 1948). It was chosen because of the strong research support for its validity and reliability (England, Thomas, & Paterson, 1953; Gilinsky, 1948; Hayes, Jenkins, & Walker, 1950; Hoffman, 1972; Peterson, 1956; Swanson & Fox, 1953), and because it could be easily adapted to measure the the reading level of written multiple-choice type items, which is necessary in the test development stage of the project.

Use of the Flesch formula results in a single numerical index which estimates the relative reading ease level of the material which is analyzed. The index ranges from 0 to 100. The higher the index, the easier the material it describes.

I believe that the actual method used to measure reading levels is less important than the way in which the results are used. Early research attempted to relate a measured reading level to a particular educational level, or type of material (Klare, 1963). Such relationships should be viewed with caution, however. For instance, attainment of a particular grade level in school does not necessarily mean the same in terms of acquisition of reading skills from school to school, or from locality to locality. Consequently, the most appropriate use of a reading level index is to compare the relative size of the measurements obtained from analysis of different materials. Differences in levels can thus be detected or adjusted as necessary, without regard to what the levels are actually stated to be.

Before the computation of the Flesch indexes can be started, a sample of all the written material included on the final list of required job reading must be compiled. These samples should be machine-copied and gathered in one central location. If the procedures described in this paper are followed completely, these machine copies will be used twice--first for counting the raw data for computing the reading ease levels, and second as the basis for the actual test questions. A copy should be made of every page of short materials (say, 10 to 20 pages), and a careful sampling plan should be followed for longer materials (say, every other page for materials from 20 to 50 pages; every tenth page for materials which contain more than 500 pages, etc.)

After the reading material samples are gathered, the second group of SMEs begins their work. They must count the raw data for computing the reading levels for each sample of the reading material. This is tedious and exacting work, and requires careful checking. The raw data is recorded right on the machine copies of the reading materials. The test specialist computes the reading level for each sample directly on the same copy, and records it at the top of the page. The individual reading levels are then combined to determine a reading level for each material on the final required reading list, and also an overall average reading level for the job.

The reading level counts cannot be carried out by an unassembled panel, as the work of sampling and computation is constantly interrelated. Also the complete sample of reading materials should be assembled in one place for security reasons.

The length of time required to do the reading level count will depend upon the amount of required reading material and the number of people who are doing this work. The work is tiring, and can only be performed for short periods of time before accuracy falls off drastically. Experience has shown us that one person can count approximately 25 to 40 samples per day, again

depending upon the complexity of the material. Then time has to be allowed for checking the counts, and computing the reading level for each sample. Finally, time has to be allowed for computing the indexes for each complete material, and the overall average for the job.

As I developed these procedures, I considered the issue of weighting the reading levels for different materials based on the relative number of pages included in each material, or on the relative frequency of use of each material, when computing the overall reading level for the job. My final conclusion was that a complicated weighting scheme was not only unnecessary, but also inappropriate. The important factors to consider are: (1) the range of reading levels required on the job, which is adequately described by determining the reading level of each material included on the final list; and (2) the overall reading level of the job, which is adequately described by determining the average of all the required materials. Because one material is used more frequently than another does not automatically mean the less frequently used material is of lesser importance. Also, because one material is shorter than another does not automatically mean it is less important than a longer material. Unless a very complicated rating procedure is followed in the identification of required reading materials step, introducing a weighting scheme at this point would not be justified. I also do not believe a complicated rating scheme would add anything to the process.

Step 3: Development of Test Questions

The final step in the process of preparing the job-related reading comprehension test is to prepare the test questions. This step can be carried out by the previously assembled group of SMEs, or by a new group.

As indicated earlier, there are two major sources for the reading comprehension test passages. One is specific, the written material used on the job, and one is general, any written material in the appropriate range of reading difficulty.

If test questions are to be written from job materials, they should be based on the passages used in the reading level analysis. If one or two of the written materials are found to be quite a bit more difficult or less difficult than the average for all the job materials, then it is not appropriate to use these reading passages as the basis for test questions. Although the reading level of an individual test item is affected by the response alternatives as well as the base paragraph, the restriction of paragraphs to those near the average reading level will help to keep the finished questions in the desired range.

If confidential materials are included in the reading level samples, they should not be used as the basis for test questions. Some care should be given to selecting other reading samples that are similar in style and difficulty to these confidential materials.

The passages should not be changed in any substantial way when test questions are written. However, clearly incorrect grammar, syntax errors, or outdated sex or race references should be corrected.

Each SME prepares draft questions based on a portion of the reading material samples. More than one question can be written on the same passage, and conversely, some passages may not be suitable as the base for even one question. A fairly even distribution across all samples is a good goal, however.

I recommend that job materials be used as the basis for the test questions. These materials are readily available, and their use adds considerably to the clear job-relatedness of the completed test. General rules for constructing objective written test questions should be followed by the SMEs. We provide a day of item writing training before the SMEs begin to write questions, and also closely monitor progress for the first few days of individual question writing.

Draft questions are reviewed and edited by the test specialist on a continuing basis to determine if the item is psychometrically sound and grammatically correct. The review also includes the identification of specific job terminology which may not be clear to non-experienced job applicants.

Reviewed questions are then either returned to the original author for rewriting, or placed into a finished question file. When questions are in final form, the reading level of the complete question is determined. Special rules are followed to apply the Flesch formula to a multiple choice test question (Payne, 1976).

If test questions are to be written from general subject matter materials, rather than job materials, it is unlikely that SMEs would be used to write these questions. They may be asked to review already prepared question files, however, to see if the questions are on appropriate subject matter, and are of the approximate difficulty level desired. If questions are written from general materials, the same procedures would be followed for construction, review, and determination of reading level.

It may seem that computation of a reading level would not be necessary for reading questions taken directly from the job materials. However, the reading level of a multiple-choice test question depends in part on the complexity of the choices, or alternatives, given in the question, in addition to the reading level of the paragraph. Also, the final test could inadvertantly include more passages from the more difficult sample materials, or vice versa.

## Determining Difficulty Levels of Items

Since we do not generally have the time or the available sample to pretest the questions developed for the reading test, we use the SMEs to estimate the relative difficulty of the questions they have written. To do this, we give the SMEs a set of reading comprehension items with known p-values to use as benchmarks, and ask them to roughly categorize the difficulty levels of the new test questions.

## Test Assembly

After the test question file is completed, the test is assembled in a standard manner, in accordance with the test plan. The test plan should state whether the test is to be used as a minimum screen or as a ranking device.

In either case, the average reading level of the test should be set at the average reading level found for all the job materials.

If the test is to be used as a minimum screen, all of the test questions should be set at or near this average reading level. Some range in reading level is acceptable, but individual questions should generally not vary by more than $\pm$ 10 points on the Flesch index. For example, if the average reading level for the job is a 57, then the reading level of questions selected for the test should fall within a range of 47-67. The range of questions should be consistent for each of the test forms assembled.

If the test is to be used as a ranking device, a wider range of reading levels is appropriate. The range can be as wide as that found in the analysis of job materials, or it can be slightly narrowed to reduce the effects of one or two extremely different job materials. The average reading level of the test must remain the same as the average reading level of the job. The use of a wider range of reading levels will allow for greater variability in scores, which is appropriate and necessary for a valid ranking device. The same range of reading levels should be used for each test form assembled.

Length of Test

Although 75 to 100 questions would be desirable to assure reliability, a 50 or 60 -question test composed of well-constructed, carefully reviewed questions is probably sufficiently long. The tests we are currently developing at OPM are 60 items long.

Time of Test

The reading comprehension test is designed as a power test, not as a speeded test. Consequently, enough time should be allowed for all except the slowest applicants to complete the test.

Setting Passing Points

If the test is to be used as a screening device for minimum reading ability, then a passing point must be set for the test. This passing point may be set rationally, as there are few completely satisfactory methods for setting passing points. One approach may be to consider a correct score of 50% of the items as the passing point. The test can also be administered to job incumbents, and the passing point set at the lowest score obtained by incumbents who have evidence of performing satisfactorily on the job. The passing point should not be set lower than the score which can be obtained purely by chance, however. Neither should passing point be set much higher than 70% of the questions, even if the lowest scoring job incumbent answers a greater percentage of items correctly. The essential fact to remember in setting a passing point is that there be some rationale or reasonable explanation documented for whatever passing point is set.

If the test is to be used as a ranking device, it should not be necessary to set a passing point. If you are required to do so (as we are at OPM), then the minimum passing point is set in much the same way as for a screening test.

## Problems

Each time we have used this procedure to develop a reading test based on job materials we have encountered some new problems. Most of these problems center around difficulties with gathering the SMEs in a central location. When the required reading materials are identified by a mailout process, it is inevitable that some rating forms are late, or completely lost. Also, without direct it is harder to communicate the purpose of the procedure, and the data is consequently not as strong. The time is greatly extended, particularly when there are differences to iron out. We have learned that when the SMEs cannot be assembled, it is worth trying to convince the agencies to pay for the test specialist to travel to the SMEs to gather the data directly.

Other problems center around the measurement of the reading level. This is exacting work, and it is hard to keep the group concentrating. The only solution solution is to attempt to get the group to see the importance of their contribution to the selection of future co-workers. This work is also very time-consuming and can ruin the best-laid project processing schedule.

Finally, if the SMEs chosen to write test questions do not have adequate verbal skills, the work may go very slowly, and the quality of the completed test questions may be very poor. This can only be solved before it occurs, by emphasizing the need for qualified SMEs to management, before they are selected.

## References


England, G. W., Thomas, N., & Paterson, D. G. Reliability of the original and simplified Flesch reading ease formulas. Journal of Applied Psychology, 1953, 37, 111-113.

Flesch, R. A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221-233.

Gilensky, A. S. How valid is the Flesch readability formula? American Psychologist, 1948, 3, 261.

Hayes, P. N., Jenkins, J. J., & Walker, B. J. Reliability of the Flesch readability formulas. Journal of Applied Psychology, 1950, 34, 22-26.

Hoffman, J. P. Readability: the measurement of reading difficulty levels of written tests and test related materials. Harrisburg: Pennsylvania State Civil Service Commission, Division of Research and Special Projects, August 1972.

Klare, G. R. The measurement of readability. Ames, Iowa: Iowa State University Press, 1963.

Payne, S. S. Reading ease level of D. C. Fire Department written materials required for entry-level job performance (TM 76-12). Washington, D. C.: U. S. Civil Service Commission, Personnel Research and Development Center, August 1976.

Peterson, J. J. Comparison of Flesch readability scores with a test of reading comprehension. Journal of Applied Psychology, 1965, 40, 35-36.

Swanson, C. E., & Fox, H. G. Validity of readability formulas. Journal of Applied Psychology, 1953, 37, 114-118.

# DEVELOPING JOB-RELATED DATABASES FOR COMPUTER-ASSISTED READING INSTRUCTION

Robert A. Wisher
NAVY PERSONNEL RESEARCH and DEVELOPMENT CENTER
San Diego, CA 92152

Summary The paper summarizes a research project that developed a generative CAI procedure for improving reading comprehension skills of Navy recruits. The procedure utilized paragraphs drawn from training manuals; the computer-based training exercises involved word prediction, sentence arrangement and recall, and paragraph organization. The philosophy behind the instructional software was to reduce the human's role in courseware development time by employing string-oriented software that operated on databases of paragraphs.

## PROBLEM

The trend towards computer technology as a training device presents opportunities for automating not only the delivery of instruction, but also the creation of instructional materials. Emphasis has deservedly been placed on automating the delivery of instruction in order to demonstrate first the efficacy of the computer as an instructor. With this in hand, the time has now come to automate, in increasing steps, the development of computer-based instructional materials. Authoring languages, *such as Coursewriter, Tutor, and Planit, are availabe* to assist the courseware developer in formatting questions, matching answers, and keeping records; or, in general, prepare pre-conveived materials into a format amenable to computer delivery. Although authoring languages are a valuable asset for "bringing up" instruction, they nevertheless require overhead expenditures related to the time to learn the authoring language, prepare the materials off-line, and input those materials, via the authoring language, to the computer.

For some instructional approaches, authoring languages can be ideal vehicles. Depending on the relationship between the CAI program, or process, and the structure of the instructional materials, or data, alternatives to authoring langauges are available. The delivery of reading comprehension instruction by computer presents unique opportunities for taking advantage of a process-data relationship due to the linear, symbolic nature of language. The concept developed here involves domain-independent, string-oriented software (process) that operates on paragraphs of text (data) extracted from training manuals.

## READING COMPREHENSION

In its most simplistic form, language may be viewed as a string of symbols; language comprehension, then, is the task of decoding this symbol string and mapping the decoded form into an internal representation, or in the parlance of linguistics, a deep structure. It is this deep structure, and not the symbol string, that a reader comprehends; the symbols give only clues to the meaning.

The task of reading comprehension is, simply stated, an exercise of applying various cognitive processes to text in order to arrive at a deep structure. Although there are other factors that distinguish good from poor readers, good readers are effective, and efficient, at executing the cognitive processes; poor readers are not. Other factors, such as the quantity and quality of knowledge (schema) one has about a topic, or perceptual processes, are not considered in this approach.

## INSTRUCTIONAL SOFTWARE

The instructional approach was to force the student to make decisions about text while interacting with the computer. If the computer-controlled decisions simulated some of the decisions made in real-time during reading, many of the cognitive processes would be exercised and, presumably, improved in terms of execution efficiency. The instructional sequence entailed having the student read a paragraph and engage in three exercises.

Missing Word Each sentence would reappear individually on the computer screen with one word randomly deleted. The task was to select, from among distractor words randomly drawn from other sentences, the word that appeared in the original sentence. This task exercised those cognitive processes related to anticipating word meaning and utilizing syntactic constraints.

Sentence Arrangement Both philosophic and linguistic theories of language have discussed how the meaning of a sentence is composed from the meanings of individuals words. This exercise involved selecting, from a 3 by 3 matrix of words and phrases, the sequence of cell entries that, when combined, constituted a previously read sentence. By interacting with a keypad, the student would "build" the sentence from its constituent words. Prior to the exercise, the computer would parse each sentence, randomly arrange the entries into the matrix, and fill the missing cells with words or phrases randomly drawn from other sentences.

Paragraph Organization The final exercise drew on the structural contraints imposed by a paragraph. The student's task was to arrange a set of randomized sentences into a meaningful order, as they appeared in the original paragraph. This exercise forced the student to recall and assemble the sequence of ideas as expressed in the paragraph under inquiry, paying attention to such constraints as pronominal reference, topic, and scope.

In order to make the instructional procedure more demanding, there were instances where the exercises would be attempted prior to reading the paragraph. This forced the student into a generative mode, making decisions about meaning and form with fewer clues. Other aspects of the CAI system, such as the use of a voice synthesizer to provide procedural and evaluative information, and the token economy invoked that gave the instruction an arcade-like, gaming environment, will not be pursued here. The system was developed on a PDP-11/34 minicomputer in the "C" language.

## Formative Evaluation

Thus far, sixty Navy recruits have participated in the course. The recruits averaged a completion rate of 21 six-sentence paragraphs in five hours of instruction. After completion of the course, they were given a criterion-referenced test that was also adminstered to a control group of recruits participating in a classroom version of comprehension instruction. The preliminary analysis indicated that, compared to the classroom control group, the CAI-recruits achieved criterion in about half the time.

## Implications for Military Training

It is important to note that the traditional CAI and workbook employment of comprehension questions was never used in this approach. Although such questions serve a useful purpose in assessing comprehension, they task the course development process: somebody must take the time to develop and evaluate the questions, specify the correct answer, and generate useful distractors. When used in the CAI mode, these questions and answers must be input to the machine.

The present concept requires only paragraphs from the course developer. These are entered into the computer by keyboard entry with a sequential file name associated with each paragraph. Currently there are restrictions as to the paragraph size, no more than seven sentences, and the number of words with a sentence, no more than twenty-seven. It is possible to automate this paragraph-entry procedure by downloading magnetic tapes containing the contents of a technical manual (as is used in the automated printing process). With the proper filtering software to divorce the tables, figures, and printer instructions from the paragraphs, and software to monitor paragraph size and sentence length, many hours of instruction can be developed in only minutes of computer-computation time; it is difficult to estimate the man-years this would require for a traditional classroom approach to comprehension instruction. With the variety and quantity of occupational specialties in the military, this automation can lead to vast savings in course development time.

## SUMMARY

By taking advantage of a process-data relationship between an instructional approach and instructional materials, an initial step in automating the courseware-development process was investigated. The evaluative data indicated that the CAI system was effective in improving reading comprehension skills of Navy recruits. Although the system used a minicomputer for course development and delivery, a microcomputer could certainly accommodate the computational requirements.

Panel: Military Compensation: A New Look at an Old Challenge.

Hale, Linda Pappas, Hay Associates, Washington, D.C. (Chair); CPT Thomas Hale, USN, Office of Chief of Naval Operations, Washington, D.C.; Peter Oglobin, Office of the Secretary of Defense/MPP, Washington, D.C.

AD P001413

The military compensation panel addressed the topic: "Military Compensation: A New Look at an Old Challenge." The panel discussion examined three essential areas - military compensation philosophy, the pecuniary rewards, and nonpecuniary rewards - and developed recommendations for adjustments in the compensation system. It addressed making proper military manpower and force management decisions based on a thorough understanding of military compensation philosophy. Additionally, an understanding of pecuniary rewards, how they can be adjusted, and their relationship to manpower supply is basic. Finally, new and creative uses of nonpecuniary (noncash) compensation were considered for a compensation system competitive with the private sector.

A thorough overview of the military compensation system's cash rewards was presented. Their definition and raison d'être provided the audience with a sound appreciation of the complexities of the military compensation system and highlighted two of the system's short comings: lack of a theory of compensation and little or ineffective use of noncash rewards. A succinct overview of major compensation changes was presented in conjunction with the relevant laws. As the discussion evolved, the way to practically address and develop a theory of military compensation surfaced.

The panel presentation on the theory and philosophy of military compensation underscored the lack of a unifying, underlying theory in military compensation today.

Questions from the audience centered on the perceived value of benefits (e.g., retirement) and the use of nonpecuniary rewards (e.g., flextime, choice of duty assignment) with some interest in the theory and philosophy of military compensation.

The discussion on the employee's perceived value of benefits and nonpecuniary rewards addressed the employer's return on investment in these areas. That is, for every dollar the employer invests in an item what is the employee's perceived value of this item. Clearly, such an examination must be part of sound management practice in the Federal government as well as the private sector.

A methodology to measure the employee's perceived value of both cash and noncash benefits was provided to the audience. Such an approach enables the employer to quantitatively measure the return on benefits and/or ascertain the perceived monetary value of nonpecuniary factors. For example, how much is it worth to an employee to work a 40-hour week but still have every Friday afternoon off? For some individuals, this work-day arrangement may be worth $1,000 a year, and hence a lower salary could be paid.

Obviously, numerous monetary and nonpecuniary factor trade-offs exist. Further information may be obtained by contacting the panel chairperson: Linda Pappas Hale at (202) 833-9250 or (703) 323-1677, or through the Army Research Institute MTA co-chairmen.

# MILITARY COMPENSATION: AN OVERVIEW OF THE CASH PAY STRUCTURE, CHALLENGES FOR THE FUTURE

By Captain T.M. HALE, USN
Head, Compensation Policy Branch,
Office of the Chief of Naval Opeations
30 October 1981

Summary. The gradual deterioriation of military pay in the 1970's because of the pay caps and reallocations of pay has been sharply reversed by major military pay acts in 1980 and 1981. A new pay standard is needed to measure military and civilian pay comparability and a tamper proof adjustment mechanism is necessary to ensure military pay does not again fall behind wages of comparable skills in the private sector. Special and incentive pays require updating and linkage to basic pay growth to maintain their value. Other compensation improvements are needed to maintain a competitive position with opportunities in industry.

The cash portions of active duty military pay can be broken down into the general categories of basic pay, housing and subsistence allowances, and special and incentive pays. Basic pay is the only cash element that is received by all members of the military service. Some members receive many of the pays, none receive all.

## HISTORICAL EVALUATION OF PAY

Before the Joint Service Pay Act of 1922 was enacted, each service provided pay for its members by means of separate pay legislation. Although the pay was roughly equivalent for the Army and the Navy, there were significant differences. For many years Navy pay was differentiated between those serving on sea duty and those on shore duty. Officers on sea duty were viewed as serving normal type duty and received normal pay. Officers assigned ashore received less pay because they were viewed as serving something other than normal duty. At times, Marines were paid on Army scales and at times on Navy scales.

The Act of 1922 changed all that. It established uniformed base pay rates for offices of all services based on a combination of rank and length of service. Rental (quarters) and subsistence allowances reflected an equivalency of need for officers by differing by number of dependents in the case of the subsistence allowance and by rank and dependency status in the case of the rental allowance. Enlisted members were provided the cash allowances for quarters and subsistence when these items were not provided in kind.

Special and Incentive pays evolved over time to reflect the need for special compensation to attract and retain the force required to man and operate the military. Enlistment and reenlistement bonuses date back to the eighteenth century, and reflect the effectiveness of money in the decision process of joining or remaining in the military. Likewise, flight pay dates back to 1913 when extra compensation was needed as an attraction device to encourage members to undertake the very dangerous practice of flying the early flying machines. Diving pay dates back to the 1880's, when it was determined that the Navy needed its own group of experts in this occupational field. Parachute pay was instituted in 1941 when it became necessary to attract large numbers of soldiers to become proficient in this new warfare tactic.

The Career Compensation Act of 1949 provided the foundation for the pay and allowances system as it currently exists. The term "basic pay" replaced "base pay" and was based on rank and years of service. The Basic Allowance for Quarters (BAQ) and Basic Allowance for Subsistence (BAS) were established to replace the rental and subsistences allowances.

Both the 1922 and 1949 Acts followed major reduction in force levels after the world wars and were designed to provide a stable and meaningful compensation system for an older, more career oriented force that would experience slower promotions during peacetime service.

Military pay was adjusted periodically during the 1950's and 1960's but generally lagged wage growth in the private sector. It was not until the enactment of Public Law 90-207 (the Rivers Bill) in 1967 that military pay increases were linked with General Schedule pay increases for federal civil servants.

This law also established the concept of Regular Military Compensation (RMC) as the rough military equivalent of civilian pay. RMC consists of basic pay, BAS, BAQ, and tax advantage associated with the tax free allowances. The three cash elements of RMC were the basis for adjusting basic pay from 1967 to 1973 and RMC itself is often used as a means of comparing military and private sector wage growth. The definition was significantly broadened in the Nunn-Warner Amendment to the National Emergency Act of 1980 (hereafter called the Nunn-Warner Amendment) by the inclusion of the Variable Housing Allowance (VHA) and Overseas Station Housing Allowance (SHA) in the definition.

## CURRENT CASH PAY STRUCTURE —

The elements that currently comprise the military cash pay structure are shown on Table 1. The length of the list is much more impressive than its effect. Most of the force does not receive more than basic pay and one or two of the special and incentive pays or allowances. The law prohibits the payment of more than one hazardous duty pay to any individual with the single exception of some special forces who may receive two hazardous duty pays. Special pays are also quite limited in application. For example, one of the costlier special pays, career sea pay, accrues to only 19 per cent of the Navy force and slightly over 5 per cent of the entire military. Submarine duty incentive pay is paid to only about 5 per cent of the Navy force, while parachute duty pay accrues to less than 2 per cent of the total military force. All of the duty-related special and incentive pays require performance in an unusually hazardous or arduous duty, either over the course of a career or in a special duty assignment in order to acquire entitlement to the pay. Because of the relatively small numbers that receive most special and incentive pays and the relative insignificance of most of the pays compared to basic pay, it has been difficult to justify increases in economic terms (that is, the effect of pay increases on accession and retention). As a result, many of these pays have been allowed to stagnate over time and have lost much of their value to the member. Table 2 shows when some of the special and incentive pays were last changed and that amount that would have to be paid today if the pay had been adjusted for inflation.

## BASIC PAY AND ALLOWANCES

The effect of the Rivers Amendment (1967) was to link military pay indirectly with wage growth in the private sector through the federal civil service pay link with the Professional, Administrative, Technical and Clerical (PATC) wage survey. Using PATC wage survey data, general schedule wages were to be annually adjusted to match wage levels in the private sector. Military pay was then to be adjusted an equivalent amount. Unfortunately, it hasn't worked as intended. Presidential pay caps in 1975, 1978 and 1979 and reallocation of pay increases from basic pay into the quarters allowance (which many members do not receive) in 1976 and 1977 had the effect of reducing military pay when compared to wage growth in the private sector. Table 3 shows the growth of military pay (expressed as RMC) from March 1971 through FY 1980 compared to illustrative standards since the advent of the all volunteer force. The relative loss of military pay compared to the other standards help explain why retention in the career force reached new lows by mid-1979 and the service began to experience severe difficulties in meeting recruiting goals.

## PAY IMPROVEMENTS

In an attempt to remedy this situation, the services, working with OSD and the Congress were able to influence the enactment of three significant pay bills in 1980. The Nunn-Warner Amendment, the 1981 DOD Authorization Act, and the Military Pay and Benefits Act of 1980 together accounted for more than twenty improvements in military pay, benefits and reimbursements. Most significantly, the laws provided for an 11.7 per cent across the board pay increase and increases in flight, sea, and submarine pay. The laws also provided much needed improvments in reimbursements for government directed travel. At this writing, there is another significant pay bill about to be enacted which, by itself, will provide the most sweeping changes in military compensation since the 1949 Career Compensation Act. In addition to an average 14.3 percent increase in basic pays, a new temporary lodging allowance will be established to help defray the cost of permanent change of station moves. Also, new travel reimbursements will be established for emergency leave for members at overseas locations and there are several new provisions which will expand entitlements to store household effects. The bill also provides for increases in flight pay, diving pay, and extends hazardous pays to additional categories of duty.

The FY 1982 pay raise will restore military pay to roughly the relative levels that existed at the beginning of the AVF. The other compensation improvements will serve to alleviate many of the finanical hardships and sacrifices that are incurred as a direct result of military service.

## CHALLENGES FOR THE FUTURE

Even the most optimistic of the economists on the national scene project inflation at an annual rate of 6 to 10 per cent for the foreseeable future. At this rate there will be a continuing need to update special and incentive pays lest they became valueless over time. More importantly, the prospect of high inflation requires that the pay standard be developed that will ensure that military pay will not lag wage growth in the private sector. An adequate pay standard would consist of wage surveys of representative skills in the military which is flexible enough to permit automatic annual adjustments in military pay and comprehensive enought to cover all of the pay grades. Clearly, the current indirect link to PATC through the civil service wage structure does not meet this test. PATC represents only about 12 per cent of the skills in the military and does not cover most combat skills, the basic function of the Armed Forces, at all. Also, the current adjustment mechanism is not tamper proof as witnessed by the presidential pay caps and reallocation. A recent study[1] determined that a combination of PATC and area wage surveys (AWS) could provide coverage to about 70 per cent of the enlisted military force and has potential as a new standard upon which to base military pay growth. Along with a new standard, an adjustment mechanism should be devised to ensure that artificial limits are not introduced to satisfy short term political concerns as have been the case in the past.

The objective of a new military pay standard would be to remove pay as a major item of concern to the military member. The demands of military service are such that even minor pay disparities tend to be magnified over time and are quickly added to the list of dissatisfiers that mitigate against continued service. Since the services have to "grow" their own personnel, each careerist lost through unexpected attrition causes severe ripples in the force structure. Not only are the years of experience lost but each lost careerist requires several additional recruits to eventually provide the one placement in the career force.

The challenge then for the future is to execute a compensation strategy that focuses on ways to minimize pay and reimbursement irritates while providing a system of compensation sufficient to attract and retain the quality and quantity of the force needed to carry out the defense mission.

1. Rader, Norvin E., et. al., "Pay Principles and Standards", General Research Corporation Report 1207-01-81-CR prepared for the Department of the Navy, General Research Corp., McLean VA., 1981, P3-1

The elements of this compensation strategy would include:

- Development of a representative pay standard upon which to periodically adjust military pay
- Establishment of a pay adjustment mechanism to ensure that military pay does not fall behind the pay standard that has been developed
- Remove artifical pay limitations, such as the senior officer pay ceiling, from the compensation system
- Update special and incentive pays and establish a linkage between the two and basic pay growth to avoid future losses of pay value
- Provide travel reimbursements sufficient to fully reimburse the member for the cost of government directed travel
- Maintain a competitive benefit program to include fully funded medical and dental care programs for the member and his or her dependents.

This program, in conjunction with a comprehensive non-pecuniary reward system, can be expected to provide the solid underpinning to sustain the defense force of the future.

TABLE 1

MILITARY PAY CASH STRUCTURE

Basic Pay Allowances

- Basic Allowance for Quarters
- Basic Allowance for Subsistence
- Variable Housing Allowance
- Family Separation Allowance
- Overseas Station Allowance
    - Housing Allowance
    - Cost of Living Allowance
- Clothing Monetary Allowance
- Officer Uniform Allowance
- Personal Money Allowance (Pay grade 0-9 and above)

Special Pay

- Diving Pay
- Continuation Pay for Nuclear Qualified Officers
- Nuclear Career Accession Bonus
- Hostile Fire Pay
- Nuclear Career Annual Incentive Bonus
- Career Sea Pay
- Certain Places Pay
- Special Pay for Medical, Optometry, Dental and Veterinary Officers

Special Pay (cont.)

- Special Continuation Pay for Medical and Dental Officers
- Variable Incentive Pay for Medical Officers
- Responsibility Pay (Pay grades 0-4 to 0-6 only)
- Proficiency Pay
- Selective Reenlistment Bonus
- Enlistement Bonus
- Overseas Extention Pay

Incentive Pay

- Aviation Career Incentive Pay
- Hazardous Duty Incentive Pay
- Submarine Duty Incentive Pay
- Parachute Duty Pay
- Flight Deck Duty Pay
- Demolition Duty Pay
- Experimental Stress Duty Pay
- Leprosarium Duty Pay

TABLE 2

Special and Incentive Pay Evaluation
(Selected Examples)

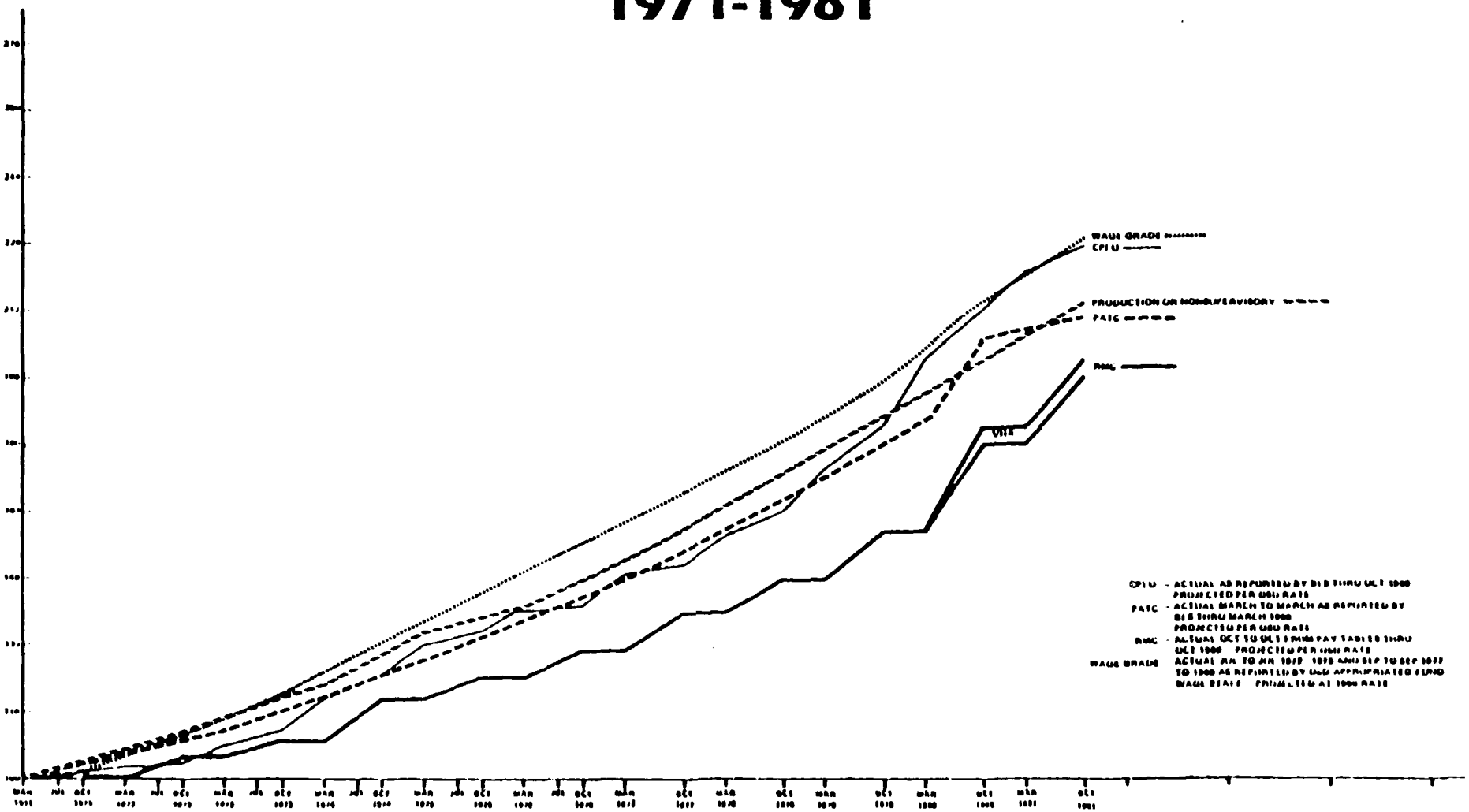
| Pay | Current Value (monthly) | Year Last changed | Value if Adjusted for Inflation (Note 1) |
|---|---|---|---|
| Hostile Fire Pay | $65 | 1965 | $192 |
| Certain Places Pay | $8-22.50 | 1949 | $31-88 |
| Responsibility Pay | $50-150 | 1958 | $161-340 |
| Proficiency Pay | up to $150 | 1958 | up to $340 |
| Parachute Duty Pay | $55-110 | 1955 | $191-383 |
| Flight Deck Pay | $55-110 | 1965 | $163-325 |
| Demolition Pay | $55-110 | 1955 | $191-383 |
| Experimental Stress Pay | $55-110 | 1957 | $182-364 |
| Family Separation Allowance | $30 | 1963 | $91 |

Note 1. Assumes inflation rate of 10% for FY 1981.

# RMC
# COMPARED TO SELECTED EXTERNAL STANDARDS
# 1971-1981

Table 3

1478




PREPARED BY GENERAL RESEARCH CORPORATION, MCLEAN, VA. FOR THE DEPARTMENT OF THE NAVY, 1981

AD P001414

# THE NEED FOR A THEORY OF MILITARY COMPENSATION

Peter K. Ogloblin
Assistant Director, Compensation
Office of the Assistant Secretary of Defense (MRA&L)
Department of Defense

Summary: The author argues that the lack of a coherent articulated theory of military compensation has been a major factor in the failure of past military compensation studies, reviews, and legislative initiatives. He urges that this intellectual void be filled as a major priority by the 5th Quadrennial Review of Military Compensation, and suggests criteria that should be followed in developing such a theory. The views are the author's own and do not necessarily represent the position of the Department of Defense.

# THE NEED FOR A THEORY OF MILITARY COMPENSATION

by Peter K. Ogloblin

## I. Introduction and Background

For the fifth time since 1965, the Department of Defense is preparing to conduct a Quadrennial Review of Military Compensation. What is such a review supposed to achieve? The law enacted in 1965 is quite explicit on this point. It states:

> "Whenever the President considers it appropriate, but in no event later than January 1, 1967, and not less than once each four years thereafter, he shall direct a complete review of the principles and concepts of the compensation system for members of the uniformed services." (Italics added.)

Principles and concepts of the military compensation system are thus what is required to be reviewed. But in what sense, and what is the President supposed to do with the results, and when? The law continues:

> "Upon completion of such review he shall submit a detailed report to the Congress summarizing the results of such review together with any recommendations he may have proposing changes in the statutory salary system and other elements of the compensation system provided members of the uniformed services." (Italics added.)

Thus, the law requires a review of the principles and concepts of the "statutory salary system and other elements of the compensation system" of the armed forces. What does this mean? As anyone marginally familiar with American military compensation knows, there has never been a "statutory salary system" for enlisted personnel at all, and an ephemeral quasi-salary system existed for officers only in the years 1870-1922. Thus, there exists no "statutory salary system" whose "principles and concepts" must be reviewed. So we must ask ourselves the question: What did the drafters of this law intend? Were they, perhaps, hinting that a "statutory salary system" be established, whose "principles and concepts" could later be reviewed?

Legislative history provides no evidence to support this contention. We know who caused the law to be drafted. It was the late L. Mendel Rivers of South Carolina, then chairman of the House Armed Services Committee. We also know his intent. What he wanted was to ensure that military pay levels kept up with the pay levels of the rest

of American society; for this we wanted annual reports on adequacy of pay, contained in another section of that law. He wanted as well a periodic review of the whole structure to see whether it still hung together in a coherent fashion, that no element or group of elements became unduly distorted, and that the system supported military operations as they were being changed by the major technological changes in weapons and support systems at the time. These changes were major: ballistic missiles, nuclear-powered submarines, space reconnaissance satellites, the electronics revolution, and command and control systems were all having a direct and disturbing effect on the armed forces of the time. It was the period of the Cold War and the early American deployment to Viet Nam. Rivers wanted to ensure that the military compensation system kept abreast of these changes and supported adequately the military personnel who would operate in this technologically changed warfare.

But that is not what his drafters wrote into the law. Their specific language, which remains when historical memory and the conditions of the time fade in our minds, has left a legacy of some confusion of what was intended. In consequence, some still believe that a military salary system was the desideratum; others believe that what was wanted was a grand review of levels of compensation, with the option of changing selected parts within the basic framework of the existing system. Nor does the law require the President to submit anything other than a report to the Congress at some later unspecified date. So what remains of the "principles and concepts" that were to be reviewed? The short answer is that they offer a challenge to compensation specialists to rethink some of their basic assumptions. Some of these assumptions are unspoken, and that is in itself a problem.

In fact, no quadrennial review, nor the various Presidential or Congressional Commissions that have studied military compensation, has ever seriously or systematically reviewed the "principles and concepts" of any kind of military compensation system. What nearly all of them have attempted, with the one exception of the Second Quadrennial Review, was to collect and manipulate pay and benefits statistical data and compare these with the private sector or the federal Civil Service, usually under the assumption that what either the private sector or the Civil Service did was what the military compensation system ought to emulate. In other words, they did not attempt to rethink the "principles and concepts" of military compensation, but compared pay with pay, and tried to make military compensation be something that it was not. Needless to say, nearly all of the results of such reviews had no effect on reality other than to perturb the morale of military personnel. The recommendation of none have been enacted, with the sole exception of the Second Quadrennial Review, which was uniformly successful, the first major legislative success by an organized review since the Hook Commission of 1949. There is a lesson in this history that should be learned by the decision-makers of the Fifth Quadrennial Review. Some of it lies in the approach taken by the Second Quadrennial Review.

## II. Compensation Theory: What It is, What Its Effects Are

If one examines the literature on the subject, nowhere will he find an explicit theory, or doctrine, of military compensation. This is strange, because nearly every other aspect of military operations has an explicit doctrine embodied in written form, ususally in a manual of some kind. For example, the Army's famous Field Manual 100-5, "Operations", articulates at the highest levels of abstraction the principles and concepts of ground operations of modern warfare. Similarly, various support activities have similar statements of doctrine or principles. If one wishes to analyze or criticize the principles or concepts of any of these activities, here are ready-made coherent descriptions at hand.

These documents have real value for purposes far more important than analysis, however. Their explicit statement of the principles and concepts of their subject constitutes what is known as "doctrine", and knowledge of such doctrine saves considerable effort in communications, since just the communication of a few key phrases sets in motion an entire chain of other actions that do not have to be explained individually. In addition, such a doctrine sets standards for evaluating events and proposed changes. Doctrine develops over the years into a coherent body of thought that fosters progress. In this it is analogous to the scientific method on the one hand or jurisprudence and systematic theology on the other. In academic areas, such coherence often distinguishes what is called a "discipline". But, unfortunately, this is precisely what American military compensation lacks. The consequences of such a deficiency have been painfully evident in recent decades.

What have been the results of this failure of thought? Here, in my opinion, are the most important.

o It has prevented the development of military compensation standards. Lack of standards, in turn, has contributed to the arbitrary treatment of individual compensation elements, leading to a severe disparity in compensation levels between military and civilian sectors, with clearly negative effects on military manning.

o It has contributed in other ways to manning deficiencies, because neither the Services nor military members can plan ahead with much assurance of certainty. Such an inability to plan with assurance is bad for retention of manpower in particular, but it affects attraction as well.

o The lack of a commonly accepted compensation theory or doctrine helps produce instability over the years in the military compensation system, when all, or nearly all, acknowledge that what is most needed is stability.

o Since the military compensation system is supposed to support the manpower structure, which in turn is a support system for military operations, the lack of a coherent body of thought with built-in standards can allow situations to develop wherein the individual, atomistic changes perpetrated not only do not support the manpower system but are often directly contradictory to its objectives. If this should sound like a bureaucratic complaint, consider what happens when the manpower system itself fails to support combat operations, as periodically it did in the Viet Nam War.

o The lack of a commonly accepted theory or doctrine promotes the treatment of each compensation item separately and on different criteria for change, usually those employed being sensitive to the ephemeral political or economic issues of the moment. This in turn leads to the politicization of a vital support system of our national security. Not only is this a malum in se, as the lawyers put it, but it contributes to further mischief. I am referring specifically to making military compensation a happy hunting-ground for economists, statisticians, actuaries, comptrollers, political scientists, psychologists, sociologists, and others. These are people who are ususally well-versed in their own academic discipline but rarely have studied military operations, let alone actually participated in or experienced them. Their mental framework is that of their own discipline, whose main orientation is to other, non-military ends and to whom military matters are but a minor sub-issue of their broader concerns. Since American senior military personnel have rarely been articulate enough to counterbalance these forays--those of the economists have been the most spectacular--they have been inordinately successful, and this has created confusion about what are, or what should be, our military compensation objectives. The harm eventually done to military operations and confidence has been significant. Let me cite the entire salary/RMC issue as one example. Its effects on the allowance structure, BAS and BAQ in particular, have been invariably harmful, until today the genuine and legal purpose of these allowances has been so dissipated that it needed complex correctives such as the new VHA (Variable Housing Allowance). And there have been other undesirable side-effects.

o The lack of a theory or doctrine of military compensation, and the undue influence of people imbued with orientations of non-military disciplines, such as economics most notably, gives a clue to why most quadrennial reviews have been unsuccessful. I will advance the thesis that the basic reason is that they have all asked the wrong questions, and even the most thorough and ingenious answer to the wrong question is going to be irrelevant and unpersuasive to the ultimate policy-makers: the President and the Congress.

## III. Present-Day Needs

Obviously the most pressing need today is to remedy the deficiencies of thought of the past several decades. The consequences of this intellectual failure have just been noted, for when a confused and contradictory babble of voices emanates from the Pentagon, higher decision-makers discount all of them and make their decisions on other bases: OMB from the point of view of cost-containment, Congress from that of political pressures. And this has recurred ad nauseam in recent years. Such history should give a strong message to the planners of the 5th Quadrennial Review of Military Compensation. The most elementary and basic requirement for that group is to develop, articulate, and get agreement on a theory of military compensation--of the very "principles and concepts" that the law requires to be reviewed, albeit under the mistaken notion that such a coherent body of "principles and concepts" exists. It will then be possible to review, criticize, and analyze them, but not until then. One solid, articulate formulation of this kind will probably be more useful than all the random tinkerings with individual compensation components or statistical comparisons with the private sector or the Civil Service put together. After all, such tinkerings and statistical comparisons have been tried repeatedly in the past, and repeatedly failed. What has the President and the Department of Defense to lose by trying another approach? A repeat of the past is highly likely to result once more in the effects of the past--failure in terms of legislative enactment, and the product of thousands of man-hours of dedicated, and often brilliant, analysis neatly filed on library shelves for the curiosity of future historians of military compensation.

The Second Quadrennial Review of Military Compensation provides the one shining exception in the otherwise dismal record of QRMCs. What did it do that the others did not? I suppose I could note, first of all, that it was small and confined itself to four manageable special pay items that could meaningfully be attacked by the staff and be understood and digested by the Congress, instead of trying to recreate the entire military compensation universe, but that, I believe, was not the truly distinguishing feature. The truly distinguishing feature is to be found, in my opinion, in its approach and staffing. Let me illustrate with the study of flight pay, which led to the present-day Aviation Career Incentive Pay after Congressional action, since I am personally familiar with the details.

The Flight Pay study started with the precept that flight pay should contribute to manning the aviation forces by being in consonance with present-day air operations. For this reason, the staff selected was not from personnel in the Service compensation or comptroller staffs, but from the operational air wings or groups directly, preferably with the most senior officers in terms of actual command in air combat hours available. Moreover, all such officers had to be qualified instructor pilots. The first iteration consisted of an Army

helicopter pilot with the most command combat time of any Army pilot in the Viet Nam War, a Navy carrier fighter pilot who not only had command combat time but was director of carrier air operations as well, and a veteran SAC bomber commander with more air command time than any other SAC pilot, and who was in SAC Headquarters at the time. In the second iteration, this group was replaced by an Army helicopter/fixed wing pilot, a Navy reconnaissance pilot, and an Air Force fighter pilot fresh out of Viet Nam, and scheduled to return there after the study was completed

The Services were not pleased to release such officers for a compensation study, but relented when the need was explained. Only one member could be considered a compensation staff specialist, and he was responsible for all compensation history, law, economic analysis, costs and statistics. Needless to say, none of these people had a preconceived notion of what flight pay should be, but all had preliminary ideas. All proved to be wrong, and what emerged was superior to what anyone had imagined at the start of the effort. This group made the necessary comparisons with what the "competition" (airlines, general aviation, and Civil Service) was paying, but these were not the governing or decisive factors. More importantly considered were the air tactics involved, lead-time in both personnel and equipment, experience of allied air forces, and, perhaps most important of all, what the operational units themselves thought and the problems they currently faced. These could be got with relative ease because the make up of that study's personnel, who could contact individual units that they knew, ask questions, and ask for these units to call us back with their results. All of this paid off handsomely in terms of the assumptions used in a later econometric analysis. That analysis proved uncannily accurate for the next two years' projections, but that accuracy had less to do with the econometric method employed than with the assumptions that underlay it. Those assumptions were heavily conditioned by the nature of air operations of the time, and not by labor economics, a fact that perhaps was not made clear with sufficient force in the report of the group. Congress enacted the structure recommended, but at lower pay rates, in the Aviation Career Incentive Act of 1979 (Public Law 93-294).

Please note the approach used: it was oriented to air operations first, and to the labor market second. Flight pay was considered as one piece of a larger military operation, and not as a part of a text-book problem in labor economics. No other QRMC used this approach before or since. I think it is time to consider this approach seriously again in formulating and articulating a military compensation theory, since military compensation is a support system for the military manpower system, in its turn a support system to military operations.

Thus, if the Fifth QRMC develops a military compensation theory, I think it should observe the following criteria.

(1) It should accept as axiomatic that military compensation should support a larger military structure in terms of that military structure, and not in terms of labor economics, or business practice, or some other civilian practice.

(2) Military compensation, as a system, must operate equally well in peace and war. The duties of military personnel, particularly combat duties that are the common responsibility and duty of all "armed combatants" and that are not found anywhere in civil life, must be given their just emphasis as the very raison d'etre of the military as an institution. Parenthetically, I would object to any military duties being called "jobs", because they are not. A "job" excludes the combat function altogether, and to call a military duty a "job" is to create a false and misleading impression of the concept in the mind of the average reader

(3) The theory underlying basic pay, as opposed to special and incentive pays, must be articulated, and related to the "comparability" and "competitiveness" arguments. Similarly, the pay and allowances structure must be functionally described, so that its differences with the salary system are made explicit, as well as the reasons for those differences. If possible, criteria for standards must be developed as well.

(4) The theory of military allowances as reimbursements must be cogently developed, and standards for allowances established.

(5) Military benefits, particularly in-kind benefits, must be explained, and their theoretical underpinnings developed and articulated.

(6) All such standards should work toward making military compensation adjustments administratively automatic, both in terms of comparability and competitiveness, rather than requiring individual positive legislative action for each adjustment, with the consequent threat of politicization of the process.

(7) Develop enough reasoning and adjustment mechanisms to promote long-term stability in the military compensation structure.

I am happy to conclude on the note that others involved in the military compensation process have also recognized this theoretical void, and we have had some success in bringing it to the attention of higher political levels. Therefore, we hope that this theoretical work will actually be undertaken as a basic and important part of the 5th QRMC.

AD P001415

# NONPECUNIARY REWARDS AND MILITARY COMPENSATION

Military Testing Association

23rd Annual Conference

This paper presents the importance and impact of nonpecuniary factors in compensation systems. Nonpecuniary rewards impact different categories of workers differently. Appropriate compensation has a substantial impact on morale, productivity and job satisfaction. In the light of decreasing budgets and increasing demands for high quality military members, the military compensation system needs to consider increased utilization of nonpecuniary rewards.

Linda Pappas
Hay Associates
1110 Vermont Avenue, N.W.
Washington, D.C. 20005
Telephone (202) 833-9250
(703) 323-1677

October 30, 1981

# NONPECUNIARY REWARDS AND MILITARY COMPENSATION

## INTRODUCTION

Noncash compensation has long been considered an important element of the reward system. Marshall, in his Principles of Economics, recognized the importance of noncash compensation as early as 1920, when he stated:

> The true reward which an occupation offers the labourer has to be calculated by deducting the money value of all its disadvantages from that of all its advantages.

This paper considers selected nonpecuniary factors and their potential impact on the Department of Defense.

## OBJECTIVES

The objectives of this paper are to:

- Identify nonpecuniary factors (NPFs) not typically considered in the military compensation system
- Present research findings on the impact of selected NPFs
- Illustrate how the military may approach identifying and evaluating NPFs to be used in its reward system

## DEFINITIONS

To facilitate communication, the following definitions are used:

- Cash Benefits: These include basic pay, allowances, special and incentive pays received by the military member.
- Noncash Benefits: These are the commonly called "fringe benefits": goods, services, or deferred money received by the service member, but paid for (at least in part) by the government. They include such items as medical insurance, holidays, paid vacations, pension plans, sick leave, and death benefits.
- Nonpecuniary Factors: Factors which are not directly tied to wage and benefit packages but may be considered as part of a job's "compensation package." Some can be physically identified and impose clear-cut costs on the defense budget (e.g., health and safety

of the work place). Other job-related factors are difficult to measure quantitatively and to evaluate and may or may not impose clear-cut costs (e.g., the effects of training on promotion and job security). Still others fall into more subjective categories which involve the measurement of attitudes and perceptions regarding work, a particular organizational setting, and the desirability of certain kinds of work. (Pay Comparability in the 1960s and 1970s, O. J. Harrison, General Research Corporation, June 1979).*

## BACKGROUND

The major research in NPFs has been conducted by psychologists, sociologists, and management on work-related attitudes of the American worker. Much of the research deals with morale, job satisfaction and dissatisfaction and motivation. This paper considers these items as well as NPF's impact on productivity.

Studies related to these areas originated in the early 1900's. The Hawthorne studies in 1924 were designed with a productivity measurement in mind but also had a by-product of identifying factors affecting job satisfaction. In 1935, Happock conducted a community survey of working adults on job satisfaction. In the 1950's, two widely referenced studies were conducted; one by the National Industrial Conference Board (NICB) and one by General Motors Corporation (GMC). The former study required respondents to identify important job-related morale factors. The latter study provided an opportunity for employees to project their attitudes with respect to their jobs

---

* NOTE: This paper draws heavily from research conducted by my deputy department director when I was employed by General Research Corporation.

and the work environment. These surveys demonstrated the importance of noncash factors in motivation. In fact a significant finding in industrial psychology research is that nonfinancial incentives play a tremendous role in satisfying workers' needs and wants. "There seems no question that, particularly as wages rise above the subsistence level and standards of living are raised, other needs and wants, satisfied by plant conditions other than pay, attain higher levels of prepotency.[1]

A literature search showing the development of NPF's from 1940 to the present is presented in Table 1. One striking feature of the table is the substantial growth of such factors with the passage of time. Another indication of the growing importance of these factors is found in the 1980 American Compensation Association Regional Conference Proceedings. Approximately 50% of the papers address benefits in some manner with considerable emphasis on NPFs.

## STATE-OF-THE-ART

We know that different NPF's have varying appeal to different categories of employees. This is clearly illustrated in a Conference Board study, Factors Affecting Employee Morale, S. Avery Raube. The purpose of this study was to determine directly from the employees what they thought was important relative to what management and labor leaders thought employees felt was important. Five manufacturing companies and one printing publishing company participated on the survey. As might be expected responses varied by employee category:

- Clerical/nonclerical
- Male/female
- Length of employment
- Age

For example clerical workers place opportunity for advancement first in the five most important job morale factors while nonclerical workers placed job security first and opportunity for advancement fourth. Such differences in prioritizing job morale factors were exhibited by other demographic groups. Even though this study was conducted in the mid-1940's, research today continues in identifying relative NPF value by employee category.

Another study, Productivity Improvement Through Incentive Management (1979), B. I. Spector and J. J. Hayes, supports not only the preceding findings but also varying effectiveness of

---

[1]Morris S. Viteles, Motivation and Morale in Industry, W.W. Norton and Co., Inc., New York, 1953, p. 385.

incentives by organizational type (e.g., military, educational, manufacturing), task type (e.g., production, clerical, training), and worker type (e.g., blue collar, white collar, uniformed).

The study was conducted at Red River Army Depot and focused on three main objectives:

- To articulate productivity improvements made in the depot system,
- To determine the effects of factors that decrease productivity, and
- To develop techniques to improve productivity and Army readiness in the future.

The study has some interesting findings even though the sample size is relatively small (N=54).

Selected findings include:

- Blue collar workers are motivated to significantly higher performance levels by recognition, privileges, and by disciplinary action.
- In training tasks, variable bonuses are most effective in assuring high quality.
- When tasks are inherently interesting, variable bonuses yield significantly more effective qualitative results. With boring tasks, however, workers are stimulated by cash-noncash mixes and by recognition or privileges.[2]

The study conclusion states "the results of this research effort confirm that incentive management techniques constitute a sound and effective methodology for improving productivity (both qualitative and quantitative)* through workforce motivation. However, different incentive strategies should be chosen for optimal effectiveness depending on the varied contingencies of the particular work situation."[2]

---

[2]Bertram J. Spector and John J. Hayes, Productivity Improvement Through Incentive Management, Cybernetics Technology Office, Defense Advanced Research Projects Agency, September 1979.

Military Compensation

The military currently has a number of noncash benefits such as:

Recreation facilities
Commissary stores
Military exchanges
Mortgage insurance premium
Annual leave; accrued leave
Medical care, members, dependents & retired members
Government contribution to Social Security
Tuition Aid
Retired Pay, nondisability and death gratuity, disability dependency and indemnity compensation
Severance pay, nondisability and disability
Survivor benefit plan
Servicemen's group life insurance
Professional and educational training
Voluntary education and training
Home owners assistance program
Veterans benefits

These benefits are in addition to over 20 different special and incentive pays (e.g., diving duty pay), nine allowances (e.g., family separation allowance) and several reserve component pays (e.g., drill pay).

*Insert mine.

The questions that are raised based on cash and noncash compensation research to data are:

- Are these military rewards being expended most effectively by typical military members (e.g., rank, rate, MOS, length of service)?
- Could alternative rewards, or reward combinations, be identified that would be more effective?
- How does the existing reward system impact productivity, readiness, manning?

These issues are of substantial importance since the rewards perceived as important by one group (e.g., economist, enlisted member, officer, force manager) may be different from those perceived as important by another group.

Relative NPF Importance

The perceived relative importance of various NPFs is clearly illustrated in the previously cited study by the National Industrial Conference Board. The research findings differed substantially by respondent category when defining the relative importance of nonpecuniary job factors. This is illustrated in the following Table.

Both labor leaders' and management's predictions of important job factors to employees differed from those actually identified by the employees.

TABLE 2
IMPORTANT JOB FACTORS
(rank ordered)

| Employee Selection | Executives' Preditions of Employee Selections | Labor Leaders' Predictions of Employee Selections |
|---|---|---|
| job security | compensation | compensation |
| opportunity for advancement | job security | job security |
| compensation | vacation & holiday | total hours worked per day, per week |
| financial benefits (i.e., insurance) | opportunity for advancement | labor unions |
| informing you of your job status | physical working conditions | |

In addition to rank ordering the relative value of NPFs, several attempts have been made to assign a monetary value to NPFs.

Monetary Quantification of NPF's

The monetary quantification of NPFs has considerable appeal to the budget watchdogs. A satisfactory system of monetizing could not only permit effective rank ordering of NPFs, but also result in fiscal savings as well as increased morale, productivity and readiness. Several studies have been conducted in this area. Selected ones are summarized briefly below:

- Quantifying Nonpecuniary Returns, L. F. Dunn, 1977; interviewed workers were asked to give quantitative evaluations of certain NPF returns on the dimensions of money and time. Textile workers were asked how much they would pay for a benefit and how much longer they would work without extra pay for the benefit.

- Dollarizing Attitudes, Meyers and Flowers, 1974; a formula was applied to the results of an employee attitude survey to convert attitude scores into financial returns on payroll investment expressed as gain, break even, or loss. This attempted to add human variables to fiscal variables such as return on investment and earnings per share.

- Pay and Benefit Preference, S. M. Nealey; a game board method was used to determine how an employee would apportion a fixed dollar amount between pay and benefits.

- Monetizing Nonpecuniary Factors for Consideration in the Pay Comparability Process, O. J. Harrison, L. D. Pappas, et al., 1980; data from company records, observation in the work place, and interviews were run through a multivariate regression and dollar values were identified for selected NPFs.

While no one study has established a precise way to monetize NPFs, initial steps have been taken and hold promise for more precise evaluation in the future. This research area is worthy of pursuit in the face of tighter budgets and fiscal restraint.

## A. POTENTIAL APPROACH

To determine the importance of NPFs to the military member, further research should be performed. One potential approach is presented here:

- Conduct a survey. The survey[3] should be designed to elicit views of the employees and, where appropriate, their families in such areas as:

  - The comparative importance of various NPFs
  - The NPFs most desired
  - Willingness to contribute toward the additional cost involved in new or improved benefits

---

[3]W.L. White and J.W. Becker, Personnel, "Increasing the Motivation Impact of Employee Benefits, June-February, 1980. (Adapted from this work.)

- Clarity of the benefit program, both fringes and NPFs

- Administrative efficiency

- Preferences concerning the relative proportion of direct compensation, fringe benefits, and NPFs

The survey should be repeated periodically to assess the changing views and needs of the service members and the effectiveness of any previous changes in benefits or in administering or communicating the program. This information would help force managers to reestablish benefit objectives or exchange existing benefits for those more in line with defined objectives.

o Use of the survey results. Analysis of attitudinal survey results can yield valuable information for the compensation planner -- for example:

- A numerical ranking of the relative importance of each benefit to service personnel as a group and a similar ranking for individual communities within the military.

- Identification of the benefits that are most in need of improvement, in order of priority.

- Identification of the new benefits most valued by the member, in order of merit.

- Exposure of deficiencies in benefits design and in administration.

- Cost-effective allocation of contributions through the determination of highly valued low-cost benefits and lesser valued high-cost benefits.

- Knowledge of whether service personnel would prefer a change in the relative proportions of direct compensation and benefits and, if so, in what direction and to what extent.

This approach is readily doable and could yield useful compensation management information that could result in reduced monetary costs and increased force management efficiency and improved morale.

PANEL: Women in the Military

Zinzer, Nora Scott, Industrial College of the Armed Forces, Washington, DC (Chair); Forestell, Diane G., LT, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario; Segal, David R., University of Maryland, College Park, Maryland; Segal, Mady Wechsler, University of Maryland, College Park, Maryland & Walter Reed Army Institute of Research, Washington, DC; Simpson, Suzanne P., CPT, Colts Neck, New Jersey; and Edwards, Henry, University of Ottawa, Ottawa, Ontario, Canada.

This panel presents and discusses three papers dealing with the issue of women in the military. Lt. D. G. Forestell, Canadian Forces Personnel Applied Research Unit, discusses women's participation in the Canadian Military during World War II, focussing on ambivalence and role conflict in those years. Another Canadian paper by Captain Suzanne P. Simpson evaluates sex differences in the performance appraisals given to noncommissioned men and women in the Canadian forces in relation to supervisors' Attitude Towards Women Scale. Socio-psychological variables were used to explain differences in results. A far-reaching paper by Drs. Mady and David Segal looks at the increasing participation of women in the U.S. Armed Forces during the past ten years within the context of three factors--the changing nature of the military institution, demographic changes, and attitudinal changes towards women's role within U.S. society as a whole. All three papers provide new dimensions to the study of women's roles in the military and society.

AD P001416

# THE VICTORIAN LEGACY
# A SOCIAL HISTORICAL ANALYSIS OF ATTITUDES TOWARD WOMEN IN THE CANADIAN FORCES*

Lieutenant Diane G. Forestell

Canadian Forces Personnel Applied Research Unit
Toronto, Canada

* The views and opinions expressed in this paper are those of the author and not necessarily those of the Department of National Defence

# THE VICTORIAN LEGACY
# A SOCIAL HISTORICAL ANALYSIS OF ATTITUDES TOWARD WOMEN IN THE CANADIAN FORCES

Lieutenant Diane G. Forestell
Canadian Forces Personnel Applied Research Unit
Toronto, Ontario

The Canadian Forces are currently conducting a five-year evaluation to assess the impact on unit operational effectiveness of integrating women into near-combat environments. One issue that has emerged during the first year of the evaluation is the circulation of rumours regarding the servicewomen who have volunteered to participate in the five-year programme.

Rumours regarding illegitimate pregnancy, sexual deviance and misbehaviour continue to circulate about servicewomen serving in near-combat roles at Canadian Forces Europe, CFS Alert and aboard the diving tender, HMCS Cormorant. The US military, embarked on a similar course of integrating women into non-traditional military roles, has dealt with similar rumours regarding sexual harrassment, sexual deviance, etc., among female personnel - e.g., last year's much publicized investigation of lesbianism aboard USS Norton Sound.[1]

While a disturbing phenomenon, historical evidence indicates that it is neither new nor unique to the military. Historically, women who have sought admission to male-dominated professions/institutions, be it the military, the medical or legal professions, or professional sports, have been the objects of rumours which have attacked their femininity and sexual behaviour.

This paper presents a research proposal for social-historical analysis of women's military participation in Canada during WWII. Women serving in the corps of the armed forces during this period were the objects of a "whispering campaign" which focussed on alleged promiscuity among servicewomen. The research will attempt to examine the "factual basis" for the widespread perception that women in uniform were "loose women". The reasons advanced for the apparent fall of servicewomen into promiscuity will be analyzed in terms of the prevailing cultural ideology which was based on middle-class Victorian definitions of women's "proper sphere" and perceptions of "femininity". It is suggested that explanation based on these notions does not adequately explain societal opposition to women's military participation. Rather, it is suggested that prevailing middle-class Victorian values and attitudes served to reinforce the status quo of the dominant political ideology. The relevance of such research for our understanding of current attempts to integrate women into non-traditional military roles will be addressed.

Common to historical accounts of women's admission to the medical and legal professions and the Canadian armed forces during WWII, are the predominance of middle-class Victorian definitions of the "proper sphere" of women and perceptions of "femininity". The Victorian definition of the "proper sphere" of women was based on "the fact that in the economy of nature or rather in the design of God, woman is the complement of man".[2] The "feminine" nature of woman, characterized by purity, delicacy and gentleness,

etc. could only find its full expression "at home amid the quiet and the peace, and the purity and the love of which she is alike the source and the receipient".[3] Thus, it was only "at home and in its co-relative situations that man finds woman to be his complement".[4] As a 19th century treatise on the subject of women's military suffrage asserts: "In the (military) camp she must either be the subordinate or the supervisor or the equal of man: she cannot be his complement".[5] Within this context, women's admission to institutes of higher education and military service was seen as antithetical to women's "feminine" nature which found its full expression only within their circumscribed sphere of activity, i.e. their role as wife, mother, and homemaker. Women's admission to these male bastions would result in a loss of their femininity with its attendant emphasis on "purity" and, consequently, their (sexual) respectability.

Women's military participation was seen as an attempt to "unsex women" - e.g., "Some Roman ladies, in the corrupt days of the empire, having exhausted ordinary means of excitement, were seized with the lust of unsexing themselves and trained as gladiators ... nothing resulted but depravation"[6]. Female soldiers were thus to be seen as "an outrageous anomoly in the body politic".[7] The term "unsex", used at that time, referred to a loss of femininity rather than to sexuality, i.e., to the loss of the proper dress, decorum and etiquette associated with middle-class Victorian perceptions of "femininity". Thus, by definition, women who attempted to gain admittance to these male bastions were "unfeminine", i.e. "unladylike" and amid the prevailing morality of a sexual double standard, promiscuous and immoral.

In her article "Ladies or Loose Women: The Canadian Women's Army Corps in World War II", Dr. Ruth Pierson argues that the ambivalence of Canadian society to women's admission to the armed forces was grounded on Victorian perceptions of "feminine respectability".[8] Paramount among societal concerns regarding women in uniform were fears of "loss of femininity" and (sexual) respectability. Both a general public opinion survey and a CWAC survey conducted in 1943 confirmed that suspicion existed in the public mind that joining the forces would embark a young woman on a life of promiscuity.[9] Service in the armed forces was seen as "an unladylike occupation" in which the young women would lose her "self-respect". The CWAC report explicitly stated that "the general public felt that young women avoided enrolling in the women's services because of the fear of association with women of poor moral standards".[10]

Women who did serve in the corps of the Canadian armed forces were the objects of a "whispering campaign" of "malicious rumours and gossip"[11] designed to discredit them on moral grounds. This "whispering campaign" was a source of concern for recruiters who were concerned with the adverse effect of the rumours on enlistment and morale. While some evidence suggests that there was some factual basis for what recruiting officers dismissed as "baseless gossip", it cannot account for the widespread perceptions among servicemen and the civilian populace that women in uniform were "loose women".

By the spring of 1942, pregnancy in unmarried servicewomen and venereal disease were viewed as a serious problem of medical treatment, welfare and "wastage" of armed forces female personnel.[12] Calculated at 32.1 per thousand unmarried women in the CWAC per year on the basis of figures for 1 Jan 1943 through 30 April 1943, illegitimate pregnancy had risen to 33 per thousand per annum by July of that year. At the end of 1944, it was reported that "the incidence of illegitimate pregnancies has remained consistent at approximately 35 per thousand strength per annum" for the past two years.[13] An examination of Dominion Bureau of Statistics Canada Year Book, however, indicates that illegitimate births ranged from 3.96% to 4.05% between 1941 and 1943, i.e., 39.6 to 40.5 per thousand for the same period.[14] One study of 95 women discharged from the CWAC for illegitimate pregnancy between 1 Jan and 31 May 1943 disclosed that 31.5% had been pregnant prior to enlistment.[15] While it is impossible to make direct comparisons of illegitimate pregnancy rates between servicewomen and civilian females, the available data suggest that the incidence of illegitimate pregnancy was not significantly higher among servicewomen than among the female population as a whole. It should be noted that the steady increase of illegitimate births in the civilian populace over the period 1941-43 was, in some measure, due to the more complete registration of children born out of lawful wedlock. Within the women's services, the high rate of illegitimate pregnancies recorded may have been more a reflection of reporting than a reflection of the "promiscuity" of servicewomen per se.

Similarily, it is impossible to ascertain if the incidence of venereal disease among servicewomen was significantly higher than among civilian females. Attempts to make inferences or draw conclusions on the sketchy data available or to examine servicewomen in isolation may result in distortion and misinterpretation of the extent of the problem among servicewomen.

The fact that "although the incidence of VD among male soldiers was higher than that among members of the CWAC*, and although in one study servicemen comprised 86.3% of the putative fathers named by CWACs discharged for illegitimate pregnancy"[16] armed forces male personnel were not the objects of a similar "whispering campaign" to discredit them on moral grounds. This is not only indicative of the sexual double standard of morality which prevailed at the time, but also reflects military socialization's emphasis on "masculinity". The common link between VD and "paternity" among soldiers is "sexual competence", an essential element of the military socialization in "masculinity".[17]

One explanation offered for the apparent fall of CWAC women into "promiscuity" was to be found in the policy of "mass recruiting" and the consequent "influx of undesirable types" into the CWAC. According to this argument there was a correlation between low level of education, skill and

* The fact that the physical effects of VD infection are more immediately apparent to males than females may account, in part, for the higher incidence of VD reported among male soldiers than among servicewomen.

intelligence, and a high incidence of venereal disease and illegitimate pregnancy.[18] However, the statistical evidence to support this contention could hardly be adduced as incontrovertible.[19] Moreover, there was evidence which pointed away from that conclusion. The "selective recruiting" introduced in February 1944, i.e., "discrimination at the recruitment level between low calibre, poor character women and women [who] have good character, trades experience or trainable qualities"[20] had no effect on reducing the incidence of illegitimate pregnancy or venereal disease.

Other explanations for the high rate of illegitimate pregnancy and venereal disease focussed on the incompatibility of the regimentation and uniformity of the masculine armed forces with "femininity". Such incompatibility led to the lowered morale of servicewomen and consequent liaisons which risked their respectability. As ludicrous as such explanation may seem, it constituted the basis for a policy/promotional campaign which emphasized that women in the services were "different" and that they should "remain women first and soldiers second".[21] In order to provide "an outlet for their feminine characteristics" and consequently raise morale and lower the illegitimate pregnancy and VD rates, it was recommended that the "minor appurtenances women usually surround themselves with" be introduced into servicewomen's barracks and recreation rooms to provide a "more home-like atmosphere".[22]

The reaction of servicemen during WWII to Canadian women donning the uniform was little different from that of male medical students at the turn of the century in both Britain and Canada to women's admission to medical school. Servicemen's salacious jibes at servicewomen, as well as jokes and lewd stories at the expense of the moral character of women in uniform were an expression of their opposition to women's admission to the military. In 1943, a directive from Ottawa advised all commanding officers to inform "their Officers and Men that any word or action on their part that might be found to reflect upon the character of girls in uniform will be punishable".[23]

One reason for the vehemence with which servicemen opposed the admission of women into the military was that it was "another male job on which women have encroached."[24] The opposition of servicemen in the 1940s echos the dominant ideology of the Victorian era which stressed that women were to complement men, not compete with them. The similarities to the contemporary situation are obvious, both within the Canadian and US contexts. The verbal and physical harrassment of women in near-combat environments in today's Canadian Armed Forces, reflects, in part, servicemen's opposition to women "doing a man's job". While servicemen readily concede that women are "capable" of doing the job, employment in these non-traditional roles (the last bastions of male exclusiveness in the military) is seen as antithetical to notions of "femininity". Additionally, the influx of women into the military, both in the historical and contemporary contexts, challenges military socialization's emphasis on "masculinity". The military, historically, has operationalized an ideal of masculinity through instrumental archtypes, i.e., the male archtype, the female archtype. Not surprisingly, the female archtype is based on Victorian perceptions of "femininity" and the distinction between "good" and "bad" women.

The question arises as to whether explanation in terms of prevailing middle-class Victorian perceptions of femininity, women's "proper sphere" and the "masculine mystique" of the military provides an adequate understanding of male reaction both in the historical and contemporary contexts.

It is suggested that a more fundamental ideological issue which focusses on Western democratic societies' conceptions of citizenship is operative here. The importance of the definition of "citizenship" is particularly relevant to both historical and contemporary analyses of women's participation in the military (i.e., women's right to bear arms). The relationship between military institutions and citizenship has been examined by Janowitz[25], Segal[26], etc. As Janowitz notes, "Military service in Western societies emerged as a hallmark of citizenship and citizenship as the hallmark of a political democracy".[27]

An 1872 treatise on the "military objection to female suffrage" stated explicitly: "...it remains true that if the defence of a country is an essential part of a citizen's duty, men alone can be full citizens"[28]. Evidence from the "persons" cases in Great Britain, suggests that the exclusion of women historically from public activities, including the right to vote, to be admitted to the bar, to hold property, etc., was based on the Saxonian proscription that confined councils to those who "bore arms".[29] The exclusion of women from male-exclusive professions and/or institutions was based on legal definitions of "persons" (in Great Britain) and "citizens" (in the US). Implicit in the generic term "persons" and the word "citizens" was the assumption that "persons"/"citizens" were male. Thus, by definition, women were "non-persons"/"non-citizens" and hence, by fiat, not eligible for inclusion in these areas.

Amid prevailing Victorian definitions of the "proper sphere" of women and perceptions of femininity, attempts to exclude women from institutions of higher education and from political participation were based on a view that women were to be shielded "from the harsh vicissitudes of public life". The exclusion of women from public life was to be seen as an "exemption" flowing from respect for the "admirable attributes of her sex, namely her gentleness, affection and domesticity". Thus, the exclusion of women from public office was to be seen as an "exemption founded upon motives of decorum, and to be regarded as a privilege rather than a disability".[30] To admit women into the military during WWII and, in the contrmporary context, to admit then to military combat-support roles, is to recognize full equality of citizenship by extending to women the right to bear arms. Evidence presented by Segal on the relationship between the concept of citizenship and attitudes toward women in conmbat indicates widespread societal disapproval for extending "military suffrage" to women. Segal hypothesizes that contemporary opposition to women in combat, and thereby extending their citizen participation, is a defence of the status quo.[31] It is suggested that the dominant cultural ideology of Western societies based on middle-class Victorian perceptions of women's "proper sphere" and the "masculine mystique" of the military has served to reinforce the status quo which denies full citizenship participation to women.

The proposed research will attempt to answer two fundamental questions. First, was there a "factual basis" for the "whispering campaign" against women serving in the women's corps during WWII? Social historical analyses, based on an examination of the personnel files of servicewomen from 1941-46 will enable us to determine the extent of the problem of illegitimate pregnancy and VD and to ascertain if indeed there was a relationship between level of education, etc. and illegitimate pregnancy and VD rates. If examination of this data indicates that there was little or no factual basis for the imputation of low moral character to women serving in the women's corps (and certainly, evidence exists which suggests that such imputation was unwarranted) it then remains to examine the reporting of the "malicious rumours and gossip". Newspaper articles dealing with the women's services will be analyzed to determine how women's participation in the military was viewed by the media. The analyses of these data will be interpreted within the context of the prevailing cultural and political ideologies.

The second question concerns what implications such research may have for our understanding of current attempts to integrate women into non-traditonal roles in the militaries of Western democratic societies. The prevalence of rumours regarding women in near-combat environments today may be dysfunctional for current attempts to recruit women to meet the manning requirements of the Canadian Armed Forces. This suggests that at a time when the CF, as well as the all-volunteer forces of other Western democratic societies, are concerned with the maximum utilization of manpower, the Victorian legacy may mitigate against the effective utilization of a substantial segment of available resources. Moreover, it suggests that the success of equal opportunity programmes may be not so much a matter of "military suffrage" as a matter of restructuring attitudes and values within the more traditional elements of the military environment to provide for a more functional perception of the role of women in the socio-political process in Western democratic societies.

## Notes

1. The Atlanta Journal, 13 June 1980.

2. Ramsey Cook & W. Mitchensen, eds. The Proper Sphere. Toronto: Oxford University Press, 1976, p. 8.

3. Ibid, p. 20.

4. Ibid, p. 20.

5. Ibid, p. 20.

6. Ibid, p. 38.

7. Ibid, p. 10.

8. Ruth Roach Pierson. "Ladies or Loose Women: The Canadian Women's Army Corps in World War II", Atlantis, Vol. 4, No. 2, p. 245-266.

9. Ibid, p.249.

10. Department of National Defence, Directorate of Army Recruiting. Why Women Join and How They Like It. Report of Enquiry, 1943, p.22.

11. Ruth Roach Pierson. "Ladies or Loose Women ...", p.249.

12. Ibid, p.250.

13. Memo of 18 Dec. 1944 to DGMS from Major G.C. Maloney, RCAMC, CWAC Consultant, PAC, Reel No. C-5296, file HQC 8972. Quoted in R.R. Pierson, "Ladies or Loose Women ...", p.254.

14. Dominion Bureau of Statistics. Canada Year Book, 1945, p.145.

15. Ruth Roach Pierson, "Ladies or Loose Women...", p.254.

16. Sexual Report on Discharged Personnel, June 1943, PAC, Reel No. C-5296, file HQC8972. Quoted in R.R. Pierson "Ladies or Loose Women ...", p.250.

17. William Arkin. "Military Socialization and Masculinity", Journal of Social Issues, Vol.34, No.1, 1978, p.156.

18. Ruth Roach Pierson. "Ladies or Loose Women...",p. 254.

19. Ibid, p.254.

20. Ibid, p.254.

21. Memo of 23 Aug 1943 to Lieutenant Colonel Margaret Eaton, AAG, CWAC from Brigadier G.B. Chisholm, Director General of Medical Services, PAC, RG24, Reel No. 5996, file HQC8972. Quoted in R.R. Pierson "Ladies or Loose Women ...", p.256.

22. Minutes of meeting in office of DGMS on pregnancy in the CWAC, 27 May 1943, PAC, RG24, Reel No. C-5296, file HQC8972. Quoted in R.R. Pierson, "Ladies or Loose Women ...", p.257.

23. Minutes of a meeting of the Combined Services Committee, 9 June 1943, PAC, RG24, Reel No. C-5303, file HQS8984-2. Quoted in R.R. Pierson, "Ladies or Loose Women ...", p.252.

24. Ruth Roach Pierson. "Ladies or Loose Women ...", p.251.

24. Morris Janowitz. "Military Institutions and Citizenship in Western Societies", Armed Forces and Society, Vol 2, No. 2, February, 1976.

25. David Segal, Nora Kinzer & John Woelfel. "The Concept of Citizenship and Attitudes Toward Women in Combat".

27. Morris Janowitz. "The All-Volunteer Military as a 'Sociopolitical' Problem", Social Problems. Vol. 22, No. 3 (February 1975): p. 435.

28. Ramsey Cook & W. Mitchenson. The Proper Sphere, p.46.

29. Albie Sachs & Joan Hoff Wilson. Sexism and The Law: A Study of Male Beliefs and Legal Bias in Britain and the United States. Oxford: Martin Robertson & Company Ltd., 1978, p.64.

30. Ibid, p. 55.

31. Segal et al., p.8.

# SOCIAL CHANGE AND THE PARTICIPATION OF WOMEN IN THE AMERICAN MILITARY*

Mady Wechsler Segal and David R. Segal
Walter Reed Army Institute of Research University of Maryland

*This is an abridged version of a paper to be published in Louis Kriesberg, ed., Research in Social Movements, Conflicts, and Change, Vol.5. Greenwich: JAI Press, forthcoming.

SUMMARY. The representation of women in the United States armed forces has increased from less than 2 percent of the force in 1971 to approximately 8 percent in 1981. Early in the Carter administration, it had been projected to reach 12 percent by the mid-1980s. However, opposition to this goal within the defense establishment became apparent in the late 1970s, and decisions were made during the first year of the Reagan administration to postpone further increases until the impact of greater representation of females among our military personnel could be more systematically assessed. It is our thesis that policies regarding the utilization of women in the American armed forces have resulted primarily from technological, demographic, and gender role changes.

## INTRODUCTION

The first set of factors affecting the increased utilization of women reflects changes in the nature of the military institution. These include changes in military technology that make warfare more capital intensive and permit a reduction in the size of basic weapon systems, a change in the definition of military mission that has deemphasized the concept of wartime mobilization and emphasized in its place the existence of a force in being to fulfill deterrence and constabulary functions even in peacetime, and a change in our philosophy of military manpower management related to the conversion from a military force based upon a mixture of conscription and voluntarism to an all-volunteer armed force.

The second set of factors is demographic, and reflects the bursting of the baby boom bubble in the late 1950s. The birth dearth of the 1960s, responsive in part to ecological concerns regarding population growth, will yield increasingly small cohorts of young men of traditional age-eligibility for military service. If we assume a force-in-being of constant or increasing size, means must be found to expand the pool of people available for military service.

The third set of factors reflects changing roles of women in the United States, including increased participation in the labor force, and broader citizenship participation generally. As military service in the United States has come to be increasingly defined as simply another form of employment in the era of the all-volunteer force, increased representation of women in the military can be seen as an outgrowth of their greater labor force participation. Beyond this, given the traditional association of the "right to fight" with other kinds of citizenship rights and obligations, the increased utilization of women in the military can be seen as a demand for, and manifestation of, advances made in the ongoing citizenship revolution. Fluctuations in policies regarding the utilization of women in the military can likewise be seen as a lack of national consensus on the extension of full citizenship to women. This lack is reflected as well in the

AD P001417

failure to ratify the Equal Rights Amendment.

## THE CHANGING NATURE OF MILITARY ORGANIZATION

The military institution has changed historically, both as a function of general technological development, and as a function of changing technologies of warfare in particular. Although conflict between political units has been an ever present characteristic of human society (Andreski 1968), the emergence of large standing armed forces, as opposed to armies composed of agricultural workers mobilized to fight wars when they were not engaged in harvesting or sowing, was dependent on the ability of a social unit to produce the economic resources necessary to maintain large numbers of people outside the domestic productive economy. In the pre-industrial world, this economic base was provided by the armed force itself, through conquest and expansion.

Modern economic systems, by contrast, are able to produce a sufficient surplus to maintain professional soldiers even when they are not engaged in imperial conquest. The return to such an investment of societal resources is low, however, relative to other potential uses. Thus, at least through the first part of the twentieth century, the industrial nations of the West utilized a mobilization model of military manpower, maintaining relatively small nuclei of military organizations in peacetime, and expanding the force in times of conflict by taking large numbers of people out of civilian roles, and making soldiers of them. This was accomplished largely through conscription.

The mobilization model assumed that in the event of war, the states involved would have time to raise, train, and field their fighting forces, and that the peacetime nucleus could fill the organizational and training functions, as well as necessary defensive functions, until the newly mobilized force was ready to take the field. Technological changes in the mid-twentieth century, however, deprived nations of the luxuries of time and distance from the battlefield that the mobilization model assumed.

## DEFINITION OF MILITARY MISSION

The increasing power of military technology, and the waning of the era of military imperialism in the West, made the mobilization model increasingly inappropriate. Not only did nations find themselves deprived of the lead time required for mobilization for large-scale wars, but as it became obvious that in a confrontation between major powers, victory would be Pyrrhic, the military mission came increasingly to be defined in terms of constabulary, or peacekeeping, rather than war-fighting operations (e.g., Janowitz, 1960: 418-441; Janowitz, 1974: 471-508; Moskos, 1976). The distinction between peacetime and wartime became less relevant for military organization, and the need to maintain a large standing force became obvious, as the deterrence concept, and the need to respond rapidly should deterrence fail, assumed primacy. The mass force, based upon the mobilization model, declined after World War II. With the emergence of a "new long term trend...toward smaller, fully professional, and more fully alerted and self-contained military forces; the direction was away from a mobilizatien force to a military force 'in being'" (Janowitz, 1975: 121).

## TOOLS OF MANPOWER MANAGEMENT

The basis of the mobilization model is the process of conscription, which brings people into the armed forces not with the intention of making career soldiers of them, but rather as

exacting a citizenship obligation in support of national security, which is assumed to be a public good. The process of conscription has appeared most legitimate in wartime, in the face of an apparent threat to national security. However, as the military mission has deemphasized war-fighting and stressed instead deterrence and constabulary operations, the external threat has been less apparent, the military mission more ambiguous, and the conscription process more difficult to justify. Indeed, Van Doorn (1975) and others have noted a general trend away from military conscription in the industrialized nations of the West, as an element of the decline of the mass army. The transition from a force combining conscription with voluntarism (some of which voluntarism was conscription-motivated) to an all-volunteer force (which some have characterized as a system of economic conscription) occurred in the United States in 1973.

The change from conscription to an all-volunteer force was part of a broader redefinition of the nature of military service in America. In a presentation focusing on enlisted personnel, at the 1973 meetings of the American Sociological Association, Charles Moskos noted, almost in passing, an "organizational shift from a predominantly institutional format (i.e., legitimized by normative values) to one more resembling that of an occupation (i.e., akin to civilian marketplace standards)" (Moskos, 1973). Policymakers in the United States have been greatly influenced by Moskos' conceptualization, and behave as though a choice must be made between these definitions. At least as far as the recruitment of personnel is concerned, they have emphasized the similarities between military service and civilian employment. Econometric assumptions of military service made by President Nixon's Commission on an All-Volunteer Force (1970), and by military recruiting strategies that have tried to compete with civilian employers for high quality personnel, have emphasized the least traditionally military characteristics of service, and have emphasized pecuniary rewards and skill training, rather than symbolic and solidary incentives. This has brought into the armed services young people who in fact think of their service as a job, and tend not to think of war-fighting as a part of that job (Gottlieb, 1980). The progressive redefinition of military service in terms of civilian labor force processes, coupled with an emphasis on affirmative action employment programs within the entire federal government structure, in turn, had implications for the perceptions by groups discriminated against in the private sector of the civilian labor force, notably racial and ethnic minorities and women, regarding employment opportunities in the military.

## THE CHANGING DEMOGRAPHIC CONTEXT

The changes in the nature of military organization, military mission, and, perhaps most importantly, military manpower policies, that we have discussed above, have taken place in the context of fluctuating demographic patterns. Indeed, it is possible that the very size of the post World War II baby boom cohorts, leading to a very small percentage of those liable to conscription in fact being drafted, contributed materially to the perceived inequity of the draft, and to its eventual demise.

Prior to World War II, the birth rates in most industrial nations of the West reached their nadirs, during the depression of the 1930s. However, they rose during the 1940s and 1950s, and indeed, while they declined again starting in the late 1950s,

they remained above the depression level until the early 1970s. The average number of births per woman in the United States had fallen to about 2.2 during the depression. It reached a zenith of 3.3 in the late 1950s, and subsequently has returned to the lower level.

The growth of the population during the 1940s and 1950s, influenced both by high birth rates and by large numbers of young women, is the period referred to as the "baby boom," and produced the cohorts that were to come of military age eligibility between the late 1950s and the late 1970s, during most of which time the United States had a system of military conscription.

The fertility decline since the late 1950s reflected the fact that people delayed their marriages longer than they had in the 1940s and early 1950s. In addition, they delayed the births of their first children longer, and spaced their children more widely. In part, this reflects the widespread adoption of new and more effective methods of contraception in the 1960s and 1970s, allowing parents to exert some choice in family size.

Declining fertility has implications for military personnel policy. The number of 18-21 year old males in the American population peaked in 1978. This is the last baby boom group. The decline from 1978 to 1982 has been modest: less than 1 percent per year. The major effect of the birth dearth of the 1960s will be observed between 1983 and 1987, when the decline in cohort size will increase to 2.5 percent per year. By 1990, the number of 17-21 year old American males will be 17 percent below the 1978 level of 10,800,000 young men.

## THE CITIZENSHIP REVOLUTION

Changes in the roles of women in the armed forces, and in the labor force more generally, can be seen as a reflection of an ongoing transformation of Western societies, which "have steadily moved to a condition in which the rights of citizenship are universal" (Bendix, 1964: 3). Marshall (1950) has noted the importance of military service as an obligation of citizenship, and Janowitz has analyzed the central role played by military institutions in the evolution of parliamentary democracy, and of military service as a component of citizenship. "From World War I onward, citizen military service had been seen as a device by which excluded segments of society could achieve political legitimacy and rights" (Janowitz, 1975: 77-78).

In twentieth-century America, this relationship between military service and citizenship has been most dramatic with regard to the racial integration of the armed forces. Through the World War II period, the incorporation of blacks into the civilian citizenry was only minimally effective, and blacks in the military served in segregated units, under a quota, for the most part limited to non-combat jobs, and with an infinitesimally small likelihood of being commissioned as an officer. It was not until 1950, under the direction of President Truman's 1948 executive order to desegregate the armed forces, and most importantly the manpower requirements of the Korean War, that segregation, the quota, and the combat exclusion truly disappeared. The racial integration of the armed forces during the Korean War preceded the gains achieved by the civil rights movement toward racial integration and equality in American civilian institutions. The integration of blacks into the armed forces anticipated the issues raised with regard to integrating women by three decades (Segal, Kinzer and Woelfel, 1977).

The relationship between military service and political

citizenship was more recently demonstrated in the case of young adults. One of the themes of the movement against the Vietnam War was that young men between the ages of eighteen and twenty-one, who were liable to military conscription, were not eligible to vote. They were, therefore, not able to participate in the political processes that selected the members of the executive and legislative branches of the federal government who determined American military policy, including policies regarding the waging of war. The contribution of the anti-war movement to the end of conscription was preceded by its contribution to the passage, in 1971, of the twenty-sixth amendment to the U.S. Constitution, which lowered the age of political majority to eighteen, allowing those liable to military conscription to participate in the electoral process. The unfortunate lesson of the Korean and Vietnam Wars with regard to the extension of citizenship rights to blacks and to young adults may be that the military serves as a vehicle for the citizenship revolution primarily during times of war.

The social strains reflecting the extension of equality to women, in both civilian and military institutions in America, can be seen as the current phase of the citizenship revolution. The major barrier to women's participation in the armed forces has been the constellation of cultural values about appropriate roles for women.

CHANGING FAMILY AND LABOR FORCE PATTERNS

Women's participation in the labor force is dependent upon the degree to which they are free from family responsibilities, their motivation to be employed, and the availability of jobs. Trends in family patterns in the U.S. show women marrying later, married women having fewer children, more women choosing not to have children, and more female-headed families (Current Population Reports, 1975; U.S. Department of Labor, 1969). These changes, as well as increased longevity, have contributed to the housewife role becoming a less exclusive role for women. Women are spending considerably less of their adult lives preoccupied with raising children and running households.

The motivation of women to be employed has been increasing as a result of both financial and ideological factors. High rates of inflation have occurred at a time when the women's movement for equality has encouraged women to recognize and express desires to work as a means of personal fulfillment, power, and financial independence.

The women's movement has also affected the availability of jobs to women and the attractiveness of those jobs. Affirmative action programs have increased job opportunities for women. Legislation, court actions, and private sector policies have helped to decrease wage discrimination against women, relative to men in the same jobs.

All of these trends have contributed to a virtual revolution in the proportion of women in the labor force. The labor force participation rates of women have increased continuously since 1920 and increases since 1950 have been dramatic. By 1980, 51 percent of all American women were in the labor force.

Despite the increased labor force participation of women, they are still concentrated in relatively few occupations: clerical work, service work, retail sales, teaching, and nursing. These occupations account for more than two-thirds of employed women (U.S. Census Bureau, 1979). In addition, the average income of full-time workers is substantially lower for women than

men. This difference persists when we control for the effects of age, education, marital status, and occupational prestige.

## WOMEN IN PREDOMINANTLY MALE SPHERES

Most roles culturally defined as appropriate for women, both family roles and work roles, are "characterized by their supportive, enabling, facilitating, and vicarious features" (Lipman-Blumen and Tickamyer, 1975: 309). These features are absent from predominantly male arenas, including not only the military, but also mathematics, science, corporate management, police work, and sports. These fields have been traditionally socially defined as male domains; girls and women have been socialized to avoid participation in these fields.

During the past two decades, there has been a great deal of social change in the actual and normatively expected roles for American women, including greater recognition of girls' and women's abilities in male-dominated areas. As stereotypes and norms are altered and the expectations communicated to girls change, we can expect more girls to pursue study and careers in previously male fields. If this happens, there would be an increase in the number of women who desire to enter the military and who have the requisite abilities to perform in military jobs. At the present time, however, the socialization experiences of the cohorts already of military age act to minimize aspirations for military service among women, and accentuate resistance among men to the notion of women in the military. While movement of females into male dominated areas is slow, and not inevitable, there are nonetheless some indications that it is occurring.

## WOMEN IN THE U.S. MILITARY

Before World War II, with some minor exceptions, women in the U.S. military served only as nurses. These nurses were under a separate command structure from regular military personnel. During World War II, each of the services established a women's unit, distinct from the nurse corps and also distinct from the rest of the force, with a separate command structure. (For a thorough history of women's participation through World War II, see Treadwell, 1954.) Only recently have the women's branches been integrated with the men's armed services: in 1978, Congress passed legislation abolishing the Women's Army Corps as a separate unit.

The number of women in the military has varied greatly, while the percentage has always been small. The largest number and concentration of women in the U.S. military, until recently, occurred in 1945, when approximately 265,000 women constituted 2.2 percent of the force of over 12 million (Goldman, 1973: 895). Legislation passed in 1947 and 1948 placed severe limitations on the numbers and functions of military women (see Binkin and Bach, 1977: 10-12). A ceiling of two percent was placed on the percentage of enlisted personnel who could be female, and female officers (not counting nurses) could number at most 10 percent of enlisted women. In the 1950s, the number of military women varied between approximately 22,000 in 1950 (1.5 percent) and approximately 35,000 in 1955 (1.3 percent). In 1967, the two percent limitation was removed, but by 1971 women still constituted less than two percent of the military (about 42,800).

From 1971 to 1980, the number and percent of women in the U.S. military has increased dramatically. At the end of fiscal year 1971, there were about 30,000 enlisted women and 13,000 officers, together constituting 1.6 percent of the total active

duty military personnel . By the end of 1980, there were about 151,000 enlisted women (8.6 percent of enlisted strength) and 22,000 female officers (7.9 percent of all active duty officers), for a total of 173,000 women in uniform (8.5 percent of the total armed forces).

The variety of jobs performed by military women parallels the pattern of their numbers. That is, during peacetime, women have played only "traditionally" female roles in the military. During World War II, although women were still concentrated in a few job classifications, the pressures of wartime necessity opened other jobs to them. The end of World War II saw a return to limitations on women's military jobs.

In addition to the recent increase in the number and percent of female enlistees, the past ten years have witnessed an increase in the number of job specialties open to women. Currently, only the combat specialties are closed to women. In 1972, the percentage of enlisted women in the nontraditional jobs was less than 10 percent. By the end of 1980, 55 percent of enlisted women were in traditionally all-male specialties.

The groundwork for the increases in the 1970s in the numbers and roles of women in the U.S. armed forces was laid in early 1973. At the start of the all-volunteer period, the number of male enlistees was falling short of the goals set by the military in order to maintain an effective active duty force. In addition, those who were attempting to enlist were coming increasingly from what the military considers "low quality" personnel: those who had not graduated from high school and those in the lower categories on the aptitude tests used by the services. The services were aware that while they were lowering enlistment standards for men, they were turning away women who wanted to enlist, despite the fact that these women were high school graduates and scored in the upper categories on the mental tests. This was because the armed forces had separate goals for the number of new enlisted men and women. At the same time, it was anticipated that the Equal Rights Amendment, which had been passed by Congress in 1972, would be ratified by the States. Military leaders and civilian policy makers within the defense establishment recognized that such ratification would make their extant policies regarding the enlistment and assignment of women unconstitutional. These pressures directly contributed to the policy of increasing women's participation in the military, the first step of which was to substantially increase the quotas for female enlistments. The number of women officers was then increased and women were admitted to the service academies for the first time in 1976.

Current laws and policies regarding the positions that may be held by women vary among the different services. In all services, women are permitted to hold all jobs that do not involve direct combat. Women are permitted to and do serve in combat support and combat service support specialties, which may involve service in a combat environment.

In the Navy and the Marine Corps, until recently, women were restricted from serving aboard most ships by 10 U.S.C. 6015 (1976) which stated in part: "...women may not be assigned to duty in aircraft that are engaged in combat missions nor may they be assigned to duty on vessels of the Navy other than hospital ships and transports." The effect of this statute was to bar women from service aboard ships (since the Navy currently has no hospital ships or transports). In Owens v. Brown, Judge John J.

Sirica ruled that the Navy could not use this statute as the sole basis for excluding women from duty aboard ship. In 1978, Congress passed a modification of this law, to permit women to serve on hospital and transport ships and other such vessels not expected to be assigned combat missions and to serve up to six months temporary duty on other Navy vessels. In the Navy, women still may not serve on vessels or aircraft engaged in combat missions" (Department of Defense, 1978: 76). Women in the Air Force are similarly prohibited (by U.S.C. 8549) from serving on aircraft engaged in a combat mission.

The Army has no statutory prohibition against women in combat. The current Army policy states: "Women are authorized to serve in any officer or enlisted specialty, except some selected specialties, in any organizational level and in any unit of the Army except infantry, armor, cannon field artillery, combat engineer, and low altitude air defense artillery units of battalion/squadron or smaller size." (HQ DA message, DAPE-MPE-C5, Washington, D.C., R082058z, 8 Sept 77.)

The Subcommittee on Military Personnel of the House Armed Services Committee held hearings in November 1979 on the utilization of women in the military. Included in their public hearings was consideration of the repeal of sections 6015 and 8549 of Title 10, which are the only laws prohibiting American women from serving in combat. While such repeal is not immediately forthcoming, it is noteworthy that the Department of Defense was in favor of such repeal at that time. Even if these legal restrictions were removed, it is still likely that Department of Defense policy would restrict the combat role of women, at least in the immediate future.

In February 1980, with Americans held hostage in Iran and Soviet troops in Afghanistan, President Carter called for the military registration of all American males and females born in 1960 or 1961. The debate over such registration centered around two major issues. First, there was (and still is) disagreement both in Congress and the general public as to whether such registration and a possible return to military conscription is necessary or desirable. Second, the proposed registration of women for the draft created a public discussion laden with emotion.

The draft registration bill enacted by Congress in 1980 excluded women, and was declared unconstitutional by a federal district court in Philadelphia as a violation of the equal protection guarantees of the Fifth Amendment. That ruling was overturned by the Supreme Court in June 1981 (Rostker v. Goldberg), with the Court basing its decision primarily on its interpretation of Congress' prerogatives on military matters.

Through the end of 1980, the military planned to continue to increase the numbers of women to about 254,000 (223,700 enlisted and 30,600 officers) by 1985, with women constituting about 12 percent of active duty personnel. At the beginning of 1981, the Reagan administration announced that it would reexamine those goals, and keep the number of women at the 1980 levels for the present time. The basis of this "pause" in the increasing utilization of women is stated to be a concern about problems encountered with integrating women into the services and anticipation that such problems would be exascerbated by increasing female representation, especially in combat support units. These problems include the following: lost duty time of pregnant women and consequent shortages of personnel in their

units; lack of acceptance and sexual harassment of women by men; physical strength limitations of women; and high rates of attrition of women from traditionally male specialties. Increasing enlistment of women would require recruiting more women for traditionally male jobs. It would also involve accepting more women without high school degrees and lower aptitude levels, who have higher rates of attrition and disciplinary problems in the military.

The pause affected the services in very different ways. The Army stopped recruiting women in mid-1981 for the rest of the fiscal year. The Marine Corps, which expanded programs for women during the 1970s "refined" these programs going into the 1980s, closing some specialties and some units to women. The Navy has been assigning more women to sea duty, and still projects a slight increase by 1985. The Air Force was already 11 percent female in 1981, with women serving in all officer career fields and all but four enlisted career fields, and 30 percent of the women serving in traditionally male specialties.

CONCLUSION

As part of the traditional gender-based definitions of social roles, the participation of women in military forces has historically been limited, except under the most extreme circumstances. The number of women allowed to serve has been kept low except during periods of wartime mobilization when, through the World War II period, their service was in an auxiliary capacity. In times of peace, the women who did serve filled traditionally female roles, such as nurses and clerks. During wartime mobilization, the women were allowed to move into traditionally male roles, and this was as true in civilian factories as in the armed forces. They were, however, excluded from combat jobs, and with the end of hostilities, once more moved into the traditionally female areas. Indeed, the only historical evidence we have of the participation of women in combat specialties occurs in instances of extreme and immediate threat to territorial integrity. This is most likely in guerilla warfare, as in the case of Israel prior to independence, and the Balkan states during World War II. Even in these instances, women seem to have been quickly relegated to more traditional roles. The only modern example of female participation in conventional combat formations was the Soviet Union during World War II. Here there was an immediate and great threat to territorial integrity, manpower resources were exhausted, and the use of womanpower was seen as a last resort.

The utilization of women in the United States armed forces during the 1970s can be seen as a divergence from historical precedent. In a peacetime period, quotas for women were raised, but not eliminated. Women were admitted to all traditionally male specialties except those defined as direct combat. Gender segregation of the armed forces was reduced through the elimination of the women's branches, which at the same time deprived women in the services of advocacy at a high level in the organizational structure. Women were admitted to the service academies. And the issues of drafting women and the utilization of women in combat became matters of public debate.

Going into the decade of the 1980s, the engine of social change appears to have slowed. The facts that the E.R.A. has not yet been ratified by the states, and that Ronald Reagan succeeded Jimmy Carter in the presidency, manifest a conservative mood in the country. With regard to military manpower, this has been

reflected in the exclusion of women from draft registration, and a levelling of the proportion of women in the armed services.

The effects of technology and of public opinion are in part in opposition to each other with regard to the utilization of women in the military. Where the material technology of a service is similar to technologies found in the civilian labor force, and particularly where that technology substitutes capital intensive automated conflict for more traditional mass face to face battles, acceptance of women is higher. Thus, women have been integrated most fully into the high technology Air Force. The Navy, with a considerably more traditional structure but a high technology base, still projects increases in female utilization. The acceptability of women in these traditionally male roles, however, is limited to those roles that have counterparts in the civilian labor force, into which women are also moving. Public opinion has not gone the next step and defined as acceptable women serving in the traditionally male ground combat specialties that do not have counterparts in the civilian labor force. Thus, the greatest restrictions to expansion are found in the Army, and the Marine Corps.

This is not to say that the pendulum will necessarily swing all the way back in a traditional direction. Changes in the direction of gender equality in the military have been made which will be very difficult to reverse. Even the reinstitution of a male only draft, for example, will not justify the exclusion of women serving as volunteers, or indeed attending the service academies.

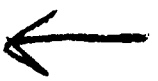
At the same time, it is important not to overstate the magnitude of the changes that have taken place. At the peak of enthusiasm about women in the services, in the late 1970s, they remained a small proportion of the force, they were excluded from combat roles, and they in fact preferred traditionally female jobs. Despite the fact that some women did enter, and continue to enter, traditionally male jobs, both in the civilian labor force and in the military, the number doing so remains small. Furthermore, such women experience serious problems on the job, including lack of acceptance by the men, intense performance pressures, social isolation, and sexual harassment (Kanter, 1977; Martin, 1980; Meyer and Lee, 1978). It appears that the degree of change in social attitudes has not been sufficient to allow full integration in traditionally male arenas, including the military. The experience of the armed services during the decade of the 1970s may well indicate that they have reached a threshold beyond which they cannot advance unless there is further social change in women's roles and in men's attitudes in American society, or unless definitions of citizen rights and obligations are once more changed to meet the mobilization requirements of a war. In the short run, it must be recognized that the expansion of women's roles in the armed forces in the 1970s has already served to reinforce equalitarian sex role attitudes in the civilian sector. If the military continues to expand women's participation, or even maintains it at present levels, it will in the process be serving as a critical agent for future social change.

## REFERENCES

Andreski, Stanislav
1968 Military Organization and Society. Berkeley: University of California Press.

Bendix, Reinhard
1964 Nation-Building and Citizenship. New York: Wiley.
Binkin, Martin and Shirley J. Bach
1977 Women and the Military. Washington: Brookings Institution.
Current Population Reports
1975 "Estimates of the population of the United States and components of change: 1974 (with annual data from 1930)." Population Estimates and Projections Series P-25, No. 545. Washington: U.S. Government Printing Office.
Department of Defense
1978 America's Volunteers. Washington: Office of the Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics.
Goldman, Nancy
1973 "The changing role of women in the military." American Journal of Sociology 79:892-911.
Gottlieb, David
1980 Babes in Arms. Beverly Hills: Sage.
Janowitz, Morris
1960 The Professional Soldier. Glencoe: Free Press.
1974 "Toward a redefinition of military strategy in international relations." World Politics 26:471-508.
1975 Military Conflict. Beverly Hills: Sage.
Kanter, Rosabeth Moss
1977 "Some effects of proportions on group life: skewed sex ratios and responses to token women." American Journal of Sociology 82:965-90.
Lipmen-Blumen, Jean and Ann R. Tickamyer
1975 "Sex roles in transition: a ten-year perspective." Pp. 297-337 in Alen Inkeles, James Coleman, and Neil Smelser (eds.), Annual Review of Sociology, Volume 1. Palo Alto: Annual Reviews Inc.
Marshall, T.H.
1950 Citizenship and Social Class. Cambridge: Cambridge University Press.
Martin, Susan Ehrlich
1980 Breaking and Entering: Policewomen on Patrol. Berkeley: University of California Press.
Meyer, Herbert H. and Mary Dean Lee
1978 Women in Traditonally Male Jobs: The Experience of Ten Public Utility Companies. Washington: U.S. Department of Labor.
Moskos, Charles C.
1973 "Studies on the American soldier." Paper presented at the annual meeting of the American Sociological Association, New York, August.
1976 Peace Soldiers. Chicago: University of Chicago Press.
Segal, David R., Nora Scott Kinzer, and John C. Woelfel
1977 "The concept of citizenship and attitudes toward women in combat." Sex Roles 3: 469-477.
Treadwell, Mattie
1954 The Women's Army Corps. Washington: Office of the Chief of Military History.
U.S. Census Bureau
1979 Population Profiles of the United States, 1979.
U.S. Department of Labor
1969 Handbook of Women Workers. Women's Bureau Bulletin No. 294. Washington: U.S. Government Printing Office.
Van Doorn, Jacques
1975 "The decline of the mass army in the west." Armed Forces and Society 1:147-57.





AD P001418

# Supervisors' Attitudes Toward Women and the Performance Appraisals Given to Men and Women in the Canadian Forces

Suzanne P. Simpson
Captain
Canadian Forces Personnel Applied Research Unit
Downsview, Ontario,
Canada

Henry Edwards PhD
University of Ottawa
Ottawa, Ontario, Canada

Hypothesizing that supervisors' attitudes towards the rights and roles of women would be related to the kinds of performance evaluations given to women in the Canadian Forces, the performance scores of a subset of women evaluated in 1979 and a matched sample of men were examined in relation to the supervisors' scores on the short form of the Attitudes Toward Women Scale (AWS). The predicted relationship held true for three out of the seventeen scales, Support of Subordinates, Supervision, and Command and Self-Assertion - that is, women were rated lower than men by supervisors expressing traditional views about the rights and roles of women, and there were no differences between the men and women's scores when evaluated by supervisors with more egalitarian views. Contrary to prediction, women evaluated by supervisors expressing more traditional views were not scored significantly lower than women evaluated by supervisors with more egalitarian attitudes. A model of self-fulfilling prophecy advanced to account for some of the supervisor/subordinate relationships which would cause the lower evaluation of women by supervisors holding more traditional views about the rights and roles of women was not supported in tests undertaken.

Research examining evaluation procedures, and other personnel decisions affecting the careers of men and women in the work place has shown that women are directed towards lower paying jobs requiring less education, and are selected less frequently for management positions (Donahue & Costar, 1977; Fidell, 1970; Heneman, 1977; Rosen & Jerdee, 1974a; Shaw, 1972). Once on the job, women are given fewer professional development opportunities, are promoted less frequently, are viewed as more capable of engaging in a consideration style of leadership, and less capable of engaging in an initiating structure style of leadership (Bartol & Butterfield, 1976; Rosen & Jerdee, 1974b; Terberg & Ilgen, 1975). In addition, employers seem more lenient with women with respect to family demands (Rosen & Jerdee, 1974c). Sex-role stereotyping and the attitudes held by persons in positions of power are frequently cited by researchers as the reasons for the apparent bias in the treatment of the sexes in the employment setting, especially in those jobs and professions not conventionally held by women (Terberg, 1977). While this is the most likely explanation, this relationship has not been clearly demonstrated in most of the research reported to date. One of the aims of this study, therefore, is to determine whether there is a relationship between the attitudes that supervisors hold towards the rights and roles of women, and the evaluations given to women on their job performance relative to those given to their male counterparts. Secondly, an attempt will be made to clarify some of the mechanisms involved in supervisor and subordinate relationships which would create differences in evaluations given to men and women.

A Model of Self-Fulfilling Prophecy

At the last MTA conference a model of self-fulfilling prophecy was advanced in an attempt to delineate the mechanisms involved in supervisor/subordinate relationships which would result in the lower appraisal of women by supervisors holding more traditional attitudes towards the rights and roles of women relative to their male peers, and relative to women evaluated by supervisors holding more egalitatian views (Simpson, 1980).

This model has as its underlying principle the concept of self-fulfilling prophecy. Jones (1977) argues that stereotypes are part of an individual's theory of personality. The act of assigning a label to an individual, or assigning an individual to a distinct social group, results in certain expectancies for the individual consistent with the label - in the case of this study, for the women to behave in a way consistent with the stereotype for their sex. Furthermore, individuals in positions of power can set up the environment, attend to behaviour, and establish reinforcement contingencies so that it is difficult for the stereotyped individual to act in a manner inconsistent with the expectancies. This is very similar to the concept of role entrapment advanced by Kanter (1976) whereby assumptions and mistaken attributions made about token groups force them into playing limited and caricatured roles. Jones (1977) further indicates that not only is the behaviour of the stereotyped individuals modified, but also the expectancies they hold for themselves (Jones, 1977).

The model as outlined in Figure 1 summarizes these processes as they apply to women in the employment setting - specifically, as they apply to women in the Canadian Forces. Briefly, the supervisor's views on the rights and roles of women in society (Box A) impacts on the supervisor's expectancies for the performance of women in the Canadian Forces (Box B), which in turn results in certain expectancies for the performance of the specific woman under his or her employ (Box C). The supervisor will, as a consequence, set up the work environment and respond to the subordinate (Box D) in a manner such that the subordinate's own work performance (Box E), and expectancies for performance (Box F), are modified. The behaviour of the subordinate on the job in turn impacts on and further corroborates the supervisor's opinion about the rights and roles of women as a whole (Box A), his or her expectancies for women as a group in the Canadian Forces (Box B), and his or her expectancies for the specific woman under his or her employ (Box C).

Whether such a system applies to supervisor/subordinate relationships is the subject of this study. The following hypotheses pertain:

1. women subordinates will receive lower performance evaluations than men subordinates when both are rated by supervisors who adhere to traditional attitudes towards the role of women in society;
2. there will be no differences in performance evaluations given to male and female subordinates by supervisors who do not adhere to traditional attitudes towards the role of women in society;
3. women subordinates who work for supervisors who adhere to traditional attitudes towards the role of women in society will receive lower performance evaluations than women subordinates who work for supervisors who do not adhere to traditional views on the roles of women;
4. supervisors who adhere to traditional attitudes towards the role of women in society will have lower expectancies for the performance and work motivation of the specific women under their employ than will supervisors who hold less traditional views;
5. supervisors who hold more traditional views about the roles of women in society will report less encouragement and reward of good performance and more leniency towards poor performance of the specific female subordinates under their employ than supervisors who hold less traditional views; and,
6. women who work for supervisors who hold more traditional views about the role of women in society will have lower expectancies for their own level of performance potential and their own work motivation, than will women who work for supervisors having less traditional attitudes.

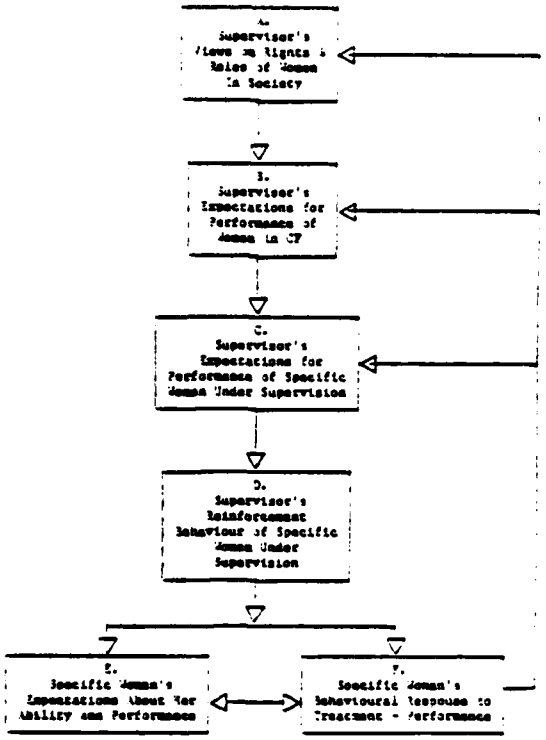



Figure 1. Schematic drawing of the impact of supervisor's expectancies for female subordinates' performance.

## METHOD

### Subjects

The group of potential subjects included all of the women of nonofficer status who had annual Performance Evaluation Reports (PERs) completed on them during the 1978/79 evaluation cycle, and who were still in the Canadian Forces in September 1980. A comparative group of men was selected using the same criteria, but in addition, were matched on a subject-by-subject basis to the group of women by rank and by trade. The result was a group of 391 men and 391 women in the potential subject sample. Their respective supervisors at the time of their PER constituted the potential supervisor sample.

The combined participation rate for the subordinates and subordinates was 59 percent. The reason for nonparticipation could not be determined in 28 percent of the cases, only 7 percent voluntarily chose not to participate, and in the balance of the cases the reasons for nonparticipation were beyond the control of the individuals involved (e.g., on training, on leave, etc.). The male and female subordinates did not differ in rank (CHI-square = .318; df = 2; $p > .05$) or in trade (CHI-square = 7.828; df = 19; $p > .05$); however, the women were significantly younger ($t = -6.39$; df = 355; $p < .001$), had less time in service ($t = -8.20$; df = 325; $p < .001$) and in rank ($t = -7.79$; df = 287; $p < .001$), had more education (CHI-square = 62.41; df = 7; $p < .001$), scored higher in clerical aptitude ($t = 6.15$; df = 253; $p < .001$), and lower in electrical ($t = -2.28$; df = 253; $p < .05$) and mechanical aptitude ($t = -5.74$; df = 253; $p < .001$). There were no significant differences in general intelligence ($t = 1.76$; df = 253; $p > .05$) and arithmetic/computational aptitude ($t = -0.34$; df = 253; $p > .05$).

### Measurement Instruments

Performance Evaluation Report for Members. The PER currently in use assesses the members on seventeen performance requirements on a seven-point scale varying from Below Standard to Rare High Standard. The most recent published report indicates that the coefficients of stability on two successive annual PERs for Corporals and Senior Noncommissioned Officers whose unit and supervisor had remained unchanged, ranged from .53 to .61 for total PER score, and from .39 to .51 for individual performance requirements (Stow, 1973). Using principal components analysis Bain, Skinner and Rampton (1980) found a three factor structure for the PER and labelled them: i) Interpersonal Skills or Influencing, ii) Individual Effectiveness, and iii) Professionalism. The performance requirements loading heavily on each of these factors are shown in Table 1.

Attitudes Toward Women Scale (AWS). The supervisors were asked to complete the short form of the AWS (Spence, Helmreich, & Stapp, 1973) so that the relationship between attitudes of the supervisors toward the rights and roles of women (Box A of the model) and the evaluations received by men and women on their PERs could be examined. The subordinates were also asked to complete the AWS, but the results do not form an integral part of the hypotheses testing and, therefore, will not be reported here.

Table 1. Performance Requirements Loading on the Interpersonal Skills or Influencing, Individual Effectiveness and Professionalism Factors of the PER

| Factor Name | Performance Requirement |
|---|---|
| Influencing or Interpersonal Skills | Delegation<br>Command & Self-Assertion<br>Support of Subordinates<br>Briefing Others<br>Supervision<br>Ensuring Understanding of Assignments |
| Individual Effectiveness | Planning<br>Performance Under Stress/Pressure<br>Cooperation<br>Knowledge of Job<br>Ability to Apply His/Her Knowledge<br>Adaptability<br>Initiative<br>Responsibility<br>Learning from Experience |
| Professionalism | Appearance and Bearing<br>Conduct |

The short form of the AWS is a 25 item Likert-type scale tapping attitudes about the vocational, educational, and intellectual roles of women, their freedom and independence, sexual behaviour, and marital relationship and obligations (Spence & Helmreich, 1972). Amongst the Canadian Military College students internal consistency (coefficient alpha) was .84 to .89, and three week test-retest reliability was .89 (Prociuk, 1980). Among members of nonofficer status internal consistency (Coefficient alpha) measured at .83 and .80 for men and women respectively (Boyce & Belec, 1980). The scale also appears to be sensitive to intervention expected to produce attitude change (i.e. mixed Cadet Basic Training at military college) indicating that the AWS is probably a valid measure of attitudes towards women (Yoder, Rice, Adams, Priest & Prince, 1979).

Expectancy Questionnaires. Both the supervisors and subordinates were asked to complete a questionnaire of very similar wording containing items with Likert-type response options. In the case of the supervisors the questions were aimed at an assessment of the specific subordinate's potential to achieve and work motivation (Box C of the model and Hypothesis 4), as well as how the supervisors would respond to different standards of performance on the part of the specific subordinate in question (Box D of the model and Hypothesis 5). The subordinates were asked to complete a similarly worded questionnaire assessing their own perceived potential to achieve and their work motivation (Box E of the model and Hypothesis 6), as well as how their supervisor would respond to various standards of performance on their part (a cross reference on the supervisor's responses). The surveys also contained other questions intended to provide a contextual framework for the hypotheses in question. These surveys will hereafter be referred to as the Subordinates' and Supervisors' Expectancy Questionnaires.

## Data Collection

The questionnaire data were gathered by sending packages of specially prepared materials for each potential participant to Personnel Selection Officers (PSOs) at Canadian Forces Bases across Canada. All potential participants were told that participation in the study was voluntary. The questionnaires were administered in a standardized fashion in classroom settings, with the supervisors and subordinates completing them in separate sessions. All participants completed the Expectancy Questionnaire followed by the AWS in the official language of their choice, French or English, on machine readable answer sheets. On return, the answer sheets were visually inspected for errors in completion, and were also submitted to computerized editing. PER data and certain biographical information were collected from existing computer files.

## Design and Statistical Analyses

The independent variables are the sex of the subordinate (SEX) and the supervisors' attitudes towards women as measured by the AWS (SAWS). Based on the AWS the supervisors were divided into the categories Traditional, Egalitarian and Moderate depending on whether their scores fell within the bottom, top or middle third of the distribution. Thus, the general design for this study is a 2(SEX) X 3(SAWS).

For Hypotheses 1, 2, and 3, the dependent measures were the seventeen performance requirements of the PER. For hypotheses 4, 5, and 6, the dependent measures were the responses to the relevant items from the Supervisors' and Subordinates' Expectancy Questionnaires. All analyses required a SEX by linear SAWS interaction. The SAWS effect was, therefore, partitioned for testing into first and second degree polynomial contrasts, and the interaction effects tested reflected this partitioning.

All analyses were carried out at the multivariate level and an effect had to meet a minimum level of significance of .05 or less before the univariate analyses of each dependent measure were considered. Dunn's Multiple Comparison Procedure was used to test planned comparisons (Kirk, 1968). In the case of Hypotheses 1, 2, and 3, the planned comparisons are outlined in Table 2.

Hypotheses 4, 5, and 6 demand a comparison of results for women working for Egalitarian supervisors; however, to satisfy the hypotheses within the context of the theory advanced, the results for women relative to those obtained for the male subordinates must be considered. In particular, the differences in results for the female subordinates working for Traditional supervisors and those for female subordinates working for Egalitarian supervisors predicted by Hypotheses 4, 5 and 6 (Prediction 1) imply that there should be no differences between the results for males working for Traditional supervisors and those for males working for Egalitarian supervisors (Prediction 2)

Furthermore, in keeping with theory advanced, it was expected that there should be differences in the results for male and female subordinates working for Traditional supervisors (Prediction 3), but no differences in results for male and female subordinates working for Egalitarian supervisors (Prediction 4). These paired comparisons are summarized in Table 3.

Table 2. Planned Comparisons for Testing Hypotheses 1, 2, and 3.

| Hypothesis No. | Comparison [a] |
|---|---|
| 1 | $SEX_f SAWS_t$ versus $SEX_m SAWS_t$ |
| 2 | $SEX_f SAWS_e$ versus $SEX_m SAWS_e$ |
| 3 | $SEX_f SAWS_t$ versus $SEX_f SAWS_e$ |

a. SEX = Sex of the subordinate; f = female; m = male.
SAWS = Supervisors attitudes towards women; t = traditional; e = egalitarian.

Table 3. Planned Comparisons for Testing Each of Hypotheses 4, 5, and 6 and Related Theory

| Comparison No. | Comparison [a] |
|---|---|
| 1 | $SEX_f SAWS_t$ versus $SEX_f SAWS_e$ |
| 2 | $SEX_m SAWS_t$ versus $SEX_m SAWS_e$ |
| 3 | $SEX_f SAWS_t$ versus $SEX_m SAWS_t$ |
| 4 | $SEX_m SAWS_e$ versus $SEX_m SAWS_e$ |

a. SEX = Sex of the subordinate; f = female; m = male.
SAWS = Supervisors' attitudes towards women; t = traditional; e = egalitarian.

## RESULTS

### Preliminary Results

A principal components analysis of the PER data was completed, which confirmed the factor structure obtained by Bain et al. (1980). Furthermore, it was found that the factor structure was stable and equivalent across the male and female samples. It was, therefore, decided that it would be appropriate to use the performance requirements listed under each of the three factors Individual Effectiveness, Interpersonal Skills or Influencing, and Professionalism (Table 1) as dependent measures in three separate multivariate analyses of variance (MANOVAs) to test Hypotheses 1, 2, & 3.

Because the men and women subordinates who participated in this study differed significantly in age, length of service in the Canadian Forces, length of time in rank, educational level, clerical aptitude, electrical aptitude, and mechanical aptitude, it was necessary to establish the relationships between these variables and the PER performance requirements. No substantial relationships were found for either the men or the women using Pearson's r (no correlation exceeded .28). Moreover, no significant canonical correlation was found for either the men or the women between these two sets of variables ($p > .05$). Based on both the bivariate and the multivariate correlation techniques, it was concluded that the relationships between the predictor variables and the PER scores were not strong enough to warrant the inclusion of the variables in question as covariates in subsequent analyses.

An analysis of the supervisors AWS scores yielded an alpha coefficient of .85, indicating that the measure is internally consistent for the sample (Hull & Nie, 1979). A principal components analysis was also performed and it was found that for this sample the AWS was not essentially unifactorial as was reported by Spence, Helmreich and Stapp (1972) with college students in Texas, and by Prociuk (1980) with cadets at Canadian Military Colleges. Limiting the analysis to three factors produced a stable and interpretable set of factors which, in their unrotated form, accounted for 36 percent of the variance. Using Rummel's (1970) criteria for the interpretation of factors the following descriptive labels were assigned to the factors: Equality of Opportunity & Division of Labour; Morality; and, Courtship and Marriage. Table 4 shows the VARIMAX rotated factor structure for these factors. This factor structure was used to generate scores for the AWS for each of the supervisors who participated in the study. The supervisors were then assigned to the categories Traditional, Egalitarian or Moderate for each of the factors by grouping according to whether their scores fell within the bottom, top or middle 33.3 percent of the distributions so that the hypotheses of the study could be tested for each of the AWS factors.

The Supervisors' and Subordinates' Expectancy Questionnaires were submitted to separate principal components analyses, the results of which were compared to determine what groups of items should be used to represent the variables of concern in the tests of Hypotheses 4, 5, and 6. The discussion of the results of these analyses is beyond the scope of this paper. It is sufficient to say that there were many groups of items which provided contextual information for the hypotheses in question; however, only those groups of items which bear directly on the hypotheses will be discussed in the following Test of Hypotheses section.

### Test of Hypotheses

Tests of Hypotheses 1, 2, and 3. As indicated previously, the dependent variables in these analyses were the three sets of PER performance requirements representing the PER dimensions Individual Effectiveness, Interpersonal Skills or Influencing, and Professionalism (Table 1). In addition, the analyses were performed using each of the three SAWS dimensions Equality of Opportunity & Division of Labour, Morality, and Courtship & Marriage. Thus, there were nine MANOVAS performed to test Hypotheses 1, 2, and 3, one for each combination of PER performance requirement subset by SAWS dimension.

The predicted interaction of SEX by linear SAWS was found for the subset of items representing the PER dimension Influencing, when the Morality factor scores were used to define SAWS ($F=2.594$; $df=6, 234$; $p < .05$; $R=.250$). At the univariate level, using Dunn's Multiple Comparison Procedure, support was obtained for Hypotheses 1 and 2, but not for 3 for the three performance requirements Support of Subordinates, Command and Self-Assertion and Supervision; that is, for these three performance requirements, Traditional supervisors of women rated them significantly lower than did Traditional supervisors rating male subordinates (Hypothesis 1), there were no significant differences in scores for men and women among Egalitarian supervisors (Hypothesis 2), but contrary to prediction (Hypothesis 3), there were no significant differences between evaluations received by women who worked for Traditional supervisors and the evaluations received by women who worked for Egalitarian supervisors. Table 5 summarizes these results, Table 6 gives the means and standard deviations for each of the three performance requirements, and Figures 2, 3, and 4 graphically depict the relationships.

Table 4. VARIMAX Rotated Factor Structure for Supervisors' AWS

| | Abbreviated Question | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|---|
| 1 | Swearing and obscenity more repulsive in speech of a woman. | .31 | .62 | -.05 |
| 2 | Women should take increasing responsibility for intellectual and social problems. | .49 | -.19 | .29 |
| 3 | Both husband and wife same grounds for divorce. | .10 | -.04 | .44 |
| 4 | Telling dirty jokes masculine prerogative. | .35 | .64 | .09 |
| 5 | Intoxication worse among women. | .16 | .65 | -.01 |
| 6 | Under modern economic conditions men should share in household tasks. | .51 | -.06 | .13 |
| 7 | It is insulting to have "obey" clause in marriage. | .26 | -.02 | .40 |
| 8 | There should be strict merit system in job appointment and promotion. | .50 | -.01 | .34 |
| 9 | A woman should be free to propose marriage. | .35 | .19 | .67 |
| 10 | Women should worry less about rights and more about being wives and mothers. | .50 | .30 | .24 |
| 11 | Women earning should bear equally expense when go out. | -.10 | .09 | .67 |
| 12 | Women should assume rightful in business and professions. | .59 | .02 | .30 |
| 13 | Women should not go to same places and have same freedom. | .40 | .39 | .23 |
| 14 | Some more encouragement to go to college. | .47 | .32 | .02 |
| 15 | Ridiculous for woman to drive locomotive and man to dance ballet. | .58 | .35 | .02 |
| 16 | Father more authority in bringing up children. | .47 | .29 | .19 |
| 17 | Women should not be sexually intimate before marriage. | .34 | .39 | .18 |
| 18 | Husband should not be favoured by law in disposal of family property. | .45 | -.20 | .17 |
| 19 | Women should be concerned with childrearing and house-tending. | .51 | .27 | .26 |
| 20 | Intellectual leadership of community should be in hands of men. | .66 | .22 | .09 |
| 21 | Economic and social freedom worth more to women than ideal of femininity. | .17 | -.31 | .29 |
| 22 | Women should be regarded as less capable of contribution to economy. | .70 | .15 | -.00 |
| 23 | Many jobs in which men should be given preference. | .60 | .23 | -.00 |
| 24 | Women should be given equal opportunity for apprenticeship in trades. | .57 | .09 | .18 |
| 25 | Modern girl entitled to same freedom from regulation and control. | .37 | .31 | .45 |

Table 5. Duncan's Multiple Comparison Procedure Testing Hypotheses 1, 2, and 3 for Performance Requirements Support of Subordinates, Command and Self-Assertion, and Supervision, using the Morality Dimension of SAWS

| Hypothesis | Performance Requirement | Duncan's Multiple Comparison Procedure[a] Minimum critical difference | Level of significance |
|---|---|---|---|
| 1 | Support of Subordinates | .499 | $p \leq .05$ |
| | Command & Self-Assertion | .538 | $p \leq .05$ |
| | Supervision | .544 | $p \leq .05$ |
| 2 | Support of Subordinates | .502 | n.s.[b] |
| | Command & Self-Assertion | .542 | n.s. |
| | Supervision | .548 | n.s. |
| 3 | Support of Subordinates | .523 | n.s. |
| | Command and Self-Assertion | .564 | n.s. |
| | Supervision | .570 | n.s. |

a. degrees of freedom = 239
number of planned comparisons = 3
b. n.s. = not significant

Table 6. Means (and Standard Deviations) for Men and Women for the Performance Requirements Support of Subordinates, Command & Self-Assertion, and Supervision by the Morality Dimension of SAWS.

| Performance Requirement | SAWS (Morality) | SEX Women | Men |
|---|---|---|---|
| Support of Subordinates | Traditional | 4.53(0.89) | 5.09(0.99) |
| | Moderate | 4.50(1.09) | 5.08(1.09) |
| | Egalitarian | 4.76(0.90) | 4.71(0.79) |
| Command & Self-Assertion | Traditional | 4.58(1.13) | 5.24(1.07) |
| | Moderate | 4.57(1.06) | 5.32(0.96) |
| | Egalitarian | 4.98(0.87) | 5.07(1.01) |
| Supervision | Traditional | 4.61(1.00) | 5.18(1.09) |
| | Moderate | 4.55(1.11) | 5.21(1.02) |
| | Egalitarian | 4.92(0.98) | 5.00(1.04) |

No other significant multivariate interactions were found among the eight other combinations of PER subset by SAWS dimension.

Tests of Hypotheses 4, 5 and 6. The dependent measures in these analyses were the subsets of variables, determined by principal components analyses, from the Supervisors' and Subordinates' Expectancy Questionnaires which best represented the variables of interest in Hypotheses 4, 5, and 6. As with the analyses for Hypotheses 1, 2, and 3, the subsets of items were analyzed for each of the three SAWS dimensions defined above. Thus, there were three MANOVAs performed for every subset of items analyzed from the Supervisors' and Subordinates' Expectancy Questionnaires.

For Hypothesis 4 the variables requiring measurement were the supervisors' "expectancies for the performance and work motivation" of their subordinates. There were three subsets of items from the Supervisors' Expectancy Questionnaire which defined these variables. The first set contained eighteen items related mostly to the supervisor's expectancies for the subordinate's future performance in the Canadian Forces, and the types of work the supervisor would be prepared to allow the subordinate to do. Collectively these items seemed to measure the supervisor's confidence in the subordinate's work capabilities and fall into the category of measuring the supervisor's "expectancies for the performance" of the subordinate, as well as, the supervisor's appreciation of the "work motivation" of the subordinate to do well at the task at hand (one item). The

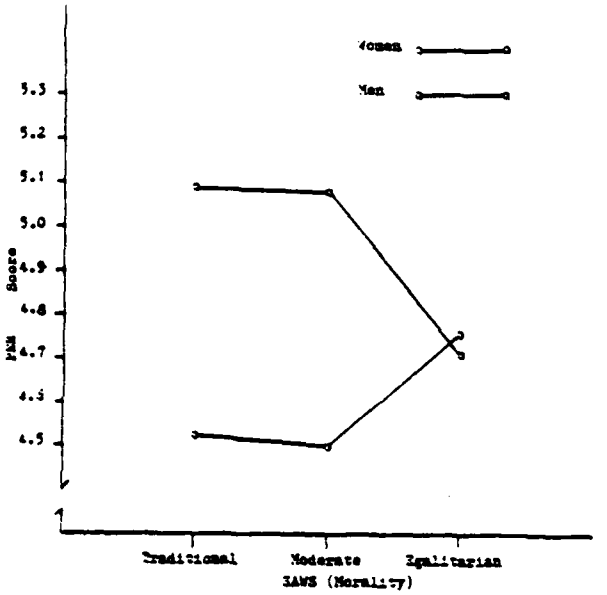



Figure 2. Performance requirement Support of Subordinates for men and women by supervisor's category on Morality dimension of SAWS.

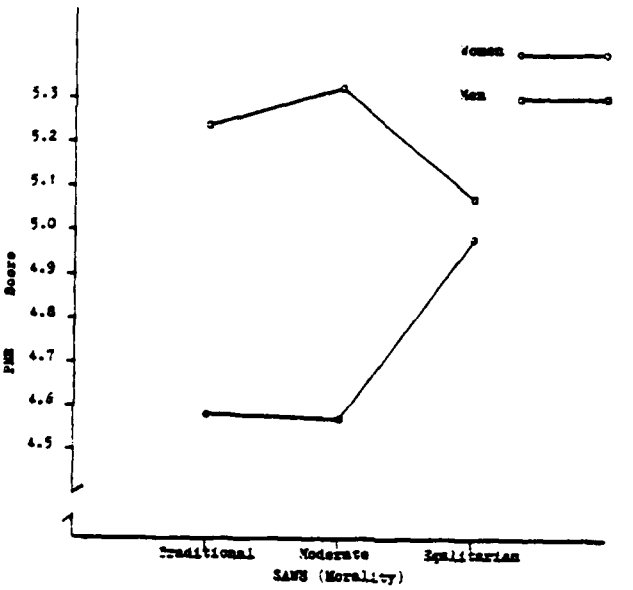



Figure 3. Performance requirement Command & Self-Assertion for men and women by supervisor's category on Morality dimension of SAWS.

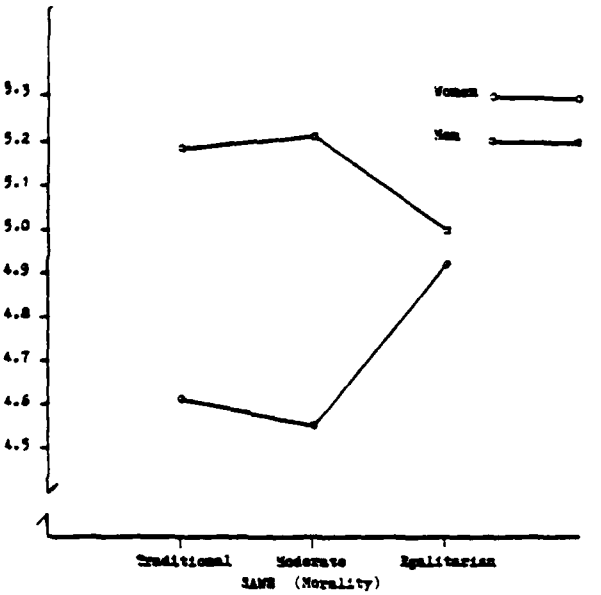



Figure 4. Performance requirement Supervision for men and women by supervisor's category on Morality dimension of SAWS.

second set contained three items measuring the supervisor's willingness to place the subordinate in "Dangerous", "Physically Difficult" and "Combat" positions, and as such measured the supervisor's "expectancies for the performance" of the subordinate in this distinctly military commitment. The third set of items included two questions assessing the supervisor's judgement about the commitment of the subordinate to a career in the Canadian Forces and measured this aspect of the subordinate's "work motivation".

A significant multivariate SEX by linear SAWS interaction was found for the first set of items for the Equality of Opportunity and Division of Labour dimension of the SAWS. At the univariate level, significant results were obtained for the three items, the job descriptors High Profile, Technical and Vital as responses to the question "I would be prepared to place this individual in positions that could be described as". The response options varied from (1) Definitely Yes to (5) Definitely No. For all three job descriptors, predictions 1, 2, and 4 were supported (see Method section) - that is, Traditional supervisors were significantly less willing to place their women subordinates in these types of jobs than were Egalitarian supervisors of women, there were no significant differences between Traditional supervisors of men and Egalitarian supervisors of men in willingness to place their subordinates in these types of positions, and there were no significant differences between Egalitarian supervisors of men and Egalitarian supervisors of women. Only in the case of the job descriptor Technical, however, was prediction 3 supported - that is, only for the descriptor Technical were Traditional supervisors found to be significantly less willing to place their women subordinates in this type of position than traditional supervisors of men.

Table 7 summarizes the univariate comparisons for these three items, Table 8 gives the means and standard deviations, and Figures 5, 6, and 7 depict the relationships.

Table 7. Dunn's Multiple Comparison Procedure - Testing Comparisons 1 to 4 for Supervisors' Willingness to Place Subordinates in High Profile, Vital and Technical Positions Using Equality of Opportunity and Division of Labour Dimension of SAWS

| Comparison No. | Type of Position | Dunn's Multiple Comparison Procedure[a] | |
|---|---|---|---|
| | | Minimum critical difference | Level of significance |
| 1 | High Profile | .501 | $p < .05$ |
| | Vital | .561 | $p < .01$ |
| | Technical | .543 | $p < .01$ |
| 2 | High Profile | .520 | n.s.[b] |
| | Vital | .489 | n.s. |
| | Technical | .467 | n.s. |
| 3 | High Profile | .535 | n.s. |
| | Vital | .497 | n.s. |
| | Technical | .438 | $p < .01$ |
| 4 | High Profile | .524 | n.s. |
| | Vital | .556 | n.s. |
| | Technical | .471 | n.s. |

a. degrees of freedom = 321
number of planned comparisons = 4

b. n.s. = not significant

Table 8. Means (and Standard Deviations) for Men and Women for Supervisors' Indicated Preparedness to Place Subordinates in High Profile, Vital and Technical[1] Positions by the Equality of Opportunity & Division of Labour(EODL) Dimension of SAWS

| Type of Position | SAWS (EODL) | Sex | |
|---|---|---|---|
| | | Women | Men |
| High Profile | Traditional | 2.64(1.05) | 2.41(1.20) |
| | Moderate | 2.30(1.16) | 2.70(1.37) |
| | Egalitarian | 2.14(1.04) | 2.55(1.24) |
| Vital | Traditional | 2.64(1.08) | 2.26(1.02) |
| | Moderate | 2.16(1.08) | 2.09(0.97) |
| | Egalitarian | 2.03(0.98) | 2.12(0.96) |
| Technical | Traditional | 3.16(0.97) | 2.54(0.94) |
| | Moderate | 2.84(1.05) | 2.36(0.92) |
| | Egalitarian | 2.34(1.06) | 2.59(1.20) |

1. Response options are: (1) Definitely Yes (2) Yes (3) Don't Know (4) No (5) Definitely No.

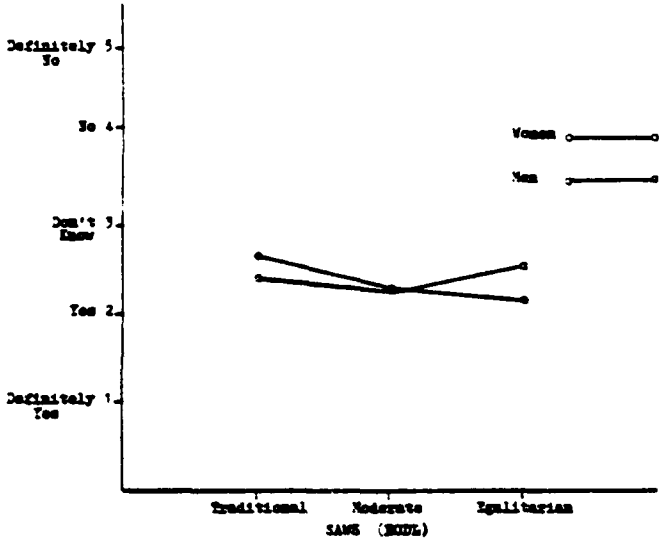



Figure 5. Means for descripter High Profile in response to question "I would be prepared to place this individual in positions that could be described as:" for men and women by the Equality of Opportunity & Division of Labour (EODL) dimension of SAWS.

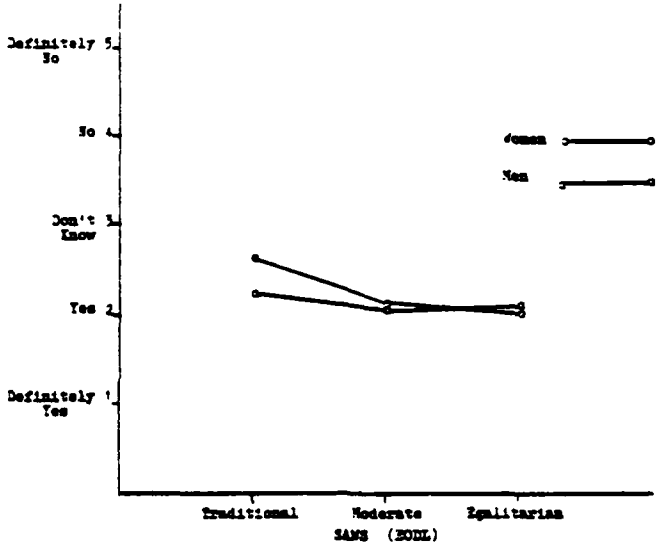



Figure 6. Means for descriptor Vital in response to question "I would be prepared to place this individual in positions that could be described as:" for men and women by the Equality of Opportunity & Division of Labour dimension of SAWS.

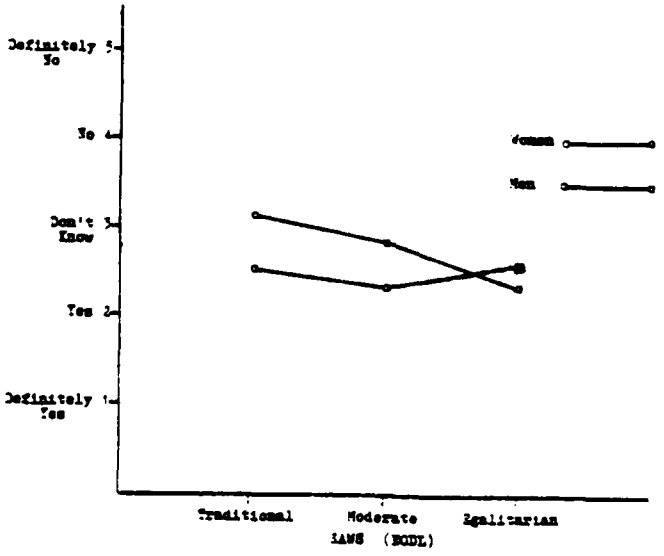



Figure 7. Means for descriptor Technical in response to question "I would be prepared to place this individual in positions that could be described as:" for men and women by the Equality of Opportunity & Division of Labour(EODL) dimension of SAWS.

There were no significant multivariate SEX by linear SAWS interactions for the second and third groups of items related to testing Hypothesis 4. Thus, only partial support was obtained for Hypothesis 4 insofar as the supervisor's willingness to place their subordinates in High Profile, Vital and Technical positions was concerned, and only with respect to the Equality of Opportunity and Division of Labour dimension of the SAWS.

For Hypothesis 5 the variables requiring measurement were the supervisors' reports of the degree of "encouragement and reward of good performance" and "leniency towards poor performance." Two groups of items from the Supervisors' Expectancy Questionnaire were used as the dependent measures in tests related to this hypothesis. The first group contained two questions measuring the supervisor's "leniency towards poor performance" and one measuring "encouragement and reward fo good performance." The second group contained two questions, both measuring the supervisor's "reward and encouragement" of good performance but were distinct in that they measured the material consequences of good performance (promotion and salary increase), whereas the first group of items dealt with the nonmaterial consequences of good and poor performance (verbal commendation, counselling, etc.)

No significant SEX by linear SAWS multivariate interactions were obtained; thus, no support was obtained for Hypothesis 5 based on the supervisors' reports on Expectancy Questionnaire. A cross check was made with the Subordinates' Expectancy Questionnaire by analyzing the same items on the Subordinates' Questionnaire as the first group on the Supervisors' Expectancy Questionnaire. The items on the subordinates' questionnaire asked how the supervisors would respond to good and poor performance on the subordinate's part. No significant multivariate SEX by linear SAWS interactions were found for these questions. Thus, there was no support obtained whatsoever for Hypothesis 5, either in reports from the supervisors or from the subordinates.

For Hypothesis 6 there were two sets of questions from the Subordinates' Expectancy Questionnaire used to examine the subordinates' "expectancies for their own performance potential and their own work motivation". The first group contained five items measuring the subordinates' "expectancies for their own performance" into the future in the Canadian Forces, and the second group contained three items assessing the subordinates' "work motivation", both in terms long term career commitment and motivation to do well at the job at hand.

In neither of these groups of items were there any significant multivariate SEX by linear SAWS interactions found. Thus, no support was obtained for Hypothesis 6 based on reports from the subordinates on the Subordinates' Expectancy Questionnaire.

## DISCUSSION

Prior research (unpublished CFPARU data) has demonstrated that there are many differences between the PERs for men and for women, with women for the most part, being scored lower than men on the PER dimension Individual Effectiveness. To some degree, at least with respect to some of the performance requirements related to leadership and influencing others, the attitudes the supervisors hold towards women are related to the kinds of PERs they give to women relative to those given to men. The differences, however, cannot be explained entirely on the basis of the supervisors' measured attitudes towards women. One possible reason for this may be that the AWS used in this study is not particularly related to the more specific attitudes supervisors may hold towards the roles and opportunities women should have in the military. Another possible explanation is that there are real differences between the performance of men and women on the job, and that these differences are due to the tendency in our society to socialize women towards a feminine ideal, an ideal which is in conflict with the kinds of behaviour and performance required to succeed in the military. This latter explanation receives some support from the Subordinates' Expectancy Questionnaire, in that women feel that there is a lower likelihood of their making a career of the Canadian Forces than do men, indicate that they have a lower capacity to take on increased responsibility and a lower potential to carry out all of the difficult tasks than is reported by men, and judge their potential to function at advanced ranks at a reduced level compared with the reports of men. That all of this is unrelated to the supervisors' AWS measures suggests that perhaps the women's expectancies for themselves were not influenced by their immediate supervisor's attitudes towards women as suggested by the model. This position is further reinforced by the fact that there was no difference in the treatment of women by the more Traditional supervisors compared with the more Egalitarian supervisors reported in either the supervisors' or subordinates' questionnaires.

## SUMMARY AND CONCLUSIONS

Support was obtained for Hypotheses 1 and 2, but not for Hypothesis 3 for the performance requirements related to leadership and influencing others - that is, for three of the performance requirements from the Influencing dimension of the PER, supervisors of women who expressed more traditional views on the rights and roles of women rated their subordinates lower than did supervisors of men expressing the same attitudes; there were no differences between the ratings received by men and women when they were evaluated by supervisors expressing more egalitarian views; but contrary to prediction, the women working for supervisors expressing more traditional views were not rated significantly lower than women working for supervisors expressing more egalitarian views.

Although some support was obtained for Hypothesis 4 with regard to more traditional supervisors having lower expectancies for their female subordinates than the more egalitarian supervisors have for theirs, no support was obtained for (Hypothesis 5 and 6) the position that supervisors act in a manner consistent with their expectancies so that supervisors with lower expectancies cause (through the kinds of reinforcement contingencies they establish and how they set up the work environment) women to perform less effectively, and have lower expectancies for their own potential to achieve.

## REFERENCES


Bain, B.L., Skinner, B.J., & Rampton, D.M. Junior Leadership Training Assessment in the Canadian Forces: Toward an Integration of Training, Research and Personnel Perspectives. Canadian Forces personnel Applied Research Unit, 80-6, 1980.

Bartol, K.M. & Butterfield, D.A. Sex effects in evaluating leaders. Journal of Applied Psychology, 1976, 61, 446-454.

Boyce, D.G. & Belec, B.B. Attitudes toward women's roles: A preliminary analysis of Canadian Forces personnel. Proceedings of the 22nd Annual Conference of the Military Testing Association, 1980.

Donahue, T.J. & Costar, J.W. Counselor discrimination against young women in career selection. Journal of Counselling Psychology, 1977, 24(6). 481-486.

Fidell, L.S. Empirical verification of sex discrimination in hiring practices in Psychology. American Psychologist, 1970, 25, 1094-1097.

Heneman, H.G. Impact of test information and applicant sex on applicant evaluations in a selection simulation. Journal of Applied Psychology. 1977, 62(4), 524-526.

Hull, C.H. & Nie, N.H. SPSS Update, New York: McGraw-Hill Book Company, 1979.

Jones, R.A. Self-Fulfilling Prophecies. New York: John Wiley & Sons, 1977.

Kanter, R.M. Some effects of proportions on group life: Skewed sex ratios and responses to token women. American Journal of Sociology, 1976, 82(5), 965-990.

Kirk, R.E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, California: Brooks/Cole Publishing Company, 1968.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. & Bent, D.H. New York: McGraw-Hill Book Company, 1975.

Prociuk, T.J. Women At Canadian Military Colleges: A Survey of Attitudes. Royal Military College of Canada, 80-1, 1980.

Rosen, B, & Jerdee, T.H. Effects of applicant's sex and difficulty of job on evaluations of candidates for managerial positions. Journal of Applied Psychology, 1974a, 59, 511-512.

Rosen, B. & Jerdee, T.H. Sex stereotyping in the executive suite. Harvard Business Review, 1974b, 52, 45-48.

Rosen, B. & Jerdee, T.H. Influence of sex role stereotypes on personnel decisions. Journal of Applied Psychology, 1974c, 59(1), 9-14.

Rosen, B. & Jerdee, T.H. The influence of sex-role stereotypes on the evaluation of male and female supervisory behavior. Journal of Applied Psychology, 1973, 57, 44-48.

Rummel, R.J. Applied Factor Analysis. Evanston: Northwestern University Press, 1970.

Shaw, E.A. Differential impact of negative stereotypes in employee selection. Personnel Psychology, 1972, 25, 333-338.

Simpson, S.P. Self-fulfilling prophecies: A model for the performance appraisal of women in the Canadian Forces. Proceedings of the 22nd Annual Conference of the Military Testing Association, 1980.

Snyder, M., Tanke, E.D., & Berscheid, E. Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. Journal of Personality and Social Psychology, 1977, 35(9), 656-666.

Spence, J.T., Helmreich, R. & Stapp, J.A. S short version of the Attiudes Toward Women Scale (AWS). Bulletin of the Psychonomic Science, 1973, 2, 219-220.

Stow, J.W. The Performance Evaluation Report (Men). Canadian Forces Applied Research Unit, 75-6, 1973.

Terberg, J.R. Women in management: A research review. Journal of Applied Psychology, 1977, 62(6), 647-664.

Terberg, J.R. & Ilgen, D.R. A theoretical approach to sex disrimination in traditionally male occupations. Organizational Behavior and Human Performance, 1975, 13, 352-376.

Yoder, J.D., Rice, R.W., Adams, J., Priest, R.F. Reliability of the Attitudes Toward Women Scale (AWS) and the Personal Attributes Questionnaire (PAQ). Buffalo, NY, SUNY at Buffalo, Dec 1979.

AD P001419

# Psychological Screening for Weapons Use Suitability

Krug, Samuel E., Institute for Personality and Ability Testing, Champaign, Illinois (Chair); Behrens, Gary, Institute for Personality and Ability Testing, Champaign, Illinois; Palese, Robert, Williams, John, CDR, & Winn, Frank, US Coast Guard Support Center, Governors Island, New York.

A pilot study was conducted for US Coast Guard 3rd District to determine the feasibility of objective psychological assessment in evaluating suitability for armed duties assignments. A linear decision model incorporating normal and clincial assessment components was constructed to facilitate the evaluation process. The results of the study are encouraging and point to possible applications beyond the issue of weapons screening.

Discussion will include four separate presentations, each focusing on some aspect of the program concepts. The initial presentation will provide background on the study objectives. The need for a cost-effective screening in relation to weapons management issues will be addressed. In the second presentation, a psychometric approach to the problem of screening will be described. Methods used in the present study will serve as an illustrative example. Results of the study will be reviewed in the third presentation. Evidence for the validity of the screening methods, with emphasis on clinical follow-up in selected individual cases, will be considered. The third presentation will explore some implications of the screening method in the larger context of ongoing personnel concerns. Particular attention will be given to potential benefits in the areas of leadership development, retention and specialty training.

# Psychological Screening for Weapons Use Suitability: A Formal Decision Model

Samuel E. Krug, Ph.D. Director, Test Services
Gary M. Behrens, M.S. Research Associate
Institute for Personality and Ability Testing
Champaign, Illinois

## Summary

A data-based systems approach to the problem of reviewing large numbers of personnel with respect to psychological fitness for weapons use was adopted. Two professionally developed and well-validated tests were administered to 709 USCG 3rd District personnel. These objective tests measured a broad spectrum of personal characteristics, such as emotional stability, anxiety, self-discipline and integrity/control, which have been shown to relate significantly to the ability to exercise appropriate judgment in high stress situations.

A formal Screening Decision Model (SDM) was developed to insure consistent and programmatic evaluation of the psychological data. The construct validity of the model was checked in several ways, including systematic comparison of sample test scores with thousands of profiles maintained in an extensive data base of normal and psychiatric protocols. A set of decision rules was established through combining the clinical judgment of IPAT staff psychologists with the results from a series of statistical analyses. Findings suggested that a small but noteworthy percentage of the personnel screened show characteristics that could significantly impair their ability to render appropriate use-of-force decisions in nonroutine circumstances.

## Scope of the Program

In September, 1980, USCG psychologists, security and readiness personnel met with IPAT psychologists at Champaign, Illinois, to define the major needs of the Coast Guard with respect to a weapons screening program and to help shape the basic dimensions of a workable program.

The population identified for participants in this project consisted of all 3rd District personnel who currently were or could reasonably be assigned to boarding party duty. A roster of 948 such individuals was drawn up by 3rd District Headquarters according to duty stations. Individual sets of test materials were then prepared by IPAT and distributed in appropriate quantities to each unit commander. Written procedures for controlled administration of the program materials were provided as well.

Test materials were received back for 730 individuals. Approximately 3% of these were incomplete and so suitability judgments could be made in 709 cases only. Demographic characteristics of the 709 participants were determined in order to better profile the group as a whole. The mean age of this sample group was 23 years, with a range from 18 to 43. With regard to gender, the sample group was predominantly male (96%). In terms of rank, about 70% of the group could be classified as junior enlisted personnel, 20% as senior enlisted personnel, and 10% as warrent or regular duty officers. Among the enlisted personnel, approximately 28% were in engineering rates, while the remaining 72% were in nonengineering rates.

## The Psychological Battery

Materials used for the screening battery consisted of two widely used and well-validated personality questionnaires, the Sixteen Personality Factor Questionnaire (16PF), Form A, and the Clinical Analysis Questionnaire (CAQ), Part II. Both instruments present objectively worded items in a multiple choice answer format. In all, a total of 331 items were included in the screening battery.

The 16PF is designed to measure key personal qualitites that determine characteristic behavior response patterns. It was first published in 1949 and has been revised five times since, most recently in 1968. It is one of the most widely used inventories in the world and an extensive normative data base has been established for it. Individual scale reliabilities range from .72 to .84.

The CAQ measures pathological behavior tendencies in terms of 12 major clinical syndromes. These include seven separate indicators of depression, in addition to such classical disorders as paranoia, hysteria and schizophrenia. More than 15 years of research and practical application have confirmed the psychometric credibility of this instrument. Individual scale reliabilities range from .67 to .86.

Together, the 16PF - CAQ combination provide reliable information on 28 personality characteristics. Three validity scales are also embedded within these tests. One checks for the tendency to "fake good," another for the tendency to "fake bad," and the third detects random or inattentive response patterns.

For accuracy, all tests were computer scored at IPAT's central office in Champaign, Illinois. If necessary, test scores were corrected for faking tendencies. On the whole, however, faking was not found to be a serious problem in the group of Coast Guard personnel who participated in this project.

## Development of a Screening Decision Model

Reliable and fair assessment of large groups of individuals requires that clear rules be consistently applied. Case-by-case judgments by individual psychologists are likely to produce great inconsistency (see, for example, Meehl, 1954), as well as prove unrealistic and uneconomical. Scientific research over the past 30 years has reliably shown that clinicians are good at determining what information is relevant to a particular decision but not at utilizing that information consistently enough in individual cases.

For these reasons, we decided to establish a formal weighting system that would generate an overall index from the many pieces of information available to us. Technically, this procedure is described as linear model building. The weights selected for the model are those which preserve a high relationship between the overall index score and the target behavior, i.e., psychological suitability for weapons use.

In cases where the target behavior is fairly simple to quantify, statistical procedures can be used to determine an optimal set of weights. But when the target behavior is complex, statistical soultions are not always possible. Instead, it is necessary to rely on existing data bases from which a rational set of weights can be derived. Although the weights selected in this manner may be less than optimal, the predictive utility of the overall approach has been shown to be better than that achieved by individual judges (Dawes, 1979).

For this project we constructed a linear decision model by incorporating some features from an existing one of demonstrated reliability. The development of this earlier model has been described by Krug (1981) in the context of screening utility industry personnel for security access to nuclear generating facilities. The concept has been extended to the problem of reactor operator certification, where the ability to make decisions under pressure plays a crucial role.

We began with this model with which we had had several years of screening experience and with additional information we had availabole from work in the area of law enforcement screening (IPAT, 1981). However, certain refinements were introduced into the model in response to unique Coast Guard needs and concerns. This was made possible through the input of a group of 12 senior officers familiar with use-of-force issues and boarding party actions.

The model actually developed for this project can be most easily understood by reference to five functional clusters among the measurement scales. The clusters, in order of importance, are labeled mental health, emotional maturity, integrity/control, intellectual efficiency and interpersonal relations. In terms of component scales, these clusters may be considered independent, but they are closely analogous in many respects to broad secondary patterns that have been factorially established for the instruments comprising the battery.

Mental health contributes the greatest amount of variance to our model, with 32%. It encompases essentially all of the pathology scales from the CAQ, Part II. The normal personality dimensionsof the 16PF are apparent in the remaining clusters. Emotional maturity, which taps characteristics of stability, self-confidence, tranquility and adaptability is next at 22% of the overall variance. It is followed by integrity/control, measuring self-control and conscientiousness while contributing 19% of the variance. Another 16% of the variance is associated with intellectual efficiency, which is reflected by characteristics of objective realism, practicality, analyticalness and precision. Interpersonal relations, a cluster consisting of

assertiveness, self-sufficiency, openness, venturousness, and enthusiasm, contributes the least amount to the total variance - only 11%.

## Psychometric Evaluation of the Model

The decision model we developed therefore consisted of a linear combination of 26 component dimensions. Two of the CAQ scales, agitation and psychopathic deviation, were dropped because the a priori evidence for their impact on the target behavior was inconclusive. An individual's score on each component dimension was multiplied by a weighting factor proportional to its importance in the overall model. These weighted scores were then added together to yield a single value that we call the Screening Decision Model index, or SDM.

We adjusted the weighting factors so that the SDM would have a range consistent with the 16PF and CAQ scores from which we started. Consequently, the SDM score values can range from 1.0 to 10.0. The expected average is 5.5 and the standard deviation is expected to be 2.0. In general, higher scores are better and lower scores are poorer in terms of psychological suitability for weapons use. On the whole, the Coast Guard sample tested well, with an average for the group of 6.26. About 64% of the cases fell above 5.5 on the index.

The reliability of the SDM index can be calculated by means of a formula given in Guilford (1954, p 393). By this formula, the composite reliability of the SDM was determined to be .91. This is far above average with respect to most personality scales and attests to the inherent stability of the model. Given its standard deviation and reliability, the standard error of measurement for the SDM is expected to be about 0.60 score units.

Several approaches were taken with respect to documenting the validity of the model. A considerable amount of validity evidence bearing on the issues of performance under stressful conditions and general adjustment was initially abstracted from published sources. In effect, this body of literature guided our thinking substantially at the time we originally set out to create a rational method for evaluating psychological fitness regarding weapons management.

We later evaluated the model against the combined 16PF-CAQ clinical data base reported in the respective test handbooks (Cattell et al, 1970; Krug, 1980). In these publications, data has been presented for about 5,000 clinically diagnosed adults. It is compiled as mean profiles for various diagnostic groups. These groups represent a broad spectrum of psychological disabilities, ranging from the confused, disorganized thought of schizophrenia to the destructively antisocial behavior of the criminal. Obviously, they are the kinds of individuals we would like to find on the low end of the SDM index.

We checked this hypothesis by calculating a statistic called the coefficient of congruence between the SDM weighting pattern and the pattern of score deviations for each of 27 diagnosed clinical samples. This statistic is defined as the quotient that results from dividing the sum of the cross-products for two patterns by the square root of the product of their respective sums of squares (Harman, 1976, p 344). In 16 of the samples, mean profile information was available only on the 16PF normal scales but not the additional evidence of the CAQ clinical scales. The associated significance levels of the congruence statistic were adjusted accordingly.

From the perspective of psychological suitability, validity of the model would be demonstrated by finding significant, negative congruences between the two patterns. The relationship between the SDM weighting factors and the pattern of group deviations was, in fact, in the expected direction.

That is, the SDM index tended to emphasize precisely the opposite qualities that characterize clinical populations. The degree of discrepancy was statistically significant in each of the 27 comparisons. Complete details of these analyses are presented in an appendix to this paper.

Thus, the SDM we developed consistently selects for those personality characteristics that promote healthy emotional adjustment. Another way of putting this is to say that the higher an individual's SDM score, the less likely is the possibility that person would act in an uncontrolled or unpredictable fashion. One reasonable conclusion is that the ability to react appropriately under pressure is better in those with higher SDM scores than those with lower scores.

In a third approach to indirect validation of the SDM, we drew over 2,000 individual cases from our existing 16PF-CAQ data base. Of these, 467 men and 457 women were normal adults who had been tested as part of the national standardization of the tests. There were 946 other individuals who were hospitalized or receiving outpatient therapy for some psychiatric disturbance at the time of testing. Of the total, 17% were diagnosed as schizophrenic, 15% as having a personality disorder, 12% as neurotic, 8% as drug dependent, 28% alcoholic. The remainder of the group showed no consistent diagnostic patterns. A third sample of 446 convicted felons was also drawn.

We calculated SDM scores for each of these 2316 individuals and contrasted their scores with the results of the Coast Guard personnel. Again, the clinical and convict populations contain the kinds of individuals we hope the SDM would screen out: maladjusted, uncontrolled, disorganized, sociopathic. The adult normal group provides an independent contrast group for determining base lines. The results are shown in Figure 1.

Scores of 1 or less on the SDM index are found about 25 times more often in a clinical population than in the total Coast Guard sample, and about 4 times more often than in the population as a whole. The situation is quite similar for the convict sample. In general, low scores occur far more often for maladjusted than for normal individuals.

By establishing a cutoff we can translate these data into a direct validity (phi) coefficient. Based on the total sample of 924 normal adults and 946 clinical cases, the correlation between the SDM index and a dichotomous "normal vs. maladjusted" criterion is .33. This is significant beyond conventional statistical levels and about typical in size of validity coefficients found in applied selection research. With respect to the convict sample, the validity coefficient is slightly lower, .26, but still highly significant.

Demographic information was available for the normal adults in our data base and so we also were able to evaluate possibly irrelevant sources of variance in the SDM score. A significant age effect was found in the population at large. Older people tended to have higher SDM scores than young people. Within the project sample specifically, this trend was confirmed, but it was not nearly as dramatic (r= .10). There were no significant findings with respect to sex or race. That is to say, sex and race do not seem to have any impact on the magnitude of the SDM score.

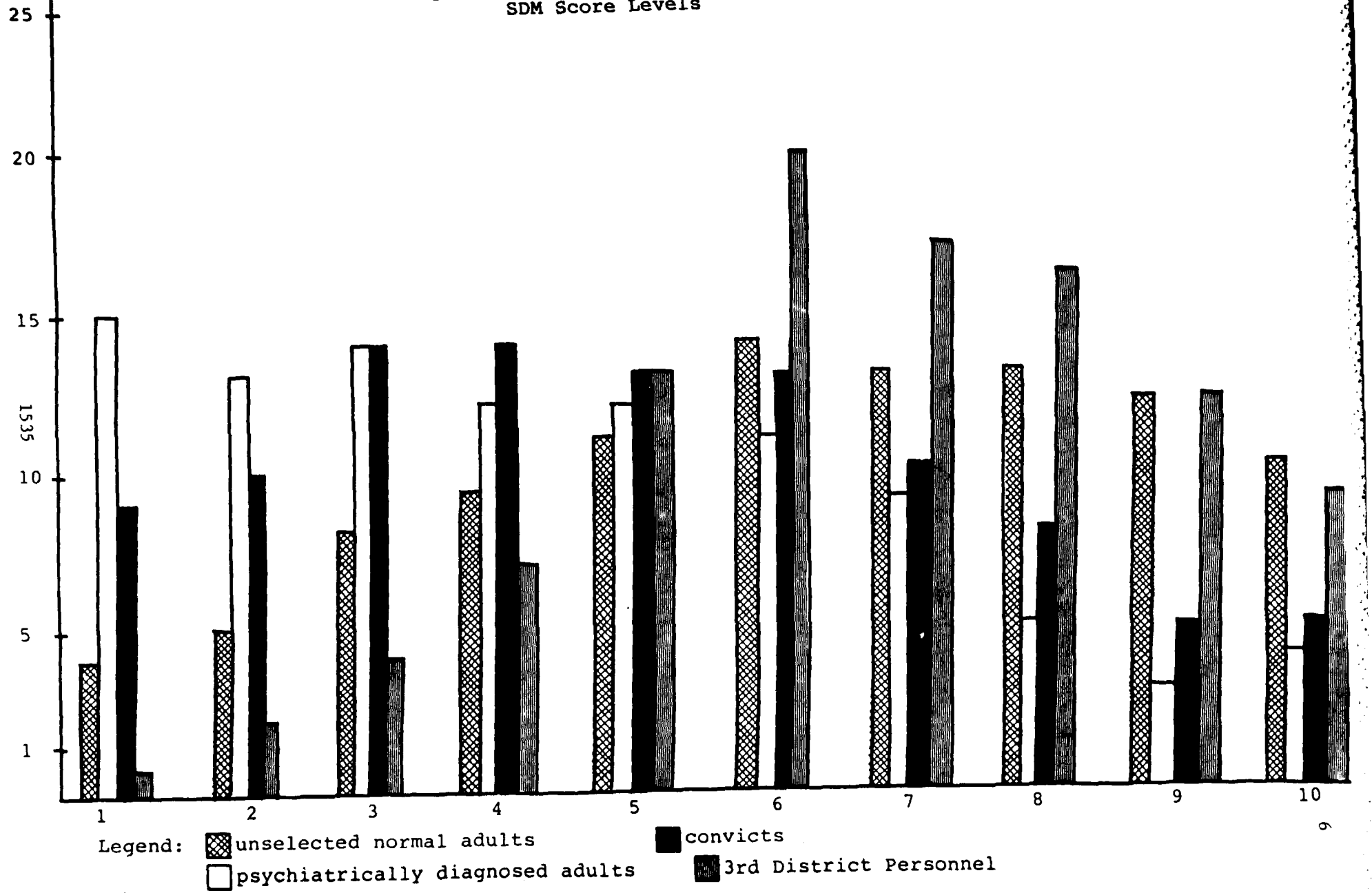
## Results

Having satisfied outselves that the SDM index was both a reliable and valid indicator of suitability for weapons use, we proceeded to render a fitness judgment with respect to each individual in the Coast Guard project sample. In this phase, three qualification levels were distinguished.

Level 1. Test scores of individuals who scored below 3.0 on the SDM index were reviewed on a case-by-case basis by two IPAT psychologists. Working independently, they nevertheless concurred that each profile showed

Figure 1

Percentages (%) of Individuals in Four Defined Populations Scoring at each of 10 SDM Score Levels



sufficient disturbance to warrant a personal interview. These were designated as Level 1, or unacceptable, risk group.

Level 2. We reviewed test results of individuals who scored above 3.0 on the SDM index and felt that some, though not all, of those also warranted a personal interview. It was our conclusion that, above 3.0, some additional signs should be present in the profile itself to substantiate a recommendation for individual follow-up. After studying the cases, we reached the decision that a Level 2, or marginal risk, assignment should be made if the SDM score was between 3.0 and 5.2 (one half standard deviation below the Coast Guard norm) and the individual had score deviations at or below the 15th percentile on more than two of the contributing pathology scales.

Level 3. Individuals who were not otherwise classified by these previous decision rules were assigned a Level 3 risk classification. Their current psychological makeup showed no evidence of factors that would appear to render them unsuitable for duties involving appropriate weapons usage judgments.

On the basis of these decision rules, 47 individuals were assigned a Level 1 classification, 47 individuals a Level 2 rating, and 615 were assigned to Level 3. Figure 2 illustrates the relative proportions of cases falling into each category. Those receiving ratings of Level 1 or 2 were referred for individual follow-up by Coast Guard mental health professionals. The outcomes of that phase of the project are discussed separately.

Earlier, it was noted that the overall Coast Guard performance is better than that of an unselected, normal adult population. Nevertheless, 13.2% of the survey group was referred for further evaluation. Since this was a pilot project, we elected to refer the borderline cases as well as the clearly unacceptable cases. By referring a larger number of "problems" for personal interview, we hoped to have a stronger empirical basis for establishing cutoffs in the event the program were later implemented throughout the service or used as part of recruitment screening procedures.

## References


Cattell, R.B., Eber, H.W., & Tatsuoka, M.M. Handbook for the Sixteen Personality Factor Questionnaire. Champaign, Il: IPAT, 1970.

Dawes, R.M. The robust beauty of improper linear models in decision making. American Psychologist, 1979, 34, 571-582.

Guilford, J.P. Psychometric Methods (2nd edition). New York: McGraw-Hill, 1954.

Harman, H.H. Modern factor analysis (3rd edition). Chicago: The University of Chicago Press, 1976.

IPAT Staff. Manual for the law enforcement assessment and development report. Champaign, Il: IPAT, 1981.

Karson, S., & O'Dell, J.W. A Guide to the clinical use of the 16PF. Champaign, Il: IPAT, 1976.

Krug, S.E. Development of a formal measurement model for security screening in the nuclear power plant environment. Multivariate Experimental Clinical Research, in press.

Krug, S.E. Clinical analysis questionnaire manual. Champaign, Il: IPAT, 1980.

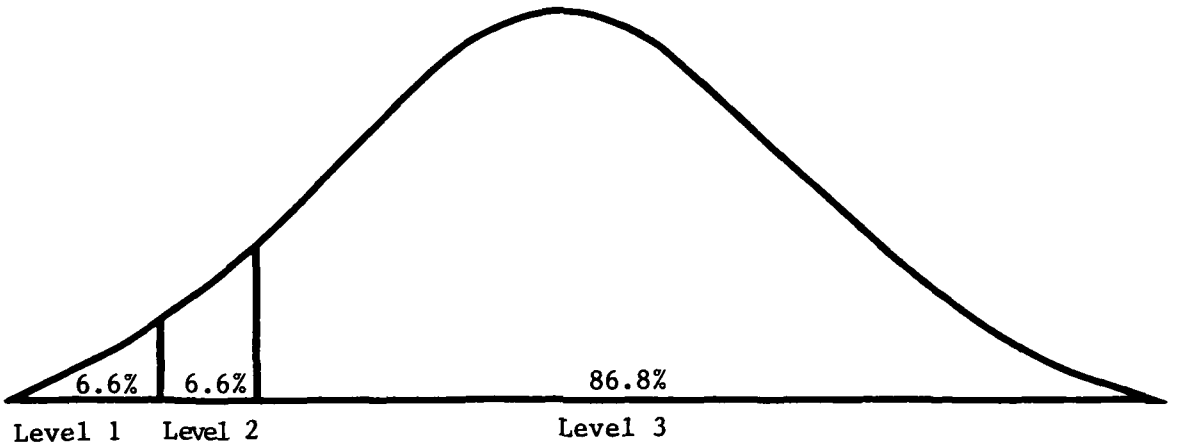
Krug, S.E. Psychological assessment in medicine. Champaign, Il: IPAT, 1977.

Meehl, P.E. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press, 1954.

Sweney, V.A. Expanded definitions of the 16PF. Paper presented at the First International Conference on the 16PF. San Luis Obispo, Ca: March, 1981.

Figure 2

Distribution of Project Participants Among Three Categories of Psychological Suitability




Disposition:

Suitability Level 1: Referred for evaluation
Suitability Level 2: Referred for evaluation
Suitability Level 3: Not referred

## Appendix

### Congruence Between SDM Weighting Factors and Profile Deviations for 27 Diagnostic Syndrome Groups

| Syndrome Description or Diagnostic Group | $r_{cc}$ | Sample Size |
|---|---|---|
| Narcotic Addiction | -82 | 64 |
| Criminals* | -81 | 891 |
| Personality Disorder-Unspecified | -79 | 76 |
| Passive Aggressive Personality | -78 | 42 |
| Functional Psychosis* | -77 | 105 |
| Acute Undifferentiated Schizophrenia* | -76 | 41 |
| Neurotic Depressive Reaction | -75 | 70 |
| Personality Disorder-Antisocial Type | -75 | 34 |
| Chronic Undifferentiated Schizophrenia | -73 | 33 |
| Alcoholism* | -73 | 1019 |
| Sociopathy* | -72 | 28 |
| Manic-depressive Psychosis Depressed Type* | -70 | 53 |
| Neurosis* | -70 | 272 |
| Neurosis | -69 | 26 |
| Suicide Attempters* | -69 | 50 |
| Anxiety Neuroses* | -67 | 80 |
| Suicide Attempters* | -67 | 50 |
| Alcoholism | -67 | 230 |
| Acute Undifferentiated Schizophrenia | -66 | 116 |
| Neurotic Depressive Reaction | -64 | 31 |
| Schizophrenia Schizo-Affective Type* | -64 | 27 |
| Narcotic Addiction* | -64 | 937 |
| Personality Disorder Inadequate Personality Type* | -63 | 54 |
| Personality Disorder Obsessive-compulsive Type* | -63 | 29 |
| Child Abusers | -59 | 78 |
| Personality Disorder Anti-social Type* | -58 | 97 |
| Paranoid Schizophrenia | -57 | 59 |

* For this sample, only 16PF scores were available.

NOTE: Significance levels associated with the distrubution of the congruence coefficient are as follows: 16PF only - .01 level = -.68 and below, .05 level = -.52 to -.67; 16PF and CAQ: .01 level = -.57 and below, .05 level = -.41 to -.56.

# Psychological Screening for Weapons Use: an Introduction

Robert P. Palese, Ph.D.
U.S. Coast Guard Training Center
Governors Island, New York 10004

In Sept. of 1980, Third Coast Guard District personnel contracted with the Institute of Personality and Ability Testing (IPAT) of Champaign Illinois to develop a screening technique to be used for selecting personnel for armed Boarding Party Duty. Preliminary determination resulted in the composition of a three-part, pencil-and-paper screening battery. The materials used for the screening battery consisted of two booklets containing objective items in a multiple-choice answer format.

Booklet one contained two widely and well-validated personality questionnaires, the Sixteen Personality Factor Questionnaire (16PF) and the Clinical Analysis Questionnaire (CAQ).

The 16PF was designed to measure key personality qualities that determine characteristic behavior response patterns. It was first published in 1949 and has been revised five times since, most recently in 1978. The CAQ measures pathological behavior tendencies in terms of 12 major clinical syndromes. More than 15 years of research and practical application have established the psychometric credibility of this instrument. Together, the 16PF-CAQ combination provided reliable information on 28 personality characteristics (factors).

Booklet two contained two experimental instruments. The first instrument was the Life History Questionnaire, developed at Cornell Medical School, which contains clinically oriented scales that measure specific aspects of assertiveness, fear, and worthlessness. The second instrument was a 13-item, achivement-type test constructed by IPAT with the assistance of Coast Guard law enforcement and training personnel. The items represented situations requiring use-of-force decisions that might reasonably be encountered in the line of duty and were based on real incidents. For each item, a series of response options were provided ranging from maximum deadly force to complete withdrawal.

The population identified for participation in this project consisted of all Third Coast Guard District personnel who were, or could reasonably be, assigned to Boarding Party duty. A roster of 948 individuals was drawn up by Third District security and readiness personnel according to duty station. Individual sets of test materials were then prepared by IPAT and distributed in appropriate quantities to each unit commander.

Test materials were received back for 730 individuals. Because 21 had failed to complete all materials, boarding party suitability judgements could be made in 709 cases only.

The average age of this sample group was 23, with a range from 18 to 43. With regard to gender, the sample group was predominately male (96%). In terms of rank, about 70% of the group could be classified as junior enlisted personnel, 20% as senior enlisted personnel , approximately 28% were in engineering rates, while the remaining 72% were in non-engineering rates.

The results of the project provided sufficiently reliable data for the construction of a 26-factor screening decision model (SDM). The SDM was comprised of all 16PF factors plus 10 of the 12 CAQ components. Neither the Life History Questionnaire nor the 13-item, use-of-force decision test contributed significantly to the decision model and so were relegated to a provisional status.

Next, individual test scores on the 26 factors of the SDM were systematically compared by computer with thousands of profiles maintained in IPAT's data base of normal and psychiatric profiles. Results from a series of statistical analyses were combined with the clinical judgements of IPAT staff psychologists to arrive at a set of decision rules. These rules allowed the IPAT staff to distinguish three levels of suitability amoung Third Coast Guard District personnel:

(1) Level 1: those who represented a serious risk of acting inappropriately with respect to use-of-force decisions;

(2) Level 2: those who represented a marginal risk of exercising inappropriate judgements;

(3) Level 3: those whose psychological makeup appeared to represent little or no risk

The application of the decision rules led to the assignment of 47 (6.6%) individuals to Level 1 and an equal number of individuals (47) to level 2. The remainder (615) of the original 709 usable participants were assigned to Level 3.

Complete psychological reports for each individual were generated by IPAT and forwarded to Third Coast Guard District mental health professionals. Those personnel assigned to Level 1 or 2 were referred for individual follow-up evaluation by that staff in order to determine their actual "fitness" for boarding-party duty.

The results of the screening program piloted in the Third Coast Guard District suggest that a simple pencil-and-paper technique may be useful in facilitating personnel utilization decisions. The final judgement on the utility of such a technique awaits the results of the validation process discussed above as well as some future criterion-validity studies whereby an individual's performance as part of a boarding part can be predicted from a knowledge of prior test performance on the 16PF and CAQ.

Perhaps more importantly, the study piloted here could serve as a model for the implementation of a more comprehensive testing program geared toward providing information which could potentially help solve several problems faced

by the Coast Guard in the areas of personnel selection, utilization, and retention.

The Coast Guard's current approach in the preliminary evaluation of enlisted personnel at the boot camp level involves medical certification and ability testing. Recruits certified medically fit for active duty who succeed in the rigorous regimen of boot camp training then use their scores on subtests of the Basic Test Battery to select a rating for which they desire training and hence job placement. The availability of information concerning a recruit's interests and temperaments and the worker traits required by particular occupational specialties would greatly facilitate the decision process as well as promote a more mutually advantageous man-job "fit". Personnel engaged in work for which they posses interest and are educationally and emotionally suited have the potential to derive deeper satisfaction from their work and therefore may be more productive.

Minimally, there is a real potential for a savings in dollars since it has been demonstrated that the intrinsic reinforcement of a particular occupational position and not other expensive perquisites (i.e., shorter work weeks, bonuses, health benefits, etc.) is the most important determinant of a positive reenlistment decision (Wehrenberg and Patterson, 1981)

## References

Wehrenberg, S. and Patterson, K. A study of enlisted attrition in the united states coast guard. Unpublished Manuscript, 1981.

# Psychological Screening for Weapons Use Suitability: The Coast Guard Law Enforcement Mission Requirements

Commander John E. Williams, Chief, Intelligence & Law Enforcement Branch, U.S. Coast Guard Third District
Governors Island, New York

## Summary

Throughout its history, the Coast Guard has been charged with the responsibility of enforcing federal statutes within the coastal boundary waters of teh United States. In recent years, the major emphasis has focused on control of drug smuggling activities as well as preventing illegal entry by foreign nationals. These potentially tense situations present the risk of hostile or uncooperative actions to which a forceful response may be necessary.

In 1979, a policy directive was issued concerning use of force in a boarding party action. Essentially, it required that all boardings be conducted on an armed basis and that all personnel assigned to such duty be properly certified to assure the maximum degree of safety to the public and individual members of the Coast Guard. This included verification by the field unit commander as to an individual's judgment and stability. Given the constraints of manning and training at all isolated stations, the requirements of the policy, and in particular the need for personal evaluation by a superior officer, posed a serious challenge to field district command staffs.

The Coast Guard has been involved in law enforcement since its conception. The Coast Guard was established 4 August 1790 for the purpose of stopping the lucrative smuggling trade that was flourishing at that time. Although the Coast Guard has assumed many other roles and missions over the years, law enforcement has remained a primary mission. At times in our history the law enforcement mission has received much more emphasis then at others; at some points it was almost non-existent.

After curbing the smuggling business which caused our conception, the law enforcement mission fell to low priority and remained so until the Prohibition Era. This era has often been referred to as the "hay day" of Coast Guard law enforcement due to the major role the Coast Guard played in the interdiction effort. With the repeal of Prohibition, the law enforcement mission again fell to one of minor importance. During the period after Prohibition until the early seventies Coast Guard Law Enforcement was primarily oriented towards Boating Safety which consists of boarding and checking pleasure craft for compliance with federal law and regulations. These boardings were conducted unarmed.

In the early seventies, with the increased drug use in the country, particularly marijuana, the Coast Guard was again called upon to interdict the flow of contraband from sea. This twist to the law enforcement mission required the occasional arming of boarding parties. The Fishery Conservation and Management Act of 1976 added yet another twist to the law enforcement mission. At the same time the drug interdiction effort continued to expand. As a result, the Commandant of the Coast Guard established a policy where-by a commanding officer could arm his boarding parties when he deemed necessary. This policy resulted in what the public perceived to be unequal or discriminatory law enforcement; e.g., "Why did you board my vessel armed and not board the other persons vessel armed?" To reduce any feeling of discriminatory law enforcement, yet deal with the realities of boarding situations, teh decision was made to arm all boarding parties no matter the reason for each boarding. However, the net result was that few Coast Guard units were able to meet the requirement. Many units did not have small arms on board nor did they have small arms trained personnel on board to conduct the boardings. A crash training program was immediately undertaken to train Coast Guard personnel in the use of small arms. Soon a standard service-wide policy on small arms training and qualifications was issued by the Commandant. To be certified qualified to carry a weapon during an armed boarding, an individual must (1) have fired a qualifying score with the type of weapon carried, (2) have successfully completed the "Shoot, Don't Shoot" course of fire (this is a course of fire where-by an individual is shown a series of situations on a screen and must make the decision whether ornot to shoot), (3) have had instruction in the Commandant's policy on the use-of-force, and (4) his Commanding Officer must have determined that he possesses the temperament and judgment necessary to make reasonable and correct use-of-force decisions under pressure. In order to retain his weapons qualification, the individual must be requalified annually.

The Coast Guard, like the other services, is plagued with personnel shortage and turnover problems; people are constantly changing jobs. Presently, tours are two to three years in length. Most of teh more junior enlisted personnel spend far less time than that at their units. This presents a major problem not only in getting personnel qualified and requalified, but many Commanding Officers feel uncomfortable about attesting to the temperament and judgment of an individual they barely know and have not observed under stressful conditions. Due to personnel manning levels about the units, the Commanding Officer cannot afford the luxury of waiting until he has had sufficient time to observe the individual before making this all important decision. If the Commanding Officer is to accomplish his mission, he must have another fully qualified boarding officer to replace the boarding officer being transferred immediately, not months down the road. The Commanding Officer thus finds himself in the proverbial position of "between a rock and a hard place."

In order to assist the Commanding Officer in making this decision, the Third Coast Guard District embarked on a project to develop a written psychological test that would provide the Commanding Officer with sufficient information about the individual's psychological profile so that he can confidently make important decisions as to the indivdiual's suitability for weapons use immediately upon the individuals reporting aboard that unit.

After a review of the commercially available psychological tests, it was determined that none were suitable for Coast Guard use in their current form. Therefore, it was decided to develop a psychological test custom tailored to the Coast Guard's unique Maritime Law Enforcement Mission.





# Psychological Screening for Weapons Use: A Clinical Validation of Measures

by

LCDR Frank J. Winn, Jr., Ph.D.

AD P001420

Every year recruits with prior psychiatric histories are allowed to enlist or are enlisted without the recruiters being aware of their condition. A recruiter may suspect a problem but, under laws dealing with medical record confidentiality, it cannot be pursued. In a few cases recruits are instructed by recruiters not to say anything about their psychiatric problems.

Coast Guard regulations state that an individual with a psychotic condition cannot be enlisted. For active duty personnel a psychotic episode or neurosis is grounds for evaluation by a Central Physical Evaluation Board with recommendations for medical retirement and referral to a Veterans Administration (VA) facility.

The cost to the government of enlisting an individual suffering from a VA ratable disability is staggering. The average life expectancy of a male is approximately 72 years. Assuming that a ratable disability is identified early in the first enlistment, that the individual is earning approximately $600 per month, and that he/she is awarded the minimum percentage required to receive a monthly pension then it can be expected that the individual will be awarded in excess of $108,000 over their lifespan. This is the amount received from the service and does not include any adjustments or compensation received directly from the VA.

Many individuals with problems are identified and released from active duty while in basic training. Others manage to make it through their basic training, especially if they react well to a structured environment. Many, but not all, of the individuals who do make it through basic training subsequently decompensate when they reach the relatively unstructured environment of the Coast Guard small boat station. It is important for the Coast Guard to identify not only those individuals with overt psychiatric problems but those whose problems are not yet overtly manifested.

Winn, Frank J., LCDR
Psychological Screening for Weapons Use

The screening of personnel with psychiatric problems became crucial with the Coast Guards' decision to arm all boarding parties. Until a few years ago the majority of Coast Guard boardings were of the search and rescue variety. In the last several years, however, the number of law enforcement boardings has increased dramatically. It was felt that it could only be a matter of time before a patrol craft was fired on since siezure of a boat ladened with drugs could cost the owners and/or crewmen millions of dollars and prison terms.

There are three possible outcomes to an armed boarding conducted by Coast Guard personnel. The first outcome is that an unstable sailor could decompensate and wound or kill an individual during a routine boarding. A second possible outcome of an armed boarding is that an unstable sailor could freeze in a tense situation that required the legitimate use of force. This outcome could lead to the injury or death of a shipmate. The final outcome of an armed boarding is that each sailor will only use his weapon under the appropriate conditions and at the appropriate times. Fortunately, this has been the case to date. Given the ramifications of the first two outcomes the Commander, Third Coast Guard District decided to explore the feasibility of using a battery of tests to aide unit commanders in the selection of personnel to carry weapons. The Institute for Personality and Ability Testing (IPAT) was subsequently selected to construct an instrument that would meet the needs of the Coast Guard and conduct a pilot study on personnel within the Third Coast Guard District to determine its effectiveness. IPAT was selected because of their work with populations involved in high stress industries, their work with firearms management, and their familiarity with military populations.

The construction of the tests, their validities and reliabilities, the administration of the test and the generation of protocols are discussed in companion papers. Basically, test respondents were divided into three risk categories; Level 1, those at serious risk of acting inappropriately on use-of-force decisions; Level 2, those at marginal risk of exercising inappropriate judgement; and Level 3, those who represent little or no risk. This report deals with the results of the clinical interviews on those Level 1 and 2 personnel identified by IPAT to be at risk to carry weapons. The purpose of the clinical interviews was to assess the accuracy of the tests in identifying potentially unstable personnel.

Subjects: The Third Coast Guard Distrct identified 948 individuals who were or could be assigned to boarding parties. From the original roster only 629, or 66% of those originally identified, could be tested. IPAT subsequently referred the names of 94 individuals they identified as being at risk to use weapons. Forty-seven individuals were from Level 1 (serious risk) and 47 were from Level 2 (marginal risk). An effort was made to contact all of those individuals identified by IPAT to be at risk. However, approximately half of the group had been transferred to other Coast Guard Districts or had been released from active duty. Individual interviews were conducted with 27 subjects (58%) from Level 1 and 26 subjects (55%) from Level 2.

Seven individuals were discarded from all consideration. Their data was not included in any analyses nor were they considered in the determination of the above sample size. Two of the seven were from the same station. These individuals were informed that they would not be granted liberty, after a weekend of icebreaking, until they had completed the tests. As might be expected, they answered the tests in a random fashion. In addition, two male caucasions and one male Phillipino were found to have difficulty in reading comprehension. Since three of the personnel

had difficulty understanding the test questions a decision rule was adopted to discard those individuals who scored between sten 1 and sten 2 on the factor that IPAT had labelled academic achievement. Of the seven who were discarded one individuals SDM sten was one, two individuals had SDM stens of two, three had SDM stens of three and one individuals SDM sten was four. The individuals who were discarded ranged from E-1 to E-7 in pay grade.

Procedure: Shortly after the test scores were returned from IPAT the various units were contacted. In the initial contact it was determined if the personnel identified as being at risk were still assigned to the unit. If the individuals were no longer assigned to the unit an effort was made to determine if the individual was released from active duty (RELAD), transferred out of the Third Coast Guard District or transferred within the district.

Most of the individuals not available for interview had been transferred out of the Third Coast Guard District or had reached the end of their normal enlistment. Because of restrictions imposed by the freedom of information act it was not possible to determine the reason for the release of all the individuals who had not reached the end of their normal enlistment. It was learned, however, that one Petty Officer was discharged for incompetency, one for disciplinary problems and three for the good of the service.

Interviews with those "at risk" sailors still assigned to the Third Coast Guard District were arranged through their group commanders. All hands, from the Group Commanders to the sailors being interviewed, were told that that those individuals to be interviewed were randomly selected by IPAT. They were instructed that the interviews were being conducted to provide a clinical validation for the computer decision, whatever that decision might be. After the interviews- except for several isolated cases where it was thought that an individual posed an immediate risk to himself or others - the people interviewed and the commands were not informed of the results.

Results and Discussion: The IPAT battery contains three validity scales. Specifically, the 16 PF and the Clinical Analysis Questionaire can be scored for an individuals tendency to fake good, to fake bad, or to provide random or inattentive responses. The clinical interviews did not uncover evidence to doubt the reliability of the fake good or fake bad scales. The interviews did provide evidence that the validity scale used to identify random and inattentive responses was not as sensitive as predicted.

Of the 27 Level 1 individuals identified by IPAT to be at serious risk we concurred in 23 cases following personal interview. Of the 26 Level 2 individuals identified by IPAT to be at minimal risk we concurred in 15 cases. The number of false positives identified by IPAT in Level 2 was to be expected because the IPAT model was conservative and allowed for a greater number of Type I errors. It is expected that as the test is put into wider use and the data base increases the number of Type I errors will decrease.

The data obtained from the clinical interviews were analyzed to see how well they fit the predictions provided by IPAT. A $X^2$ analyses was conducted to determine if the number of Level 1 and Level 2 individuals found not fit to carry weapons, through clinical interviews, differed from that expected. The number of Level 1s and 2s that were found not fit to carry firearms did not differ from what was expected ($X^2$=1.684, $p>0.05$). The distributions of those found fit and not fit to carry firearms, across 5 SDM sten categories, was analyzed. The results of the analysis tend to indicate that the distributions were drawn from the same population ($X^2$=6.658, $p>0.05$). A third analysis was conducted to determine if the distribution of those found not fit to bear arms was different from that identified by IPAT. An analysis of the distribution across the 5 SDM sten categories indicated that it did not appear to be significantly different from that identifid by IPAT ($X^2$=7.65,

$p > 0.05$). A final analysis was conducted to determine if the distribution of those found fit to carry firearms, after clinical interview, was the same or different from the distribution identified by IPAT. An analysis of the frequency distribution across the 5 SDM sten categories indicated that the distribution was different from that identified by IPAT ($X^2 = 9.96$, $p < 0.05$). The reason for the possible differences in the distributions is that the IPAT model overestimated the pathology because of their conservative decision rule. The overestimation of the pathology across the 5 SDM sten categories resulted in an underestimation of the number capable of handling firearms appropriately and, because of the small sample sample size of healthy individuals within the sample, the shape of the distribution.

In summary, it would appear from the analyses of the clinical data that IPATs' battery of tests can be used as a valid indicator to determine who is at risk to carry firearms on a boarding party.

Panel: Analysis of the Impact of the Organizational Effectiveness (OE) Program of the Army

Chair: Laurel W. Oliver, US Army Research Institute

Presenters: U. S. James, Arthur Young and Company; L. W. Oliver, US Army Research Institute; M. D. McCorcle, Southern Methodist University; J. R. Mietus, US Army Research Institute

AD P001421

The US Army Research Institute for the Behavioral and Social Sciences (ARI) has been conducting research on the impact of the Army's Organizational Effectiveness (OE) program. The OE program provides assistance to commanders by internal consultants who have been trained in the utilization of management and behavioral science skills and techniques to improve combat effectiveness. The model underlying the collecting of data to assess the impact of the OE program is presented. The methodology of the data collection is outlined, and two illustrative case studies are described. The problems of measuring change in Army organizations are further delineated by the presentation of a sociotechnical systems intervention in an Army organization in Germany. The attempts to document change in an organization with high turnover are described, and implications for future field research are noted.

# ASSESSING THE IMPACT OF THE ARMY'S ORGANIZATIONAL EFFECTIVENESS (OE) PROGRAM: MODEL, METHODOLOGY, AND ILLUSTRATIVE CASES

AD P001421

U. S. James

Arthur Young and Company

Laurel W. Oliver

US Army Research Institute for the Behavioral and Social Sciences

Mitchell D. McCorcle

Southern Methodist University

Paper presented at the meeting of the Military Testing Association

Arlington, Virginia, October 1981

# ASSESSING THE IMPACT OF THE ARMY'S ORGANIZATIONAL EFFECTIVENESS (OE) PROGRAM: MODEL, METHODOLOGY, AND ILLUSTRATIVE CASES

## Introduction

The Army's Organizational Effectiveness (OE) program uses behavioral science technology to improve the effectiveness of Army organizations. In the civilian community, these management and behavioral science skills and techniques are known as Organization Development, or OD. In the Army, OE is the application of selected OD methods in a military environment. The objective of the OE program is to provide assistance to commanders for improving mission performance and increasing combat readiness. This assistance is provided by consultants--Organizational Effectiveness Staff Officers (OESOs) and Organizational Effectiveness Noncommissioned Officers (OENCOs) who have been trained in a 16-week course at the Organizational Effectiveness Center and School (OEC&S) at Fort Ord, California.

The Army Research Institute (ARI) has been conducting research to assess the impact of the OE program. The purpose of this paper is to discuss that research. Specifically, we offer a conceptual model of organizational change, describe the methodology of the research, and present two illustrative case studies.

## Model of Organizational Change

An organization, or any social system, is always in a state of change. The type of change this model addresses is planned change that requires the support of a User's (commander's) subordinates for implementation or acceptance of the change. This model is concerned with those planned changes which are carried out as a part of the Army's Organizational Effectiveness (OE) efforts.

There is some confusion about the use of the term "Organizational Effectiveness." Unless it is used in conjunction with its historical (and more broadly accepted) predecessors such as organizational development and process consultation, the OE term can be misused. Thus, we are addressing only those planned changes which require subordinate commitment to obtain improved organizational effectiveness.

The conceptual model which is emerging from our research on organizational change is composed of several components: actors linked by their roles to a set of processes which lead to a hierarchy of outcomes.

### Actors

There are three types of actors:

1. The User or client--the person who is in charge of the organization (the commander).

2. The OESO (Organizational Effectiveness Staff Officer) or OE consultant.

3. User subordinates--the individuals in the client organization below the user. More specifically, those individuals who are involved in an intervention or change project because their future support is required for the change to work.

Processes

There are three types of critical processes which must be managed during the change. These processes include applying a theory of practice, supplying structure, and insuring diffusion of information. While there may be user responsibilities related to the processes, the consultant or OESO is primarily responsible for ensuring that these processes occur.

Applying practice theory. The OE consultant must have a set of alternative strategies through which a change process can be executed. The overall strategy is generally based on an action research model and, depending on User needs, may emphasize certain aspects of this or another model more than others. It is of great importance that an outcomes orientation be a part of this practice. The OE consultant furnishes the practice theory concepts through which the User can develop a desired future progression. As the change process progresses, the plan can be altered in any dimension. However, the plan must be there for the User to know where to proceed and where he/she has been. Thus, intended outcomes for each step of the strategy are defined and assessed in conjunction with the next step. Desired organizational outcomes are clearly identified and defined in measurable terms by the User and User subordinates as early in the change process as possible.

Supplying structure. The consultant clarifies the User's intended outcomes, organizes them in a meaningful manner, and provides any needed training in order to eliminate unnecessary confusion and obstacles as the intervention progresses. The role of the OE consultant is in this case analogous to that of a construction engineer who must build a bridge between the organization's present condition and the User's intended outcomes for that organization.

Insuring diffusion of information. Finally, the OE consultant must become a communications engineer, developing and energizing systems for exchanging information about the OE operation. This information exchange progresses from areas of greater concentration of information to areas of lesser concentration of information and back again.

Roles

The primary means by which the actors are connected to the process is through their roles in the change process.

User's role. The role of the User is to seek to understand the data and practice theory strategy provided through the consultant in order that future actions can be taken to remedy problem areas. This requirement necessitates

significant commitment since frequently the User is part of the identified problem. The User must also be willing to involve subordinates both in finding a solution and in the implementation process for those solutions which require their support.

OE consultant's role. The consultant's role is one of considerable complexity. It includes three major responsibilities:

1. Assisting the User in choosing and defining issues/problems suitable for an OE operation. The OE consultant must be able to clarify the potential benefits and risks involved in the use of OE methodology.

2. Actively integrating both User and subordinate needs in a way which uses valid information as a basis for all activity.

3. Providing an appropriate and flexible practice theory strategy which stresses outcomes, is supported by necessary structure, and is diffused throughout those parts of the organization affected by the intended change.

Role of User subc 'inates. The subordinates' role is reactive at first. Their role becomes more active as individuals are given opportunities to influence the way in which the organization functions. As people are given these opportunities, they develop expectations about the future return on their contribution. The many cycles of contributions and returns are referred to as psycho-economic transactions. Throughout an operation, these transactions have an important cumulative effect. For the greatest subordinate commitment to occur, individuals must experience returns on their contributions which meet or exceed their expectations. If their expectations are met and they are permitted to assist in finding solutions to important organizational issues/problems using valid information with the freedom to make what they believe are their best choices, it is very likely that they will be committed to the resultant solutions.

## Outcomes

The model depicts a causal flow which implies that improved organizational effectiveness resulting from an OE operation is based on innovation and change to which those affected are committed. The commitment occurs because the OE practice-theory strategy met subordinate expectations through positive psycho-economic transactions, involved an opportunity to influence important organizational matters, and resulted in solutions which were based on valid information and permitted the use of free choice.

## Mediating Factors

For an intervention to be successful, all of the processes described above must be present. In addition, there are factors which mediate the relative success of the change process: a need for change within the organization, a change that is within the control of the unit involved, a goal orientation on the part of the actors, and a supportive environment.

Need for change. The success of the change effort will be greater to the extent that there is a legitimate need for change within the organization. The success of the change will be further enhanced if members of the organization have perceptions that a change is needed.

Change within control of organization. Unless the desired change lies within the control of the organizational unit, the success of the change is problematical. While it is possible that a change might be effected, the probability of success is necessarily much lower.

Goal orientation. A goal orientation on the part of the actors will increase the probability of a successful change.

Supportive environment. The OE process must take place in an environment which is supportive of change. If the environment proves to be nonsupportive, the probability of successful change is diminished.

## Method

### Design

This is a study to produce grounded theory (Glaser & Strauss, 1967) through the collection of a sizable number of OE operation cases obtained throughout the Army. Each cell of the Case Selection Matrix (Figure 2) was to have been filled with four cases. If feasible, two cases were to have had successful outcomes, and two less successful outcomes. The successful and less successful cases in each cell were to have been compared to determine those differences which led to improved outcomes in an OE operation of that cell type. The data were to be collected through structured interviews and from relevant and easily accessible organizational records. The resultant information was then to be coded using an extensive codable variable scheme (Dunne & Swierczek, 1977). A case report describing the case details and outcomes was to be prepared, and the cases were to be assessed using a grounded hypothesis emerging from the data collection. The resultant revised and refined grounded theory was to address the goals of the project:

1. To determine those situational variables which appear to affect the outcomes of an OE operation in a significant manner.

2. To identify a set of replicable OE consultant actions which were highly correlated with successful operations, regardless of organization type.

### Sample

The desired sample size was 48 OE operations with more cases collected from the larger major commands (MACOMs)--Forces Command (FORSCOM), Training and Doctrine Command (TRADOC), and Department of the Army Readiness Command

(DARCOM). Thirty-five cases have been collected and will be used to prepare the findings of the study. The larger MACOMs--FORSCOM, TRADOC, DARCOM, and the United States Army in Europe (USAREUR)--are well represented.

Selection of individual cases in the field was determined primarily by case availability and the access permitted by MACOMs to field sites. In general, our field units were cleared by the MACOM OE Office, and we were given OE offices and OESOs to contact. Upon contact, these OESOs were provided with desired case characteristics. We preferred operations which:

- Had been completed long enough for some effects to have taken place.
- Were not transition workshops or some sort of abbreviated application but somewhat representative of action research based operations.
- Occurred in larger organizations and related to the organization's technical functions.
- Involved a User and an OESO who would agree to be interviewed.
- Finally, were representative of both successful and less successful operations.

FORSCOM policy restricted data collection to designated locations and time periods. Thus it was necessary to collect six cases at a single site within a one-week period. This constraint in many instances precluded the selection of the most desirable research cases. Some MACOM OE offices made fairly obvious attempts to provide us only with their "winners." At this point, however, it does not appear that these constraints made a noticeable difference in overall operation quality or outcomes in the sample.

From a purely statistical perspective, the resultant sample is probably not representative of the entire Army. From a practical view, it appears to us that our sample is representative of OE methodology being employed in the field. As can be seen in Figure 2, the cases tend to cluster in Organization Types I and II (smaller and less complex) and Intervention Types A and B (simpler, shorter-range impact), as we had anticipated in the original research plan. In those cells, it appears that we will have enough data to meet the objectives of our study, and attempts will be made to generalize from the more limited number of cases in the remaining cells.

## Measurement of Variables

The instruments used draw upon the multi-variable coding scheme of Dunn & Swierczek (1977). This instrument consists of approximately 1600 variables which describe behavioral or perceptual elements which can be discretely assessed when information is obtained from the User,

OE consultant, and subordinate individuals involved in an operation. The structured interviews with both User and OE consultant are driven by the coding scheme and function as a crosscheck on information provided by these two key participants.

The coding scheme and interviews cover all aspects of the OE operation from entry through evaluation. While some portions of the coding scheme cannot completely represent the great number of potential varieties of OE applications, the scheme does appear to incorporate most significant elements.

It should be emphasized that when interviews are conducted, the interviewers are careful not to skip or ignore excursions in the operation's process. All actions in the operation are tracked to conclusion in the interview process to the depth permitted by the interviewee. While there are no empirical data for the actual coding scheme being used, we believe that it is a reasonably reliable instrument because it describes small, verifiable behavioral events or outcomes.

## Data Collection

Separate structured interviews with the User and OE consultant were generally conducted and tape recorded by persons with previous extensive military OE experience. Following the User interview, at least one focus group interview with user subordinates was conducted. The group normally consisted of 5 to 12 persons who had a direct relationship with the operation's intended outcomes and who were present in the command prior to and throughout the operation. The groups were generally horizontal or diagonally structured to avoid chain of command relationship conflicts and encourage openness.

The focus group interview provided information about those outcomes perceived by the subordinates -- the extent to which the subordinates attributed the outcomes to the OE operation, the importance they assigned to the changes, and the potential location of recorded information substantiating the perceived changes. When such locations were identified, an attempt was made to collect hard data.

Cases were then coded, and a summary of the OE methodology and findings written. Outcomes information was organized using Kirkpatrick's taxonomy of reaction, behavior, and hard outcome measures (1967) for both intended and unintended results.

## Analysis

The cases are currently being analyzed. The coded variables for each OE operation are being entered into a computer, and statistical procedures will be employed to determine the relationship of the coded demographic and behavioral variables (or groups of variables) to evaluations of success. The written case reports are being reviewed along with interview notes in order to assess these data in relation to the grounded model presented in this paper.

At this point in the study, the following hypotheses are emerging:

- OE operations which purposefully concentrate on both the social organization and the technical aspects of an organization accomplishing its mission are more likely to bring about tangible results.

- When the process outcome of generating commitment is not understood/adhered to by an OE consultant or User, subordinates frequently adopt a negative perception about the purpose and use of OE in the Army.

- The process outcomes occurring during an OE operation, and the final outcomes of the operation are largely determined by the goal orientations of the User, OE consultant, and--to a lesser degree--subordinates.

- The success of an operation, regardless of organizational type, first appears to be based on the consultant's ability to respond to the organization's needs with appropriate practice-theory strategy which (1) is results-oriented, (2) insures that there is appropriate structure, and (3) permits the diffusion of appropriate resultant process information to those expecting it.

- Even though substantial, desirable outcomes result from the operation, organizational members below the User will generally view the operation as a failure if the process is conducted so that subordinate psycho-economic transactions are negative. Consequently, to avoid damaging the OE Program's image and to accomplish outcomes which result in subordinate commitment, OE applications must be chosen and conducted with great care.

## Illustrative Cases

### Description of Cases

Two cases, one very successful and one less successful, are briefly described below using the grounded theory described in this paper. Ratings of structure and diffusion are on a five-point scale: Very High, High, Medium, Low, Very Low. Comments concerning the combat support battalion are on the left, those pertaining to the combat battalion are on the right.

ORGANIZATION TYPE

| Combat Support Battalion | Combat Battalion |
|---|---|

SIZE

| 185 Persons | 735 Persons |
|---|---|

## PURPOSE OF USING OE

Sincere attempt to improve battalion wherever possible. Outcomes not clear.

General assessment for battalion commander who just took over the unit. Recent past of battalion indicates high level of turbulence in leadership and problems within unit.

## DATA COLLECTION

GOQ[1] - 75% of organization surveyed. Individual and Group interviews - 75% of organization covered. Extensive field observation and successful attempt to establish empathetic relationship with User subordinates in field.

Structure: High
Diffusion: High

Individual interviews with officers Unstructured, open-ended questions used.

Structure: Low
Diffusion: Low

## FEEDBACK

Individual session for User

Individual session for officers with User.

Individual session for Senior NCOs.

Additional sessions offered to any officer/NCO desiring individual feedback

Great care taken in display and conduct of these sessions.

Structure: High
Diffusion: Very High

Individual feedback given to User

Feedback essentially "data dump" of interview comments.

Structure: Low
Diffusion: Very Low

[1]The General Organizational Questionnaire (GOQ) is a military adaptation of the Survey of Organizations (Taylor & Bowers, 1972).

## PLANNING

User conducted own planning process with his officers and NCOs. OE consultants were not involved. Planning occurred immediately after feedback sessions.

Structure: Unknown
Diffusion: Very High

Battalion problems increased four months after User feedback. User asked for assistance. OE consultant offered to provide one-half day issue identification and problem-solving workshop. OE consultant designed unstructured process for two groups: staff group and command group.

Structure: Low
Diffusion: Very Low

## IMPLEMENTATION ACTIVITY

User, with officers and NCOs, developed action plan to address nine issues identified in feedback/planning processes. A task team was established to ensure that momentum carried over after User's imminent transfer.

Structure: Very High
Diffusion: Very High

A one-half day workshop for staff and command officers was conducted. Command group refused to do tasks. Had OE consultant intercede on their behalf with User. At confrontation meeting between command group and User, User vowed to increase meetings with command group and staff and change his behavior to reduce crisis management within unit. Workshop employed had very little structure.

Structure: Very Low
Diffusion: Very Low

## OUTCOMES

Significant behavioral and hard outcome results achieved. These included accomplishment of all nine intended outcomes of the operation. While these intended outcomes were never described in measurable terms by the User prior to implementation activity, they did result in several significant positive changes including the following in IG reports:

- Reduction of personnel complaints 600%
- Reduction in deficiencies requiring report 438%
- Increase in laudatory comments 600%
- Reduction of motor pool deficiencies 311%

Additional measurable changes included:

- Special battalion training qualification improvement 13%
- Equipment readiness improvement 71%
- SQT improvement
  - Verification 48%
  - Certification 30%

In all instances, changes were attributed to the OE operation by command members. Further, the changes, taken as a whole, were considered to be very important to the primary mission of the Battalion.

Some moderate behavioral outcomes achieved. It was generally perceived by the officers in the battalion that communications improved among the battalion commander staff and the command group. The initiation of the improvement attributed to a minor extent to the half-day workshop. However, change was not viewed as being very important because underlying problems of crisis management still remained.

## Other Considerations

- User was near the end of his command tour in a very small organization with a combat support role.
- Original User transferred prior to the implementation of action plans. New battalion commander did not know much about OE operation--only that those subordinates concerned were very committed to accomplishment of desired changes.
- User was at beginning of his command tour in a very large organization with a combat role.

General Comments on Cases

The combat battalion appeared to be affected more than the combat support battalion by outside pressures. These pressures were caused by the response of the commanding general of the division to the battalion's poor performance and the requirement for the battalion to prepare for the Ready Deployment Force. OE was perceived by both the User and his subordinates as a "nice-to-do" activity rather than as a way of solving pressing problems.

Both battalions indicated that there was ample reason for change and that most of the desired changes were within the control of the organization. The Users of both battalions indicated that the environment did support the use of OE. An assessment of the OE consultant and User goal orientation, as well as a summary of their ability to provide structuring and diffusion, is provided below:

| Goal Orientation | | Structuring | Diffusion |
|---|---|---|---|
| | Combat Support Battalion | | |
| OE Consultant | Moderate | High | Very High |
| User | High | | |
| | Combat Battalion | | |
| OE Consultant | Low | Very Low | Very Low |
| User | Moderate | | |

At the moment, the above ratings are based on a subjective assessment of the operations described. The analysis of the coded variable scheme will provide greater objectivity, although the general conclusions are expected to remain essentially the same.

## Discussion

The research approach described here represents an attempt to understand the organizational change in the Army by sampling a large number of cases across a range of OE interventions and organization types. An effort will also be made to correlate characteristics of the cases with the degrees of success or failure which are attributed to an OE operation. From this study, a preliminary grounded model of successful organizational change has been developed. Further analysis will provide opportunities to test that model. Additionally, findings will provide information which can be used to guide OE program policy and the training of OE consultants.

The study is limited in a number of ways. These limitations include:

- Lack of access to combat unit cases in USAREUR.

- Being forced to collect six cases in one week at times and at installations designated by FORSCOM. This constraint probably reduced the quality of some of our data collection methods and resulted in the collection of cases of lesser value to the study.

- Being forced to rely on what happens to be available in each case. As noted in the Interim Report for this project, even though evaluation is emphasized in the OE curriculum, there are a great many reasons why most OESOs ignore this activity. Our data collection substantiated this situation. Of all the cases we collected, we found only one which had been conscientiously evaluated. In many cases, neither the OESO or the User knew what outcomes had resulted from their operation.

- The general lack of an outcome orientation in most OESOs and Users imposes the most serious limitation on the study. If those involved in an operation fail to specify the outcomes they desire, it becomes a matter of conjecture when anyone attempts to link causally the OE process with results. The conjectures are further weakened when they are based on individual historical reflections which may often be inaccurate. Consequently, there are instances in which we could have been misled. It is also possible that in some cases, there has been more positive change than the User of subordinates was aware of or willing to attribute to OE. Unfortunately, this is a "Catch 22" situation. Even if we had discovered these elusive changes ourselves, it is likely that those involved would not have agreed that OE caused them. The long-term solution to this problem is a sincere attempt on the part of OESOs and Users to adopt an outcomes-oriented or goal-seeking change strategy so that they know "where they are, where they want to go, and when they get there."

## REFERENCES


Dunn, W. N. & Swierczek, F. W. Planned organizational change--toward a grounded theory. Journal of Applied Behavioral Science, 1977, 13, 135-147.

Glaser, B. G. & Strauss, A. L. The discovery of grounded theory: strategies for qualitative research. Chicago: Alding, 1967.

Kirkpatrick, D. L. Evaluation of Training. En R. L. Craig & L. R. Bittel (Eds.) Training and development handbook. New York: McGraw-Hill, 1967. 1967.

Taylor, J. C. & Bowers, D. G. Survey of organizations. Ann Arbor, Mich.: Institute for Social Research, 1972.

# A MODEL OF THE ORGANIZATIONAL CHANGE PROCESS IN ARMY ORGANIZATIONS

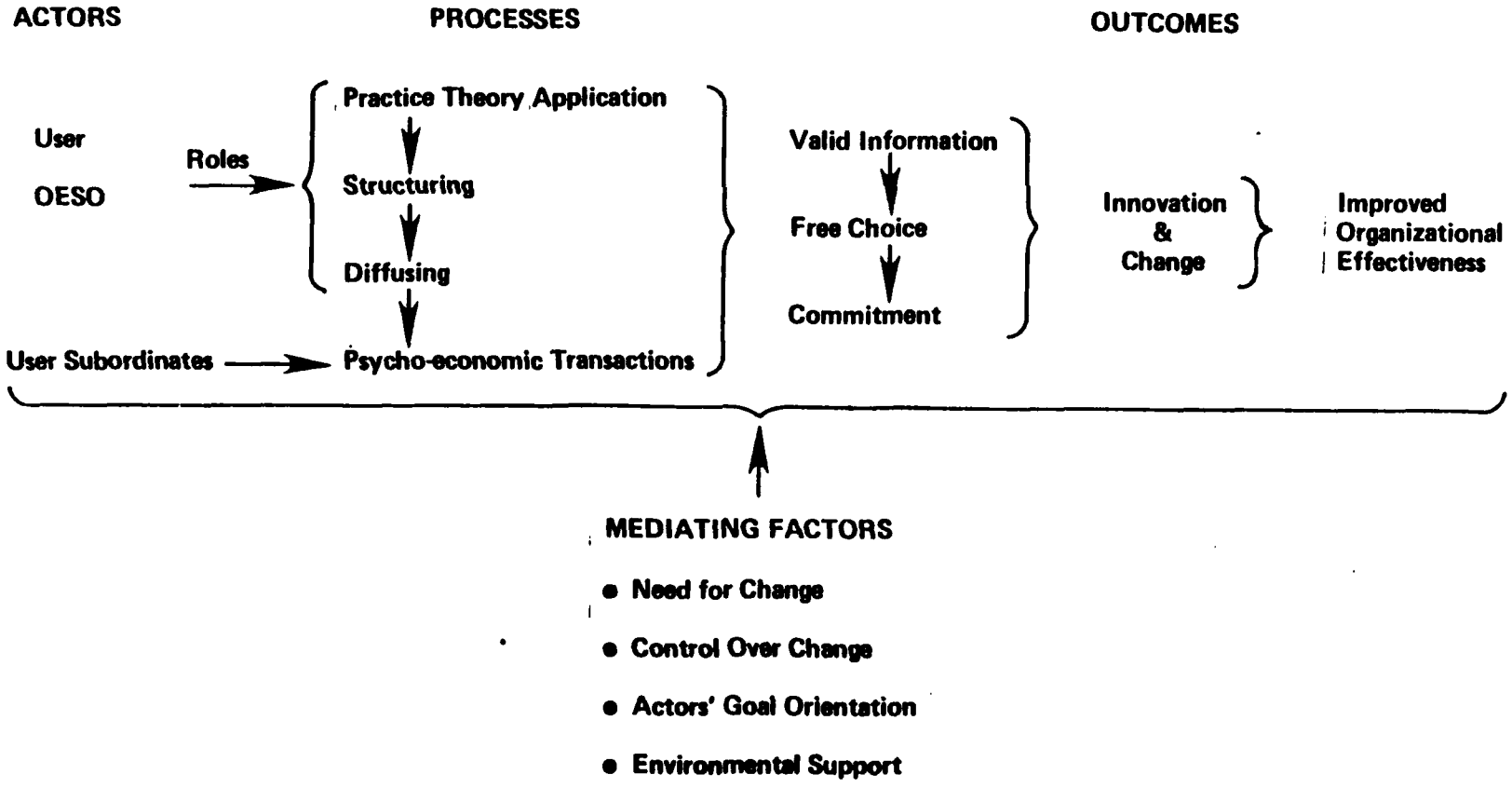



Figure 1. A model of the organizational change process in Army organizations

# Case Selection Matrix for Evaluation of OE Operations

COMPLEXITY OF OE OPERATION OBJECTIVE: LEVEL OF COORDINATION

OE OPERATION CLASSES

| COMPLEXITY OF USER: ORGANIZATIONAL INTERDEPENDENCE / USER CLASSES | A<br>• Interpersonal or inter-group relationship or process objectives<br>• Short term impact (<1 year)<br>• Requires coordination through standardization | B<br>• Process issues interface with technological system issues<br>• Long term impact (>1 year)<br>• External policy, resource or support not needed from environment<br>• Requires coordination by planning | C<br>Process issues interface with technological system issues<br>• Long term impact (>1 year)<br>• External policy, resource, or support needed from environment<br>• Requires coordination by extensive planning and mutual adjustment |
|---|---|---|---|
| I. Small line combat, combat support, combat service support units (≤BN) or military staffs (Primarily pooled interdependence) | | | |
| II. Staffs, internal components of large organizations or components of user's organization where decision-making and coordination is complicated by subgroups of significantly different characteristics. (Primarily sequential interdependence) | | | |
| III. Large combat, combat support, combat service support units (≤DIV) where decision-making coordination is complicated by a lack of direct control over policies, implementation processes, and rewards and sanctions. (Primarily sequential interdependence) | | | |
| IV. Complex systems involving decisionmaking or coordination across boundaries of the organizational hierarchy of the change sponsor. (Primarily reciprocal interdependence) | | | |

Figure 2. Case selection matrix for evaluation of OE operations.

AD P001422

# Problems in the Measurement of Major Change in Specialized Organizations

John Mietus, Ph.D.
US Army Research Institute for the Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

## The Problem

The environment for the military in general is becoming more complex and faster moving. There is a need for more military organizations which themselves are more complex, faster acting and reacting, which design themselves for a mission and environment which is unique today and then redesign themselves for a changing situation in the future. These organizations generally utilize high-technology and their membership is educated. What methods can these units use to sense and adapt rapidly and properly to a changing environment? Which combinations of technology, work structures, organization designs and procedures are appropriate in which environment? In studying this problem, how can the field researcher systematically study both the development of techniques and procedures to help these organizations self-adapt and also study the impact these new techniques have on organizational performance?

To put it differently, what measurement methods are appropriate in a formative and summative evaluation study in which the sample size is one, environmental variables affecting the sample are changing, there are no comparable sample organizations, and the treatment itself is designed to impact on all aspects of the sample and its relationships with its environment?

If the organization is more effective, whatever that is, after the treatment, should the treatment be considered as possibly having had an effect? More importantly, how does one know if one properly and competently administered the treatment?

## Illustration of the Problem

To illustrate the problem and provide us with material for discussion, I shall relate a recent Army Research Institute project in sociotechnical systems analysis as applied to the World-wide Military Command and Control System, Data Processing Center - Europe (WWMCCS, DPC-E).

The military has, like the rest of western society, an increasing number of specialized, high-technology organizations which must adapt to changing technology, staffing, and mission environments. Examples are the data processing centers, signal units, and the air defense units in West Germany. In 1978, sociotechnical systems analysis methods seemed appropriate as a way of helping these units assess and redesign themselves, but practical research remained. A contract was let to Dr. Bill Pasmore at Case Western Reserve University's School of Management. Pasmore was to develop and try

[1]The views expressed in this paper are those of the author and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army,

out a model of sociotechnical systems analysis adapted to the military. The target unit would be the WWMCCS, DPC-E. I was the Contracting Officer's Technical Representative and was on site the first year of the project as a consultant. Note that the purpose was to develop and try out a model adapted to the military, not demonstrate the effectiveness of an already developed model.

Sociotechnical systems theorists argue that the human and technological subsystems of organizations operate according to different sets of laws, and that making one as efficient as possible may have unfavorable effects on the operation of the other; therefore, for the organization to operate smoothly overall, it is necessary to design each system with respect to the other. To accomplish this joint optimization, the Commander, with the help of recommendations from an expert consultant and unit members, over time, reassesses and modifies where necessary the unit's organization structure and policies, technologies, task structures, and work methods.

The WWMCCS, DPC-E is located in Heidelberg, West Germany in the Headquarters of the US Army - Europe. It serves as a communication link for the Joint Chiefs of Staff and as a command and control information systems developer for the headquarters. The director of the facility saw the following challenges in the years ahead: (1) decreasing number and skill level of personnel; (2) need for a mobile capability in addition to the present fixed station posture; (3) increases in total demand for services and a changing type of demand from users.

The overall project strategy could be summarized as a recurring cycle of analysis, recommendations, and implementation. Organization members at all levels would be heavily involved, with the consultant passing on necessary skills to members so they could continue the process after the consultant was gone.

First Data Collection

At the beginning and end of the 20-month time frame during which the consultants periodically visited the unit and carried out the intervention, an intensive data gathering operation was conducted using a survey questionnaire, interviews, observation, and archival data. The survey instrument, very broad ranging yet in-depth, measured organizational climate, work and job characteristics, satisfaction, formalization of procedures, distribution of power, and demographics. The initial data suggested the unit was in fairly good shape. Employees were satisfied with the design of their jobs, supervision, co-workers, rewards. They were motivated by their work, were learning useful skills, and liked their location.

Seventy-eight percent of the unit members were interviewed. Questions were developed by both the consultants and two "core groups." The core groups were parallel organizations within the unit formed to provide an additional sensing/feedback mechanism to the command group. They were composed of 8 to 12 members, recruited from all levels and parts of the unit. Interview data showed individuals working in different parts of the unit held widely divergent views about it. While the interviews again demonstrated a general

satisfaction with the organization, they did also show, in various sections of the unit, that there was dissatisfaction with the work itself, training, fairness, measurement of productivity, cooperation across boundaries, relations with users, and instability of personnel.

The third major assessment was of the technical systems. Adapting methods used primarily by industrial engineers, the consultants assessed role descriptions, skills needed, team and individual goals, task activities, information used, problems encountered, and recommendations for changes in task procedures. A variance matrix was developed which showed impact of problems in one part of the organization on other parts. From this analysis, needed changes in organization design, policies, and procedures could be identified.

Thus, while there was a clear need in the director's mind to prepare for changes in the future, the present state of the unit was reasonably satisfactory. It was operating effectively and to the general satisfaction of most of its members. The need for change was not readily apparent nor was change desired by many personnel. The critical issues were to prepare it for the future and to help it improve its productivity without lessening its quality of work life. The unit would be ideal for the development of a model for sociotechnical systems analysis, but far less desirable as a site to demonstrate the power of the model.

## Changes Made

Several major and many minor changes were made in the unit. Each change was thoroughly thought out and talked over by the core group, by management and the consultants. Decisions were made to move from a fairly standard organization design to a matrix design; this in turn would allow the formation of small temporary project teams and more formal emphasis on skill training. Performance contracts for both project and skill development were to be agreed upon, and rewards would be tied to performance. The core group was to be continued.

Interview data were gathered by the core group members and the consultants throughout the time period of the change, with a formal interviewing of about half the unit one year after the start of the project. It was obvious that the change process here was a slow and difficult one, and the unit was still learning and deciding about itself.

## Later Data Collections

About five months later, the WWMCCS DPC-E took part in a major exercise. This is viewed as the real test of the system's capability. Compared to years past, user demand exploded. New technology put the unit's services in much greater demand for programming, training users, and installing telecommunications equipment. At the same time, there was a slight decrease in unit personnel strength and skill level. The unit trained and serviced significantly more users and increased programmer support of exercise activities by 15 percent while maintaining existing duties.

Four months later, a second survey was readministered to the unit. Note there was about 80 percent turnover in respondents between the first and second survey. There were no signiifcant changes in organization climate, satisfaction, or other attitudes. The unit will be looked at again in the Spring of 1982 in much the same way.

My colleague, Ul James, and an associate, looked at the unit intensively for about 2½ days in August 1981. Using structured interview methodology, they found little awareness of change or improvement among unit members. Unit members attributed the hard outcomes of improved exercise performance not to the sociotechnical systems operation but to the good management and hard effort of the unit. There was some question as to whether the change to the matrix was either very smooth or didn't occur.

Thus we have a difference of opinion on what occurred and why it occurred. Different methods of data gathering at different points in time revealed different aspects of the organization. To what extent are these different viewpoints correct? How does one measure an organization as this one is?

## Restatement of the Problem

To sum up, field research in organizations is costly in time, money, and the potential for error. Often, it is desirable to get a formative and summative evaluation of a new treatment at the same time. Yet in this problem the unit is chosen because it is changing in its technology, personnel, and environment, and is unique. What generalizations can be made from such a study? The problem area seems a proper field of inquiry, yet there are no accepted or proper methods of inquiry. What measurement methods are appropriate to (1) determine how well the treatment is being administered and (2) determine impact of the treatment, when the sample organization is chosen because it is unique, constantly changing, and important.

## Major Reference

Pasmore, W.; Shani, R.; and Kaplan, M. Sociotechnical Approaches to Organization Change in USAREUR. US Army Research Institute for the Behavioral and Social Sciences, Draft Final Report, 1981.

## Alternative Approaches to the Design and Analysis of Objectives-Referenced Competency Tests

O'Neil, Harold F., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia (Chair); Baker, Eva L., Choppin, Bruce, & Quellmalz, Edys S., UCLA Graduate School of Education, Los Angeles, California. (Thurs. P.M.)

The need to describe trainee competency in relation to specific tasks and skill levels requires approaches to the construction and analysis of criterion test situations quite different from current methodologies. In "Matching Tests with Instruction," Eva L. Baker will describe the design of a system for matching explicit content requirements of instructional and testing tasks, estimating their "fit" and assessing the cost of the process. Edys S. Quellmalz will draw upon cognitive learning research to describe a continuum along which features of test-like tasks can approximate the information bases and cognitive operations trainees must activate in actual situations. Alternative validation procedures will also be described. Bruce Choppin describes how latent trait modeling of test responses can be used to define a scale of performance in terms of specific tasks and to locate an individual examinee within a region of the scale.

AD P001423

# Matching Instruction and Tests

Eva L. Baker
Director, Center for the Study of Evaluation
UCLA Graduate School of Education

This paper puts forth the thesis that much of extant testing practice is wasteful owing more to habit, routine, and appeal to dysfunctional theory than to the actual decision needs the tests are intended to serve. An analysis of tests by decision purpose and focus is presented. Articulation across different decision foci may be achieved by attending to the task level. This level is critical both to the assessment and certification of competency and to the design of instruction. Other decision needs may be served by sampling and aggregating analyses from the task level. A specific procedure is presented for dealing with the match between instruction and test tasks, and a research agenda to resolve critical issues is discussed.

## Introduction

The arguments addressed in this paper are three: 1) that much of practice in testing is inefficient; 2) that focusing attention on the task level of performance will give us both better control and understanding of student performance; 3) that the critical work needs to be done in matching instructional with performance features in a non-trivial way.

That present testing practices can be improved is axiomatic. However, the particular focus for improvement is not necessarily the refinement of various tests and indicators that already exist. The root of the problem goes back to the litany of validity: that tests have different purposes. Unfortunately, the practical consequence of this statement has been to create a separate test for every purpose. Justification for the profusion of tests comes from reliance on the authority of fading psychometric theories, habits, and penchant for "scientific" numbers rather than from

clearheaded analysis of decision needs. Without a doubt, tests are needed for different decisions. Figure 1 presents a list of frequently used test purposes.

Figure 1

Applications of Test Information

| A. Uses Directly Affecting Students | B. Uses Affecting Programs |
|---|---|
| 1. Admissions | 1. Modification (formative) |
| 2. Placement | |
| 3. Diagnosis | 2. Validation (summative) |
| 4. Branching | |
| 5. Certification | 3. Comparison (summative) |

Test purposes in column A directly bear on the student's educational options. Different opportunities will be available to him depending upon his own test performance. In column B, effects on students are indirect. Findings for students matter only as they are aggregated to permit inferences about program operations. A subsequent cohort of students, of course, will be directly affected by such program-oriented decisions. But not those who provide the data at the time of testing.

In most educational systems, students are, by necessity, sorted and assigned into educational offerings of differing complexity. At issue is whether separate tests ought to be used to make these decisions. Referring again to Figure 1, column A, using a typical university as an example, we can illustrate the inefficiencies of many current systems.

For example, admission is determined by test performance on the Scholastic Aptitude Test (SAT), a measure nominally of acquired educational expertise. On the basis of SAT, the student is either admitted or rejected. Next, he is given a test for the purpose of placement in a special program targetted toward basic skills. The purpose of this test is to determine whether basic skill special services are needed...whether and not what ser-

vices are required. After placement into a special course to improve writing, for example, the individual course instructor administers a set of tests to diagnose which particular deficiencies are most acute. Although the university setting is not known for branching instruction (except on the basis of student self-selection), some decisions, e.g., to receive tutoring, will be made based on students' performance on classroom assignment and within-course tests. The certification decision, however, may include a course-specific final examination, a departmentally administered test, or readministration of the original general placement examination.

Decisions about programs in this same example may be made on data remote from the actual learning focus. Program improvement needs may be gauged by attendance, other measures of student compliance, instructor satisfaction, and professorial review of the "content" of courses. Validation of most instruction remains a tautological exercise, since measures and instruction rarely are judged in ways that are not inextricably bound ...a repetition of the chicken-egg dilemma. A more refined analysis of testing includes the focus (or insides) of the instruments themselves in order to determine whether information at the appropriate level of specificity has been collected as intended. Figure 2 expands the Figure 1 analysis by adding test focus and illustrative indicators.

Figure 2

Test Purpose, Focus, and Indicator

| PURPOSE | | |
|---|---|---|
| 1. Admission | Global aptitude/achievement | Norm-referenced tests |
| 2. Placement | Content in topical area | Content subscales of norm-referenced tests |
| 3. Diagnosis | Task | Domain-referenced achievement tests |
| 4. Branching | Task/methods interactions | Adaptive testing/"Aptitude" treatment interaction (ATI) |
| 5. Certification | Tasks sampled under range of conditions approximating job | Domain-referenced tests |

Test purposes 1, 2, and 3 simply reflect a successive series of refined sortings, moving from gross prediction to more precise description of what students can and cannot do.

An efficient testing system should attempt to collect data as infrequently as necessary in a resource conserving manner. Let's return to the university example. If we were truly interested in students' ability to use writing to communicate ideas, then such a measure could be given as part of the basis of the admission decision (since such tests correlate at about the same levels as grades, and other concurrently academic tests). The resulting student performance could be scored diagnostically to eliminate the need for additional testing for placement and to reduce greatly the burden placed upon the instructional system itself (the instructor, resources, etc.) for finding out about the student's competencies. By focusing on the task level, one can achieve test purposes for admissions, placement, diagnosis, and certification. The need for "branching" decisions will depend principally upon the adaptive capabilities of the instructional system itself (noted to be at the university not very good). Yet, the logic used to identify tasks for some sort of ATI assignment would be consistent with the analyses of the task in developing a domain-referenced test. Needed subdivisions would depend clearly on the range of alternatives available for students and the need for detailed information to match assignments. (For most instructional systems, branching options are undertaken at levels considerably grosser than our state-of-the-art analysis can handle.)

What is manipulated, then, in this system is the number of samples of

performance collected, and, to some degree, the level of detail for review of students' work. When relatively gross decisions need to be made, such as admissions or placement, then only brief samples of test performance might be required, and testing time greatly conserved. When decisions for remediation, bypass, or certification were addressed, more attention could be given to the representativeness of the tasks sample or the details of achievement of subtasks. When inferences are to be made about the quality of instructional programs, then aggregation in analyses and student and task sampling would be required.

The savings for an approach of this sort are clear, for the most critical resources we have is the time and energy needed for careful development. That investment would be concentrated to measures of important tasks of learning rather than diffused into the development, administration, analysis and interpretation of sets of different measures. This recommended procedure would abolish the "levels" problem inherent in successive sorting requirement and change the question from "articulation" among levels to the decomposition and aggregation of data about the same task set.

By investing a good deal of attention at the task level, we may make the system more rational; at the same time, we increase the risk of error. If our analysis of tasks and the indicators used to infer accomplishment is wrong, then the costs to the system are higher. The third point of this paper is to outline broadly an orderly and serious line of inquiry that examines task requirements. What is special about this research is that task examination proceeds from the outset with concerns for instructional effectiveness. Although other systems for task and instruction matching are in use and are being refined, the need remains for more intensive study of the deeper connections between outcomes and instruction.

The benefits of good fits between test outcomes and instruction are many: 1) They permit us to describe instructional programs; 2) they can serve as a heuristic for design or revision of programs; 3) they can reduce the cost of iterative developmental trials.

Available procedures

Much of the extant research on matches of instruction to testing do not make integrated use of findings from cognitive studies of learning and appear to depend to a great extent on refined behavioral analyses. Some work has been conducted using matrices of content and process and applying them to both test content and instructional material and texts. These efforts attend to the topic overlap, and using the analysis in Figure 2, relate more to the attempt to describe imprecisely developed systems, (e.g., commerical tests and texts) focused on "topic", the level at which students are sorted or placed, rather than on the learning task itself. Even in more rigorously developed systems, matches of test and instruction are based upon relatively gross analysis of cognitive requirements, or on technology and structures proposed by Bloom, et al., 1956; Gagne, 1973, Merrill, Reigeluth, and Faust, 1977. At best these techniques function as heuristics. They employ procedures of self-interrogation by the developers, e.g., "Is X really an instance of this concept?" The actual content requirements of the learner, in terms of cognitive demands is relatively untouched. In some cases, where trainees are to learn one and only one procedure, then such levels of analysis are probably cost-effective. But if we move to concern with general thinking and problem solving skills, and concerns for students' ability to comprehend the re-

lationship of relatively complex information, then such analysis represent only a starting point. I will outline the agenda for research on instruction and testing, not by trying to acknowledge all the partial attempts that might have relevance, but to focus most directly on my view of the least number of components that require study.

Instructional-Task Configurations

A complete model of an instructional system would include the sorting and placement decisions outlined in Figures 1 and 2, as well as descriptions of learners' abilities, (schema and experience), teachers, settings, materials, hardware and task measures. Since the case has been argued for task focus, we would enter the system at the level of tasks to serve as outcome indicators of manipulable instruction. Even within this relatively limited area, a great many research options exist.

Features. Our research operates from the assumption that an integrated and comprehensive set of descriptors (specifications) guides the development of both tasks and instruction. What features should be used to set these important boundaries? A first candidate is the cognitive processing requirements for the to-be learned skills, with attention to "metacognitive" demands as well as requirements for quick, accurate retrieval, reference, and networks. These features may also need to be characterized in terms of the specific content used. This research may be less relevant for those idyosyncratic tasks that require the application of one procedure to one set of content. But if our goal is to produce individuals capable of performing a range of tasks rather than one only, it is a more efficient learning strategy to attempt to connect these tasks on some common basis, rather than teaching by rote a great number of specific

routines. To characterize tasks, for instance, we might want to develop some rudimentary theory of the content of tasks, looking at level of detail, specialized meaning, embeddedness in other meaning networks, or hierarchical structure. So the boundaries or "features" of the matching task include both cognitive processing demands and type and complexity of information.

Fit. A second critical component of this research is the match or fit between instances of test task or instruction set by the boundaries of the above features. What constitutes a fit or match? How do you go about finding one? Two particular areas for study are 1) the degree and confidence of congruency; 2) the vantage point from which the match is made. Most efforts at fit or match are based on casual inspection and an on-off choice, i.e., it matches or it doesn't. Agreement among raters is rarely computed, since only one or two people might be making the judgment. Our research in the analogous and partial problem of matching test instances to domain-referenced test specifications suggests that matches often involve selective attention by different raters to relatively trivial cues, e.g., the number of response alternatives in a test item. Using the concept of fuzzy set, (that is, belongness is not an on-off proposition) game-like notions of probability (that is, how much do you bet...?), and providing cues to attend to the critical features' dimensions of cognitive processing requirements and content, the fit issue can be explored. Keep in mind that we are talking about matching both test tasks and instruction to the same set of parameters.

A second concern with vantage point addresses the process by which whatever procedure for identifying congruency above is actually applied.

"Front-end analysis" as its name implies often begins with looking at the outcomes and tasks. Other researchers, Hively et al., look first at instruction, studying what "successful" instructors do. An entire line of "school effectiveness" studies is based on this same proposition. If instruction is taken as a vantage point, one would have to attend to the extent to which instruction intents (whatever they may be) get diffused and deflected by the processes in the classroom. Focusing on instruction raises the complexities of studying actual processes rather than plans. What information and skills <u>are</u> transmitted, how they become refracted by teachers' misunderstandings, diffused by inaccurate translations, are part of what instruction is about.

<u>Verification</u>. To this point, planning has been a focus of this paper. In the acknowledgement of the importance of verification, we move directly to the instructional and testing settings. We are interested in what has been delivered and the extent to which our characterization of features and fit is useful. Two specific questions are addressed here. The first is "To what extent have we identified the critical features shared by testing and instruction tasks? The second is "Have we characterized the fit accurately?" Both questions address different aspects of validity. We would suggest that these verification questions need to be studied by close-up encounter with the actual events of instruction and testing. Such detailed descriptions can help refine our understanding of the translation from plans to action. We may also need to revise what we call critical features. We also need to reconsider the tasks originally identified by verifying the necessity of on-the-job requirements.

Only by relatively independent verification cycles can we break the tautological links between our views of important tasks and instruction. An additional concern in the verification process is the extent to which the task under study is worth the cost of careful front end analysis, including feature identification, etc. The feedback from the verification activities could be used to change the degree of confidence required (either raise or lower it) for the match itself. In certain areas, it may not be worth the cost.

## Conclusion

This process of focusing on tasks and the match of outcomes and instruction is not recommended unreservedly for most or even much of the curriculum of training for any institution in the same way that no particular routines of test administration or analysis are appropriate across the board. The attempt here is to identify the critical elements for the development of the test/instruction connection for those tasks that are most important. Importance is a matter of policy and not research.



PAGES 1584 - 1594 LEFT INTENTIONALLY BLANK

AD P001424

# CRITERION-REFERENCING OF PERFORMANCE BY LATENT-TRAIT SCALING

by
Bruce Choppin
Center for the Study of Evaluation
University of California, Los Angeles

## ABSTRACT

Conventional true-score approaches to criterion-referenced tests break down when such tests are used for classificatory purposes. Latent trait modeling can help in this situation, but it is appropriate to question whether this classificatory use of criterion-referenced tests is what is needed. Latent trait modeling of test performance can enable straightforward criterion-referenced interpretations of test scores. As the use of testing for individual diagnosis increases, criterion-referencing with latent trait models is likely to become even more attractive.

## Criterion-Referencing of Performance by Latent-Trait Scaling

Over the last 15 years the two major developments in the field of educational measurement have been criterion-referenced testing and the development of latent-trait models. Each topic now commands a substantial proportion of the available space in professional journals and each has benefitted from the close attention of some of the most creative minds and clearest thinkers in our field. Yet, remarkably little of this thinking has been directed towards bringing these two developments together, or into positions from which they could support each other. (A notable exception to this is van der Linden's article in the current issue of the Review of Educational Research to which I shall refer later.)

Criterion-referenced measurement was originally envisaged as a system which permitted the interpretation of test performance in terms of some external criterion. The test score would be used to relate the individual's performance to a meaningful scale of tasks rather than merely conveying information about the individual's standing vis-a-vis other persons who had taken the same test. Performances from the field of athletics, and examinations such as the driving test and measures of typing proficiency were often cited as examples.

In recent years, however, this conception of criterion-referencing has been steadily replaced by another, better suited for use in decision making situations. Educational decisions (in common with many others) frequently depended on which of two categories or

classifications contained a particular object or event. In education once the concept of mastery learning had become fashionable, the classification was often according to whether or not a mastery state had been achieved (Meskauskas, 1976). If achieved, then the instructional program could progress; if not, then some remedial work was indicated. This situation was modeled with idealized domains of items all of which were focused on this particular discrimination between masters and non-masters. One psychometric model on which a considerable amount of work has been published makes the intrinsically satisfying assumption that all the items comprising a domain are of equivalent difficulty, and that mastery of the domain was to be equated with the ability to respond correctly to all the items which comprised it. Other approaches, which allowed for item difficulty variation within the domain, sought to estimate the proportion of the domain which individual students had mastered. Of course these models are not realistically portrayals of actual educational measurement situations. Domains are rarely so well defined. Items are related in varying degrees to the variable to be measured and contain differing amounts of irrelevant and contaminating content. They are not all of equal difficulty. Neither, in many situations, are they selected randomly from the domain for inclusion in tests. Yet the models should not be rejected for these reasons. All models necessarily simplify in order to represent the crucial features of a complex situation, and should not be expected to be a true reflection of reality. Indeed it is precisely this simplification of a real situation that give them a

a potential value.

The attempts to extend classical item analysis procedures to derive indices such as reliability, validity, and discrimination for criterion-referenced tests are carefully considered by van der Linden in the article to which reference has already been made (van der Linden, 1981). He points out that, particularly in the area of test validity, serious objections can be made to all the methods that are in common use. These objections include:

(a) the effect that variations in item difficulty have on the pattern of responses in the pretest-posttest situation which is often used for determining criterion-referenced test validity;

(b) the lack of quality control for the intermediate instructional experiences in the same pretest-posttest design;

(c) the dependence of many of the models (and the parameters they contain) on the size of the variance of performance among the sample of people who provide the validating data.

At issue, of course, are the characteristics of the test instrument at and immediately around the cut-off point. Further, it has been argued (e.g., Glaser & Nitko, 1971) that criterion-referenced validity of a test score requires that the score itself has an external referent, i.e., that it be interpretable in terms of specific levels of behavior in a defined domain. Conventional statistical analyses of criterion-referenced test data do not provide this.

In his review, van der Linden takes on the problem of mastery classification by considering a criterion-referenced test's measurement characteristics: validity, reliability and discrimination at the cut-off point. In this context, he argues for the use of an item information function derived from a latent trait model, and then he demonstrates the application of this approach to criterion-referenced test construction by identifying items that do not discriminate well at the cut-off point.

This approach, employing the item information function, works in theory with all the latent trait models, but for the models with two or three item parameters there may be great difficulties of interpretation. Although such models order a set of items uniquely in terms of their difficulty parameters, this ordering is not in general the same as the ordering of the probabilities for a correct response to each item by any particular student, and in particular by a student at or near the cut-off point. These complex models are proving valuable for detailed study of test-taking behavior under varying conditions, but they do not provide a convenient basis for measuring students, nor for making dichotomous classifications. For these operations we need sets of items that fit one parameter models where such can be developed (Choppin, 1981).

I wish to take issue with only one aspect of van der Linden's review which is implicit throughout his whole treatment, and made explicit in his final paragraph. This is the notion that since criterion-referenced testing is now almost always used for decision making, it is appropriate to treat criterion-referenced tests as instruments designed to classify (into "mastery" and "non-mastery" groups)

rather than to measure. Unfortunately we have no precise methodology in general use for establishing the cut-off point beyond which mastery can be confidently assumed to have occurred. Methods such as the Nedelsky procedure for organizing expert judgment in an objective fashion mask the fact that the necessary bridge between test item response and learning theory has yet to be constructed (Glaser, 1981). I recall from the very early days of the Mastery Learning Movement a time when 80% or more of the students scoring 80% or more on the test was the necessary and sufficient condition for the teacher to infer "mastery" of the material, and to move on to the next unit. As far as I know, there was no empirical evidence to support the use of these percentages. I do remember that rather more than 10 years ago Ben Bloom asked me how many questions you had to ask a student to be sure that he had mastered any single objective. My reply went something along the lines, "For typical test items, you should be looking for a score of 3 out of 3, or if you want to allow the possibility of an accidental error then ask for a score of at least 4 out of 5". I hope that that was not the beginning of the 80% criterion.

To clarify the point, let me turn to my favorite everyday latent-trait, temperature, and the practical realization of its model for measurement - the thermometer. The thermometer is a measuring instrument calibrated so that we can give consistent external meaning to a range of different temperature observations. Where we need only to classify temperatures into two classes around a defined cut-off point, we substitute for the thermometer a degenerate form of it, the thermostat. Thermostats are effective, but they have a

very limited range of applications - and their sensitivity at the cut-off temperature is perhaps their most important feature. The validity, reliability and discrimination (sensitivity) of the thermostat is established empirically by repetitious use. If we were able to perform similar experiments in education to validate cut-off points for mastery learning we might be able to discard measurement. Then when we had established where the cut-off point was, it might be appropriate to use criterion-referenced tests that operated like thermostats.

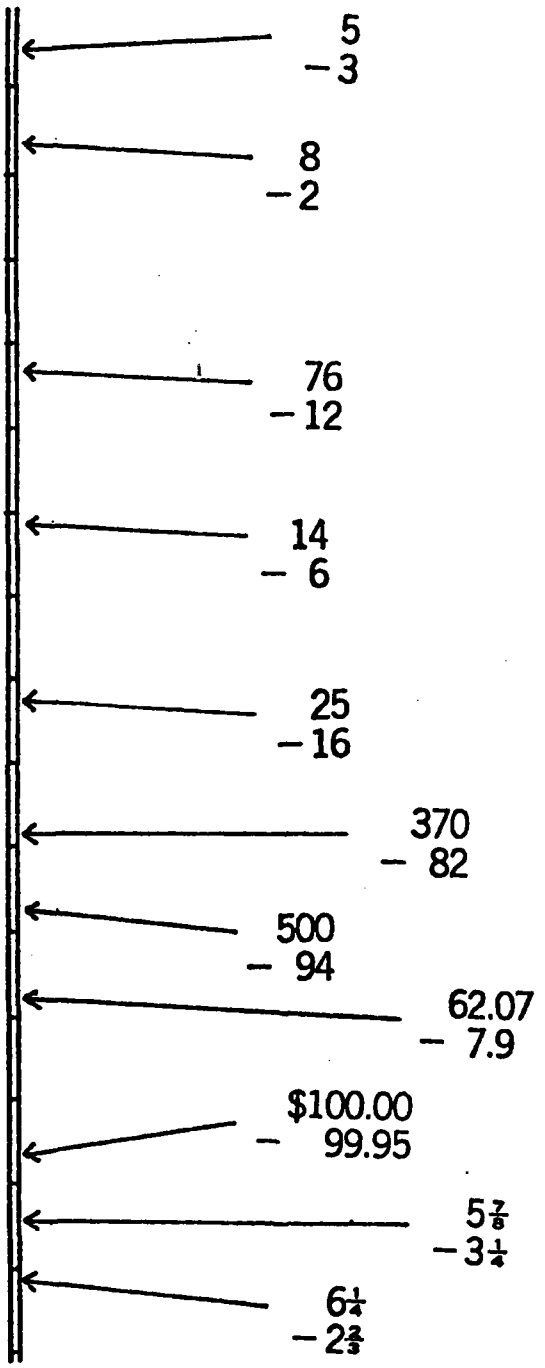
Until that time, we need to use tests as measuring instruments and to calibrate them just as we calibrate general purpose thermometers. We need them to explore the possible effects of using alternative cut-off points on clearly defined scales. Further, these alternative cut-off points need to be defined in terms of some external reference such as sample behaviors or tasks. Latent trait models, since they relate achievement level and item difficulty directly on the same scale, can give such a reference. Two examples which I would bring to your attention are scales for measuring achievement level in arithmetic subtraction (taken from the Keymath Diagnostic Test, Connolly, Nachtman & Pritchett, 1971) and "operations" from the Mathematics Profile Series (Cornish & Wines, 1975). In each, the location of an individual's achievement on this scale gives an immediate visual indication of what material has been mastered, and what remains to be learned.

Another implementation of this approach is to be found in the "KIDMAP" reports on student achievement used (among other places)

| MAPS Scale of Difficulty and Ability | Small Numbers | Large Numbers | Pronumerals |
|---|---|---|---|
| 25 | $3+4=4+\triangle$<br>$6\times 1=\triangle$ | | |
| 30 | $7\times 8=8\times\triangle$ | | |
| | | $123+456=456+\triangle$ | |
| 35 | $15\div 3=\triangle\div 3$ | $44\times 125=125\times\triangle$ | |
| | | $(23\times 24)\times 25=\triangle\times(24\times 25)$ | $p+q=q+\triangle$ |
| | $5-2=\triangle-2$ | $864\div 432=\triangle\div 432$ | $r\times s=s\times\triangle$ |
| | $(3\times 2)\times 5=\triangle\times(2\times 5)$ | | |
| | $9+1=\triangle$ | $984\times 1=\triangle$ | $h\div j=\triangle\div j$ |
| 40 | $5+0=\triangle$ | $876+0=\triangle$ | |
| | $(5+4)+6=\triangle+(4+6)$ | $654-543=\triangle-543$ | $(u\times v)\times w=\triangle\times(v\times w)$ |
| | $8\times 0=\triangle$ | $578+1=\triangle$ | $m-y=\triangle-y$ |
| | | $(89+67)+33=\triangle+(67+33)$ | $(h+j)+k=\triangle+(j+k)$ |
| | $(8-4)+4=\triangle$ | | $g+0=\triangle$ |
| 45 | | $(987-321)+321=\triangle$ | |
| | $15\div 5=30\div\triangle$ | | $y\times 1=\triangle$ |
| | $(12\div 2)\times 2=\triangle$ | $769\times 0=\triangle$ | |
| | | $(625\div 25)\times 25=\triangle$ | $w\times 0=\triangle$ |
| 50 | $8\div 4=\triangle\div 8$ | | |
| | $7-4=\triangle-7$ | | $v+1=\triangle$ |
| | $4+5=(4+6)+(5+\triangle)$ | $654-543=\triangle-654$ | |
| | $(24\div 6)\div 2=\triangle\div(6\div 2)$ | $240\div 15=480\div\triangle$ | |
| | | $123+456=(123+789)+(456+\triangle)$ | $(j-k)+k=\triangle$ |
| 55 | | $468\div 234=\triangle\div 468$ | $s\div t=rs\div\triangle$ |
| | | $(72\times 25)-(60\times 25)=(72+\triangle)\times 25$ | $p+q=(p+r)+(q+\triangle)$ |
| | $(7\times 2)-(3\times 2)=(7+\triangle)\times 2$ | | $(m\div g)\times g=\triangle$ |
| | $(12-6)-4=\triangle-(6-4)$ | $(900\div 30)\div 10=\triangle\div(30\div 10)$ | |
| 60 | | | $mp-np=(m+\triangle)\times p$ |
| | | $(89-56)-21=\triangle-(56-21)$ | |
| | $(40\div 8)\times 4=(40\times 4)\div(\triangle\times 4)$ | | |
| | | | $f-e=\triangle-f$ |
| 65 | | | |
| | | | $k\div p=\triangle\div k$ |
| | | $(72\div 36)\times 9=(72\times 9)\div(\triangle\times 9)$ | $(y\div z)\times w=(y\times w)\div(\triangle\times w)$ |
| 70 | | | |
| | | | |
| 75 | | | $(u-v)-t=\triangle-(v-t)$ |
| | | | $(p\div q)\div r=\triangle\div(q\div r)$ |

OPERATIONS SCALE

(Mathematics Profile Series)
Cornish & Wines, 1977




SUBTRACTION SCALE

(Keymath Diagnostic Arithmetic Test ; Connolly, Nachtman & Pritchett, 1971)

within the Los Angeles County schools, in which the curriculum, expressed in terms of objectives, is mapped out on a single sheet of paper, and the student's achievement level is indicated by his location vis-a-vis the objectives (Wright, Mead & Ludlow, 1980).

It seems probable that the future will see a considerable expansion in the use of testing to diagnose the learning problems of individual students and trainees. Much of our testing technology has grown out of group testing situations whether for program evaluation or for the ranking of testees. The context for individual diagnostic testing will be very different. Such testing necessarily requires criterion-referencing of test performance, and indeed of each item response. Latent trait modeling becomes especially attractive in this context because of its efficient integration of diagnostic evidence from different items. We have some of the techniques for doing this, but much more development work remains to be done.

## REFERENCES


Choppin, B.H. The use of latent-trait models in the measurement of cognitive abilities and skills. In D. Spearritt (Ed.) The improvement of measurement in education and psychology. Australian Council for Educational Research, 1981.

Connolly, A.J., Nachtman, W. & Pritchett, E.M. Keymath Diagnostic Arithmetic Test. American Guidance Service, 1971.

Cornish, G. & Wines, R. Mathematics Profile Series: Operations Test. Australian Council for Educational Research, 1977.

Glaser, R. The future of testing: A research agenda for cognitive psychology and psychometrics. Technical Reprint No. 3. Univ. of Pittsburgh, LRDC, 1981

Glaser, R. & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.) Educational measurement, American Council in Education, 1971.

Meskauskas, J.A. Evaluation models for criterion-referenced testing. Review of Educational Research, 1976, 46, 133-158.

van der Linden, W.J. A latent trait look at pretest-posttest validation of criterion-referenced test items. Review of Educational Research, 1981, 51, 379-402.

van der Linden, W.J. Decision models for use with criterion-referenced tests. Applied Psychological Measurement, 1980, 4, 469-492.

Wright, B.D., Mead, R.J. & Ludlow, L.H. KIDMAP: Person-by-item interaction mapping. Research Memorandum No. 29, University of Chicago: MESA Press, 1980.

AD P001425

# Issues in Designing and Validating Alternative Performance Indicators

Edys S. Quellmalz
Director, Test Design Project
Center for the Study of Evaluation
UCLA Graduate School of Education

A major flaw of many achievement measures is their failure to present ecologically valid tasks. In public schooling, adult education and military training systems, educators are recognizing that multiple choice tests often can not sufficiently represent competence in real-life or on-the-job settings. McNeill (1980), for example, found that general reading comprehension test scores did not represent mechanics' ability to read repair manuals. Consequently, adult education systems now are providing alternative test situations for demonstrating competence.

This presentation will draw upon cognitive learning research to describe a continuum along which features of test-like tasks can approximate both the information base and the cognitive operations trainees must activate in actual situations. Features will include task problem type, display modality, levels of required cognitive processing, criteria for uniformly judging performance and the functional context in which the task is placed. The paper will also discuss alternative procedures for validating that these measures actually do characterize mastery of targetted competencies.

# Issues in Designing and Validating Alternative Performance Indicators

Edys Quellmalz
Center for the Study of Evaluation
UCLA Graduate School of Education

A major problem in military training is documenting that courses build job competencies. Training programs must determine the range of military occupational specialities (MOS) for which a recruit's entering capabilities are most relevant, accredit the acquisition of specific job skills and validate the adequacy with which these skills are applied in the actual job setting. Currently, multiple choice tests are the most common measure of these levels of competence. Such selected response formats derive their popularity primarily from their economy, as well as from claims for their statistical relationship to "other" measures of job success. However, many of these tests are being criticized because they fail to present psychologically and ecologically valid tasks.

In public schooling, adult education and military training systems, educators, psychologists and evaluators are recognizing that multiple choice test scores often do not represent competence in real-life or on-the-job settings. For example, McNeill (1980), found that general reading test scores belied mechanics' ability to comprehend repair manuals. Adult education programs are providing alternative test situations for demonstrating competence (Allenpress, 1980). Military educators, too, recognize the mismatch between the content of many tests and particular job training.

Begland (1981) reports that the Army's functional approach to its Basic Skills Educational Program "recognized that the presence or absence of these skills in any given individual could not be determined on the basis of standardized reading test scores and that programs designed to raise standard test scores lacked focus sufficient to bring individuals to the desired level of competence in the time available. Consequently, TRADOC proposed to develop skills requirements based on detailed analyses of the learning requirements of each MOS." (p. 8)

In their handbook on criterion-referenced testing, Ellis and Wulfeck (1981), too, affirm the need to design criterion tests that mirror demands presented in the occupational speciality. The general problem of designing adequate competency measures, then, is to match the test to the job. However, the more specific problem lies in identifying features which are critical for a fit and determining if the ubiquitious selected response format can provide that fit. In the remainder of this paper, I will reference findings from cognitive research suggesting that selected response formats are very often highly questionable as valid representations of MOS skills, and that the question of fit involves not only the information processing components of cognition, but also the functional ecology of the context within which the competency must operate.

## The Cognitive Task Dimensions

Cognitive studies of the demands of reading, writing, math, physics and chess problems consistently highlight the care with which assumptions about learner's capacities for assimilation, generalization and transfer

must reference the information base and solution routines required to solve tasks as well as the learner's existing knowledge structures or schema (Bransford, 1971; Quellmalz & Capell, 1979; Brown & Vanlehn, 1980; Greeno, 1977; Chi & Glaser, 1980). Masters can call upon more information and activate connected sets of procedures, while novices often stumble when remembering content or orchestrating strategies. The master has automated the steps for operating a piece of machinery, identifying a malfunction and correcting it; the novice laboriously works through the steps. Certainly, the notion of analyzing a job or criterion performance into component, enroute tasks is not new (Skinner, 1958; Gagne, 1977; Markle, 1964). But appreciation of the precision required to merge the targetted course information and procedures with learners' entering schema is new. Furthermore, an issue in the design of competency tests becomes the appropriateness of levels of expertise targetted for end-of-course appraisal vis-a-vis competency levels demanded on the job. In both civilian and military courses, many end-of-course assessments, particularly those using multiple choice tests, tend to ask for component pieces of information or steps, e.g., "list the steps for operating the equipment," but they neglect the end product or combination of the components in an applied situation. Begland (1981) cites the marginal quality of the job analyses presented in Soldier's Manuals, and Ellis, et al. (1980), also acknowledge the reliance on the artistry and intuition of the course developer and test maker in Navy Training courses. The cognitive research strategy of contrasting the configurations of information and the sequences of routines used by novices with those used by experts on the job, offers

an alternative methodology to state-of-the-art procedures for designing competency measures. By specifying the features of the equipment or problem the master must attend to and have information about, it becomes possible to build criterion problem situations for end-of-course competency assessment. Furthermore, a more precise identification of the procedures the master uses permits elicitation of these routines in the test task. When application tasks such as use of a procedure or a rule are required on the job, some form of test task other than a selected response is in order. Although asking trainees to describe a procedure is better than giving a selected response item, it still differs in important ways from actually using the procedure to operate actual equipment. Of course, the more realistic the task, the closer the match between the assessment situation and the required job performance. It is hard to imagine many jobs where multiple choice responses elicit anything but processes enroute to criterion performance. If competency tests are to correspond to job performance, then they must provide comparable stimulus and processing requirements (Quellmalz, 1981).

Display Modalities

The intellectual processes test-like events require can vary along a continuum on at least two dimensions. The first dimension would be a part/whole axis where tasks might elicit component pieces of information and enroute steps, then to proceed to requiring their integration to solve the criterion task. A second, related, dimension is the approximation of simpler tasks to criterion levels of processing, moving from

recognition and recall to application.

Even when learners must produce a response, test stimulus situations can range along a continuum of closer approximations to the actual task. Many end-of-course tests still rely on diagrams or other two dimensional representations of a three dimensional task; others present static renditions of moving phonomena. Compupter simulations and 3-D simulators are increasingly being used to approach realism. Use of computer simulations, however, introduces an additional set of requirements into the trainees task: computer literacy. Little research has examined the complexity of the visual, motor and symbolic demands placed on trainees by computer simulations. Moreover, 3-D simulations can range greatly in their presentation of the stimulus situations that obtain in the workplace. Anyone who has hit tennis balls in a practice alley knows that hitting the ball through the little ring still permits considerable latitude in where it would have landed if it had been hit into a real court.

Clearly, concerns for expense, logistics, equipment damage or safety must be weighed against the need for trainees to demonstrate their competence in a realistic "dry run" situation. For those MOS's where newly trained recruits have difficulty actually applying their newly acquired skills, further research could provide empirical evidence about the short and long term cost effectiveness of designing more job-like performance tasks.

Rating Performance Adequacy

Not only are selected response tests more economical to administer,

they are also less costly to score. When trainees produce written descriptions of "what they would do if..." or when they demonstrate or simulate their skills, a set of logistical and technical problems arise that are common to all performance ratings. Observations of job performance are time consuming, but if they are precisely focused on job relevant criteria, they are clearly the most valid index of job competence. Ideally, these precise, operational criteria should derive from a careful analysis of the performance of masters, referencing components of the process as well as its outcome. Criteria may specify levels of accuracy, speed, and flexibility under variable conditions. While some occupational specialities have clear indices of success, e.g., the pump works, the bomb does not explode, in most jobs there are gradations of mastery. Careful analysis of the aspects of performance that distinguish levels of expertise would provide information useful for instruction and, on the job, for feedback to workers and supervisors about areas in which performance is strong or weak. It is questionable not only whether many MOS job checklists really provide precise diagnostic criteria, but also whether the criteria are operational enough to be applied uniformly. When criteria are unclear or applied unevenly the assessment may be viewed as too subjective and useless by administrators and as biased and unfair by workers.

A second issue in performance ratings, then, is whether even the most relevant criteria will be applied uniformly. Explicit criteria contribute to scale stability; they do not guarantee it. For example, intensive research on issues affecting the stability of raters' judgments of students' writing has identified a number of methods for establishing and maintain-

ing rater reliability within a rating session as well as scale stability across sets of raters and scoring occasions (Quellmalz, 1980a, 1980b). Examination of issues of rating stability as they pertain to military training courses and on-the-job performance and promotion might well reveal the need to strengthen and to systematize existing practices.

## The Functional Context

Learning research is also dramatizing the critical functions context plays in learners' engagement and command of cognitive tasks. While various forms of simulations may permit competency tests to approximate the cognitive demands of job tasks, there is little research in military training and occupations on the features of job contexts or ecologies that mediate cognition and performance. Extrapolating from the methodology of social-anthropological research, educational studies of what Doyle (1977) has termed the "classroom ecology" have shown that the extra-classroom and classroom social organizations significantly mediate student's cognition (Mehan, 1978; Dorr-Bremme, 1980). Participant-observers using micro-and-macro-ethnographic techniques have also examined adult educational systems to identify the constitutive events in the participation structures of both classroom and workplace settings (Finnan, 1979). Findings from these studies describing features of the classroom that influence performances, features of the work context that influence performance and the often gross mismatches between the two settings suggest that competence appraisal must provide for contextual variables.

## Alternative Validations for Performance Measures

I have argued that cognitive research findings cast serious doubts on the validity of selected response tests and advocated that courses and job assessments focus on actual performance. Traditional notions of validity have accepted statistical relationships beween an end-of-course test and "some other" job related measures. Often these correlations are based on total test scores, holistic ratings or measures of attrition. Clearly, the most valid indicator of an end-of-course performance rating would be its relationship to on-the-job ratings. If we believe that end-of-course competence may continue to improve, then some form of analytic ratings of component skills as well as overall performance would probably best distinguish one level of mastery from the next. Accuracy, speed, and flexibility under varying work conditions could be validated by comparing the performance of end-of-course trainees with that of recent course graduates and by verifying that criteria specified for higher skill levels describes advanced personnels' performance. Many "standards of excellence" have been invalidated empirically when workers deemed functional and skilled did not meet conceptually derived criteria. Ecologically valid criteria might include ratings for trainees' abilities to negotiate such contextual problems as the interpersonal and systemic social orders in addition to the technical, cognitive manipulations of the job itself. To assure robustness of course training, validation data that is descriptive and quantative might look far more closely at trainee performance across occupational contexts. Thus validation procedures for competence would be descriptively and psychologically convincing as well as numerically high.

## References


Allenpress, J. Assessing life skills competence. Pittman Learning, Inc., 1979.

Begland, R. R. A multi-faceted approach for the development of the Army's functional Basic Skills Educational Program. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA, April 1981.

Bransford, J. D., & Franks, J. J. The abstraction of linguistic ideas. Cognitive Psychology, 1971, 2, 331-350.

Brown, J. S. & VanLehn, K. Repair theory: A generative theory of bugs in procedural skills. Cognitive Science, 1980.

Chi, M., & Glaser, R. The measurement of expertise. In E. L. Baker & E. S. Quellmalz (Eds.), Educational Testing and Evaluation. Beverly Hills: Sage Publishing Co., 1980.

Dorr-Bremme, D. W. Behaving and making sense: Creating social organization in the classroom. Unpublished doctoral dissertation. Harvard Graduate School of Education, 1980.

Doyle, W. Paradigms for research on teacher effectiveness. In L. Shulman, (Ed.), Review of Research in Education. Itasca, IL: F. E. Peacock, Publishers, Inc., 1977, pp. 163-198.

Ellis, J. A. & Wulfeck, W. A. Handbook for Testing in Navy Training Programs. Navy Personnel Research and Development Center, San Diego, CA, 1980.

Ellis, J. A., Wulfeck, W. H., & Fredericks, P. S. The Instructional Quality Inventory, Volume II. Navy Personnel Research and Development Center, San Diego, CA, 1981.

Finnan, C. R. The development of occupational identity among Vietnamese refugees. Unpublished doctoral dissertation, Stanford University, 1980.

Gagne, R. M. The conditions of learning. New York: Holt, Rinehart and Winston, 1977.

Greeno, J. G. Process of understanding in problem solving. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), Cognitive Theory, Vol. 2. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

Markle, S. M. Good Frames and Bad. New York: John Wiley & Sons, 1964.

McNeil, J. D. Competency Based Reading Skills and the Reading Demands of Minority-Bilingual Auto Mechanics. ED 195 790

Mehan, H. Structuring school structure. Harvard Educational Review, 48(1), February, 1978, pp. 32-64.

Quellmalz, E., & Capell, F. Defining writing domains: Effects of discourse and response mode. Report to the National Institute of Education, Los Angeles: UCLA Center for the Study of Evaluation, 1979.

Quellmalz, E. Designing Instructionally Relevant Scoring Balancing the Ideal and the Real. Invited Symposium paper presented at the Annual Meeting of the National Council of Measurement in Education, Los Angeles, CA, April, 1981.

Quellmalz, E. Controlling Rater Drift. Final Report submitted to the NIE, November 30, 1980b.

Quellmalz, E. Problems in stablizing the judgment process. Paper presented at the Annual Meeting of the American Educational Research Association, April 1980, Boston (Grant No. OB-NIE-G-79-0213 to the UCLA Center for the Study of Evaluation, 1980).

Skinner, B. F. Teaching Machines. Science, 1958, 128, pp. 969-977.

# Profiling the Aptitudes of the Current Youth Population

Sellman, Wayne S., Office of the Assistant Secretary of Defense/MRA&L, Pentagon, Washington, DC (Chair); Eitelberg, Mark J., Laurence, Janice & Waters, Brian K., HumRRO, Alexandria, Virginia.

The National Opinion Research Center (NORC), under contract to the Department of Defense (DOD), admnistered the Armed Services Vocational Aptitude Battery (ASVAB 8A) to a nationally representative sample (N = 11,878) of youth 16-23 years old to: (1) aid in determining the aptitude profile of the current mobilization population and (2) renorm the ASVAB. Demographic data on variables such as age, sex, race, and geographical region were also gathered. Scoring patterns were compared with military and civilian historical data to assess aptitude test score trends over time.

Results of the present study enable a comparison of current military recruits with the current youth population as well as with the 1944 mobilization population. The analysis of these data will make a substantial contribution to recruiting management and mobilization planning throughout DoD.

The symposium will consist of three papers:

1. The 1980 Defense Mobilization Population: Mental Testing and American Youth. Presenter: Wayne S. Sellman

2. Military and Civilian Test Score Trends (1944-1980). Presenter: Brian K. Waters

3. Subpopulation Analyses of the Current Defense Mobilization Pool. Presenter: Janice Laurence

AD P001426

# SUBPOPULATION ANALYSES OF CURRENT YOUTH APTITUDES

by

Mark J. Eitelberg
Janice H. Laurence
Brian K. Waters

Human Resources Research Organization

and

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Manpower, Reserve Affairs and Logistics)

## ABSTRACT

This paper describes the subpopulation analyses that will appear in a forthcoming report on the Profile of American Youth. In 1980, the Department of Defense and the Military Services, in cooperation with the Department of Labor, sponsored a large-scale research project to assess the vocational aptitudes of American youth. A national probability sample of approximately 12,000 young men and women, selected from participants in the National Longitudinal Survey (NLS) of Youth Labor Force Behavior, were administered the Armed Services Vocational Aptitude Battery (ASVAB). The results will be analyzed to identify subgroup differences in test performance. The subgroup variables selected for analysis are age, sex, race/ethnicity, level of education, socioeconomic status, and geographic region. Subgroup comparisons will be made on the basis of Armed Forces Qualification Test (AFQT) scores, ASVAB composite scores, and an estimate of reading ability. The ASVAB scores will also be used to estimate the numbers and percent of 1980 youth population subgroups eligible for military enlistment, based on 1981 Service aptitude standards.

# SUBPOPULATION ANALYSES OF CURRENT YOUTH APTITUDES [1]

by

Mark J. Eitelberg
Janice H. Laurence
Brian K. Waters

Human Resources Research Organization

and

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Manpower, Reserve Affairs and Logistics)

In 1980, the National Opinion Research Center (NORC) of the University of Chicago administered the Armed Services Vocational Aptitude Battery (ASVAB) to a national probability sample of approximately 12,000 young men and women. The procedure and methods used to select the sample were designed to yield a data base of youth that could be statistically projected (within known confidence intervals) to represent the entire population (and important subgroups) born in 1957 through 1964 (Frankel & McWilliams, 1981).

The results will be analyzed to identify differences in test performance among population subgroups and the corresponding qualification rates for military service. The demographic variables selected for analysis are age, sex, race/ethnicity, level of education, socioeconomic status, and geographic region. Subpopulation comparisons will be made on the basis of the Armed Forces Qualification Test (AFQT), four aptitude composites, and an estimate of reading ability. The subpopulation analyses have not been completed. Therefore, this paper contains only a description of the background, methodology, and scope of the demographic comparisons.

## COMPARISON MEASURES

Mean AFQT percentile scores are used since AFQT results are typically reported in terms of this metric. The raw AFQT scores of individuals will be converted to AFQT percentile scores, and the mean percentile scores for each investigated subgroup will then be calculated.

The four ASVAB aptitude composities selected for analysis are Mechanical (M), Administrative (A), General (G), and Electronics (E). The ASVAB subtests comprising the Administrative, General, and Electronics composites are the same in all four Services. The Air Force version of the Mechanical composite was used for the subpopulation analyses. The individual subtests that comprise these composites are shown in Table 1.

---

[1] A paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D.C., October 27, 1981. The views expressed in this paper represent those of the authors and do not necessarily reflect the policy or opinion of the Department of Defense.

Table 1

Common Aptitude Composites and Their Component Subtests

| Mechanical | Administrative | General | Electronics |
|---|---|---|---|
| Mechanical Comprehension | Coding Speed | Arithmetic Reasoning | Arithmetic Reasoning |
| Automotive - Shop Information | Numerical Operations | Paragraph Comprehension | Electronics Information |
| General Science | Paragraph Comprehension | Word Knowledge | General Science |
| | Word Knowledge | | Mathematics Knowledge |

Estimates of reading ability will be obtained for the profile study subgroups by converting ASVAB General composite scores to comparable scores on the Adult Basic Learning Examination (ABLE) (see Mathews, Valentine, & Sellman, 1978). ABLE is a battery of tests (vocabulary, spelling, reading, arithmetic/computation, and arithmetic/problem solving) designed to measure the educational achievement of adults who have not completed high school. ABLE covers 12 years of school achievement through the use of three separate levels of test batteries. Since the ASVAB General composite (which combines Paragraph Comprehension, Word Knowledge, and Arithmetic Reasoning subtests) correlates so highly (.85) with ABLE, it was possible to convert the General composite scores to ABLE scores, and then use these measures as estimates of reading ability expressed in terms of scholastic grade levels.

## SUBPOPULATION ANALYSES

### AGE

Background. The Army Alpha tests from World War I provided some [illegible]e first documented evidence of population differences based on chro[illegible]al age. Since that time, numerous cross-sectional studies have support[illegible] the finding that mental ability (a) reaches a peak in early adulthood (the mid-twenties); (b) declines gradually to about age 50; and (c) drops sharply thereafter. Longitudinal studies conducted since the early 1950s, however, indicate that the pattern of intellectual growth and decline is somewhat different from that which is found in cross-sectional research. Although there is still little longitudinal evidence concerning the shape of the so-called "age curve," the data now imply (a) a pattern of intellectual growth through early adulthood; (b) general stability during the middle decades of life (with increases in certain abilities and decreases in others); (c) a gradual and minor decline beginning after the age of 50; and (d) increased decline during the 70s and 80s.

Two-year groupings will be used to separate the 1980 youth population by age. The age categories, by years of birth and age at testing, are as follows: 1961 and 1962 (ages 18 and 19); 1959 and 1960 (ages 20 and 21); 1957 and 1958 (ages 22 and 23).

The analysis of age differences will concentrate on mean AFQT percentile scores and measures of reading ability.

## SEX

Background. Many standardized tests of general aptitude are designed to eliminate (or counterbalance) items or subtests that result in systematically higher scores for one sex over the other. The belief that differential factors should be minimized or balanced is based on the assumption that (a) there is no clear understanding of which specific test items are the best indicators of general aptitude and (b) no special "advantage" in measured performance on these tests should be given to either sex.

Nevertheless, the consistent trend has been that males tend to excel on tests of mathematical reasoning (or quantitative ability), spatial abilities, and mechanical/science aptitudes; and females tend to excel on tests involving verbal fluency or the mechanics of language, memory abilities, perceptual speed, and manual dexterity (Tyler, 1965; Maccoby & Jacklin, 1974).

The AFQT measures verbal and quantitative abilities in approximately equal proportion. This balance reduces the likelihood of sex-related differences in test performance.

The analysis will present data on the mean AFQT percentile scores of males and females by the three age groups. Mean percentile scores of males and females on the four aptitude composites and the estimated reading grade levels of young men and women in the general population will also be presented.

## RACE/ETHNICITY

Background. In this country, most studies of racial/ethnic group test performance focus primarily on the differential abilities of white and black children and young adults. Published evidence suggests that, on standardized tests of mental ability, (a) whites, on the average, score higher than blacks; (b) average group differences remain fairly constant during the school years (the smallest differences occur at the very young ages); (c) blacks perform relatively better on verbal tests than on non-verbal tests; (d) the socioeconomic, geographic, and educational correlates for racial minority groups and whites are generally similar (though there are some differences in the magnitude of correlation); and, further, (e) the differences between individuals of the same race exceed in magnitude the average differences between separate races.

Attempts to measure racial differences in test performance in the civilian sector can be traced back as far as the late nineteenth century. As in the military testing experience (Eitelberg, 1981), there is a remarkable unanimity of results in civilian testing: at each age-level and under a variety of conditions, blacks, on the average, regularly score below whites (Jensen, 1980; Scarr, 1981). There are regional variations; nevertheless, these variations are similar for blacks and whites, and the racial differential remains fairly constant from one region to another.

Although the majority of studies involving racial/ethnic groups in this country currently concentrate on the differences between whites and blacks, there is a long history of research regarding the relative abilities of different "ethnic" (i.e., national origin) groups. The volume of scientific research on the topic of ethnic differences (or race differences other than those between whites and blacks) has lessened greatly since World War II. Nevertheless, there is some degree of consistency in the data on test performance. For instance, study results in this country show that (a) white school children of European ancestry score, on the average, considerably higher than children from racial and ethnic minority groups (with the notable exception of certain Asian-American groups) -- and especially those that are socioeconomically disadvantaged (e.g., Hispanics, native Americans, as well as blacks); and (b) the test performance of racial/ethnic groups perform noticeably better on certain kinds of tests than on others, and the extent of group differences will change according to the types of tests that are emphasized).

The profile study results classify the population into three groups: white and others (including all non-Hispanic and non-black racial/ethnic subgroups), black (non-Hispanic), and Hispanic. These three groups are used since they represent the largest relative racial/ethnic subgroups within the general population. Yet, it should be noted that the hispanic category includes several separate ethnic groups (e.g., Mexican-Americans, Puerto Ricans, Cubans and other Latin Americans, Spanish and Portuguese) variously described simply as being of "Hispanic" origin. Furthermore, the category defined as "white and others" includes Native Americans, Pacific Islanders, and persons of Asian ancestry. (Since the data are weighted, and the proportion of "non-white" groups in the general population is so small in comparison with whites, the differences between the combined group and a "white only" group are negligible). For the purposes of the profile study, then, references to the "white" and "white and other" racial/ethnic groups are synonymous.

The mean AFQT percentile scores for whites, blacks, and Hispanics will be analyzed (by total subgroup and two-year age categories). In addition to the AFQT score comparisons, the mean score for males and for females within the separate racial/ethnic subgroups will be compared on the basis of common aptitude composites and estimated reading ability.

## LEVEL OF EDUCATION

Background. There is a strong positive correlation between aptitude test performance and the amount of formal education. There are, however, several problems involved in using years of schooling as a focus of analysis. For example, there are differences in the quality of education from geographical region to region, school to school, and other related factors. In addition, education variables are not easily isolated or separated from other variables (e.g., age and socioeconomic status).

For the profile study, educational attainment is defined according to high school graduation status. The three categories of graduation status are: (a) non-high school graduate (including, in some cases, high school students as well as drop-outs); (b) recipient of the General Educational Development (GED) high school equivalency certificate; and (c)

Table 2

**1981 Service Enlistment Aptitude Standards**

(Required Operational Score on ASVAB 8 - 10)

| Service/Education | Males | | Females | |
|---|---|---|---|---|
| | Operational Standards | | Operational Standards | |
| | AFQT | Aptitude Composites | AFQT | Aptitude Composites |
| **Army** | | | | |
| High School Diploma Graduate | 16 | 85 on 1 | 16 | 85 on 1 |
| Non-High School Graduate (Including GED) | 31 | 85 on 2 | 31 | 85 on 2 |
| **Navy** | | | | |
| High School Diploma Graduate | 17 | - | | School Eligible[a] |
| GED | 31 | - | | School Eligible[a] |
| Non-High School Graduate | 38 | - | | Not Eligible |
| **Marine Corps** | | | | |
| High School Diploma Graduate | 21 | GT[b] = 90 | 50 | - |
| Non-High School Graduate (Including GED) | 21 | GT[b] = 95 | | Not Eligible |
| **Air Force** | | | | |
| High School Diploma Graduate | 21 | G[c] = 30; MAGE[d] = 120 | 21 | G[c] = 30; MAGE[d] = 120 |
| GED | 50 | G[c] = 30; MAGE[d] = 120 | 50 | G[c] = 30; MAGE[d] = 120 |
| Non-High School Graduate | 65 | G[c] = 45; MAGE[d] = 170 | 65 | G[c] = 45; MAGE[d] = 120 |

[a]Department of the Navy, "Criteria for selection of recruits and new accessions for formal school training." NAVMILPERSCOM Instruction 1236.1A. Washington, D.C.: Naval Military Personnel Command, Jan. 1981.

[b]General-Technical Composite

[c]General Composite

[d]Mechanical, Administrative, General Electronics Composites

THE PROFILE OF AMERICAN YOUTH marks the first time that a vocational aptitude battery has been given to a national probability sample. Up to this time, such research has not been conducted due to the great difficulty and expense involved in obtaining data. The subgroup analyses of current youth aptitudes and the projected qualification rates will soon be completed and released by the Department of Defense.

## REFERENCES


Anastasi, A. Differential Psychology: Individual and Group Differences in Behavior (3rd Edition). New York: The MacMillan Company, 1958.

Bock, R. D. and Moore, E. G. J. Advantage and Disadvantage: Vocational Prospects of American Young People. Chicago, IL: National Opinion Research Center, October 1981.

Eitelberg, M. J. Subpopulation Differences In Performance on Tests of Mental Ability: Historical Review and Annotated Bibliography. Technical Memorandum 81-3. Washington, D. C.: Directorate for Assession Policy, Office of the Secretary of Defense, August 1981.

Featherman, D. L. "Schooling and Occupational Careers: Constancy and Change in World by Success." Constancy and Change in Human Development. Edited by G. Brian and J. Kagan. Cambridge, MA: Harvard University Press, 1980.

Jensen, A. R. Bias in Mental Testing. New York: The Free Press, 1980.

Maccoby, E. E. and Jacklin, C. M. The Psychology of Sex Differences. Stanford, CA: Stanford University Press, 1974.

Mathews, J. J., Valentine, L. D., and Sellman, W. S. Prediction of Reading Grad Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB). AFHRL-TR-78-82. Brooks AFB, TX: Air Force Human Resourcs Laboratory, December 1978.

Scarr, S. Race, Social Class, and Individual Differences in I.Q. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981.

Sewell, W. H. and Hauser, R. M. Education, Occupation, and Earnings. New York: Academic Press, 1975.

Tyler, L. E. The Psychology of Human Differences (3rd Edition). New York:

AD P001427

# APTITUDE TESTING IN DOD AND THE PROFILE OF AMERICAN YOUTH STUDY

by

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Manpower, Reserve Affairs, and Logistics)

and

Janice H. Laurence
Human Resources Research Organization

## ABSTRACT

This paper provides a brief discussion of aptitude testing in the Department of Defense and the rationale for Defense sponsorship of the Profile of American Youth Study. Also described is the historical development of the current version of the Armed Services Vocational Aptitude Battery (ASVAB), as well as its enlistment eligibility composite, the Armed Forces Qualification Test (AFQT). The aptitude profile study involved administration of the 1980 version of ASVAB to a national probability sample of approximately 12,000 young men and women ages 16 to 23. The young people sampled were participants in the National Longitudinal Survey (NLS) of Youth Labor Force Behavior sponsored by the Department of Labor. The methodology, sampling procedures, test administration, and data quality controls used in the execution of the Profile of American Youth Study are described.

# APTITUDE TESTING IN DoD AND THE PROFILE OF AMERICAN YOUTH STUDY [1/]

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Manpower, Reserve Affairs, and Logistics)

Janice H. Laurence
Human Resources Research Organization

## INTRODUCTION

Discussions of present or future military manpower procurement policies consider the way in which individuals are selected for service, assigned to military jobs, and trained to perform those jobs. Philosophically, there is consensus that enlistment standards are essential for manning an effective military. Beyond that broad agreement, the type and kind of enlistment standards (i.e., medical, moral, educational, and aptitude) are topics for ideological, legal and scientific debate.

The Armed Services have devoted considerable effort to develop reliable and valid methods for assessing persons prior to their entering military service. One focus of these efforts has been on the development of tests which measure the aptitudes of individuals. Aptitudes have historically been defined as measures of trainability for the various military jobs.

Aptitude levels within the military have been referenced statistically to the extensive testing of adult males that took place during World War II. This World War II "reference population" has been the baseline for comparing aptitudes of military examinees and recruits across time. Recently, questions have been raised concerning the appropriateness of retaining the World War II reference population as the sole basis for today's military personnel decisions. Accordingly, it was decided that the contemporary youth population should be examined to facilitate the Department of Defense's understanding of the quality and representativeness of its new enlistees.

An aptitude profile of current youth is important for managing recruiting and evaluating recruiting results. The Department of Defense (DoD) should be able to compare the characteristics of today's population with DoD requirements for military manpower. Information is also needed for mobilization planning. If a national emergency required the resumption of conscription, DoD must be able to establish entrance standards compatible with available manpower resources that meet the personnel needs of the Services. Decisions on who should be drafted or permitted to volunteer require an accurate knowledge of the aptitudes of contemporary youth.

---

[1/] Paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D.C., October 1981.

The 1980 Profile of American Youth. The Profile of American Youth was designed to assess the vocational aptitudes of young people, ages 16 to 23, and, at the same time, to develop a new reference population against which scores on the DoD enlistment test can be interpreted. To achieve these goals, DoD contracted with the National Opinion Research Center (NORC) of the University of Chicago to administer the enlistment test to a nationally representative sample of about 12,000 young men and women.

Beyond its value to military manpower planning, aptitude profiles from a national sample of young people are a significant contribution to scientific research. Such aptitude profiles have not been previously available due to the difficulty and expense in obtaining representative data.

## APTITUDE TESTING IN DOD

The Armed Services Vocational Aptitude Battery. The test used in the 1980 aptitude profile study was the Armed Services Vocational Aptitude Battery (ASVAB). ASVAB was introduced on January 1, 1976 as the single DoD test to replace the various aptitude test batteries then in use by each Service. Replacement forms were subsequently implemented on October 1, 1980. A 1980 version (Form 8A) of ASVAB was administered in this study.

ASVAB scores serve two important purposes in the enlistment process. First, they help determine eligibility for enlistment. Second, they are used to establish qualifications for assignment to specific military jobs.

The ASVAB consists of the following 10 subtests: arithmetic reasoning, numerical operations, paragraph comprehension, word knowledge, coding speed, general science, mathematics knowledge, electronics information, mechanical comprehension, and automotive-shop information. These subtests are included because research and experience have shown them to be valid predictors of success in military training.

The scores of four subtests (word knowledge, paragraph comprehension, arithmetic reasoning and numerical operations) are combined to produce an Armed Forces Qualification Test (AFQT) score. The AFQT score, supplemented by the scores on various aptitude composites, are used in conjunction with educational, medical, and moral standards to determine applicant enlistment eligibility. The scores on the aptitude composites also determine eligibility to enter specific military skills.

The Services combine a variety of subtests to form aptitude composites. Table 1 shows the subtests that comprise two selected composites.

TABLE 1

**Examples of Selected Aptitude Composites Derived From Combinations of ASVAB Subtests**

| Administrative Composite | Electronics Composite |
|---|---|
| Paragraph Comprehension | Electronics Information |
| Word Knowledge | General Science |
| Numerical Operations | Arithmetic Reasoning |
| Coding Speed | Mathematics Knowledge |

The Armed Forces Qualification Test. During the early years (1940-1942) of World War II, men were accepted for service if they had completed the fourth grade or were able to pass literacy screening tests; in later years (1943-1945), minimal literacy was no longer required for induction (Ginzberg, Anderson, Ginzburg & Herma, 1959). After service entry, the primary test instrument for job assignment purposes was the Army General Classification Test (AGCT). A test of general trainability, the AGCT was composed of questions which measured verbal, arithmetic, and spatial abilities. After World War II, it was used by the Army for enlistment screening. Modeled after the AGCT, the Armed Forces Qualification Test (AFQT) was introduced in 1950 to determine the eligibility of draftees and volunteers to enter any of the Services (Uhlaner & Bolanovich, 1952).

To minimize test compromise and to update test language and content, the AFQT has been revised periodically. Until 1973, each new AFQT was calibrated back to the AGCT so that successive AFQT scores would have a constant meaning in terms of the level of trainability. In 1972, the use of a common AFQT was discontinued. From 1973 through 1975, each Service estimated an AFQT score from its own test battery. The ASVAB became operational as the single DoD enlistment test in 1976, and AFQT scores have been based since then on a common test. The AFQT composite of ASVAB used in this study (Form 8A) was calibrated against an earlier version of AFQT (Form 7A) used operationally from 1960 through 1972. This calibration established the linkage to the World War II reference population, thereby enabling percentile scores from the new AFQT to have the same interpretive meaning as scores from predecessor tests.

AFQT Categories. For reporting purposes, AFQT scores have traditionally been grouped into five broad categories. Persons whose scores place them in Categories I and II are above average in trainability; those in Category III, average; those in Category IV, below average; and those in Category V, markedly below average and, under current Service policy, not eligible to enlist. The Services prefer enlistees in the higher AFQT categories because training time and associated costs are lower, and such recruits are more likely to qualify for specialized training in a greater number of occupational areas. Table 2 shows the percentile scores for the various categories and the percent of the World War II reference population in each. AFQT percentile scores are based on the World War II population of officers and enlisted men who were on active duty as of December 31, 1944 -- 11,694,229 males.

TABLE 2

**AFQT Scores by Category**

| AFQT Category | Percentile Score Range | Percent of Reference Population in Each Category |
|---|---|---|
| I | 93 - 100 | 8 |
| II | 65 - 92 | 28 |
| III | 31 - 64 | 34 |
| IV | 10 - 30 | 21 |
| V | 1 - 9 | 9 |
| | | 100 |

## STUDY METHODOLOGY

The Profile of American Youth is closely related to the five-year National Longitudinal Survey of Youth Labor Force Behavior (NLS) in three ways. The first and most important relationship is that the profile study used for its sample young people who completed the first annual interview of the NLS in 1979. The profile study used the NLS sample because it was an already-existing nationally representative sample of young people in the age group of interest. Second, the data collection for both studies was carried out by the National Opinion Research Center (NORC). Third, there would be a sharing of data between the two studies. Demographic data collected by the NLS were added to the ASVAB test information obtained in the profile study.

The purpose of the NLS is to study the behavior within the labor market of a large and representative cross-section of American youth. Information about youth born from 1957 through 1964 is being collected through annual personal interviews. The NLS is primarily concerned with problems relating to employment and unemployment. The interviews also gather a great deal of supplemental information about the characteristics, experience, plans, and attitudes of the young people.

## STUDY RESEARCH DESIGN

The Sample. The NLS sample was designed to represent the national population of youth ages 14 to 21 as of January 1, 1979 (Frankel & McWilliams, 1981; McWilliams & Frankel, 1981). Civilian members of the youth population were obtained by screening approximately 80,000 households during the fall of 1978. This screening identified approximately 14,000 eligible youth of the appropriate age. Members of the youth population serving in the military were selected in the fall of 1978 from lists provided by the Defense Manpower Data Center (DMDC). Youth in the military were eligible for selection if they were (a) serving in the Armed Services as of September 30, 1978 and (b) would be between the ages of 17 and 21 as of January 1, 1979. In the spring of 1979, NORC interviewed 12,686 civilian and military youth for the first annual (baseyear) NLS survey.

The NLS baseyear sample contains youth from both urban and rural areas, youth from all major census divisions, and approximately equal proportions of males and females. The sample overrepresents, in a statistically appropriate way, certain key groups, such as Hispanics, blacks, economically disadvantaged whites, and women in the military. This overrepresentation allows for more precise analyses of these groups than would otherwise be possible.

The profile study used for its target sample the 12,686 young people who completed the first annual (1979) interview of the NLS. The 11,914 tests administered represent a completion rate of approximately 94 percent. Table 3 shows the composition of the completed profile sample by sex and race/ethnicity.

TABLE 3

**Description of the Profile of American Youth Sample**

| Racial Ethnic Group | Male | Female | Total |
|---|---|---|---|
| White[a] | 3,531 | 3,496 | 7,027 |
| Black | 1,511 | 1,511 | 3,022 |
| Hispanic | 902 | 927 | 1,829 |
| Total | 5,944 | 5,934 | 11,878 |

[a]"White" is used throughout this report to refer to non-Hispanic, non-Black persons, (i.e., white and others).

Since the Services primarily recruit individuals who are 18 years of age and older, the profile study focused upon young people born between Janury 1, 1957 and December 31, 1962. Thus, the age range for the profile study sample is 18 through 23 years at the time of testing. Table 4 shows the profile study sample of 9,173 people of enlistment age. Table 5 displays the corresponding size of the 1980 national youth population (weighted sample) by year of birth, race/ethnicity, and sex.

Table 4

**Profile of American Youth Enlistment Age Sample**
**By Year of Birth, Race/Ethnicity, and Sex**

| Year of Birth | Age at Time of Testing | White | | Black | | Hispanic | | All | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Male | Female | Male | Female | Male | Female | Male | Female | Total |
| 1962 | 18 | 458 | 401 | 213 | 210 | 108 | 145 | 779 | 756 | 1,535 |
| 1961 | 19 | 363 | 418 | 207 | 211 | 129 | 116 | 699 | 745 | 1,444 |
| 1960 | 20 | 445 | 448 | 197 | 206 | 123 | 110 | 765 | 764 | 1,529 |
| 1959 | 21 | 490 | 519 | 169 | 195 | 108 | 109 | 767 | 823 | 1,590 |
| 1958 | 22 | 477 | 505 | 190 | 167 | 92 | 102 | 759 | 774 | 1,533 |
| 1957 | 23 | 521 | 488 | 167 | 166 | 93 | 107 | 781 | 761 | 1,542 |
| TOTAL | | 2,754 | 2,799 | 1,143 | 1,155 | 653 | 689 | 4,550 | 4,623 | 9,173 |

Table 5

**Profile of American Youth Enlistment Age Sample**
**Estimated Size of National Youth Population by**
**Year of Birth, Race/Ethnicity, and Sex**

(In Thousands)

| Year of Birth | Age at Time of Testing | White | | Black | | Hispanic | | All | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Male | Female | Male | Female | Male | Female | Male | Female | Total |
| 1962 | 18 | 1,677.9 | 1,616.1 | 295.4 | 292.1 | 139.5 | 123.5 | 2,112.8 | 2,031.7 | 4,144.5 |
| 1961 | 19 | 1,701.6 | 1,643.9 | 296.6 | 293.1 | 140.0 | 124.3 | 2,138.2 | 2,061.3 | 4,199.5 |
| 1960 | 20 | 1,729.6 | 1,669.8 | 295.7 | 290.2 | 134.8 | 127.8 | 2,160.1 | 2,087.8 | 4,248.0 |
| 1959 | 21 | 1,753.2 | 1,675.3 | 285.2 | 289.3 | 120.1 | 131.8 | 2,158.8 | 2,096.4 | 4,255.1 |
| 1958 | 22 | 1,755.5 | 1,708.7 | 284.1 | 289.5 | 122.0 | 131.7 | 2,161.6 | 2,129.9 | 4,291.4 |
| 1957 | 23 | 1,762.8 | 1,700.4 | 275.7 | 282.9 | 121.2 | 127.5 | 2,159.7 | 2,110.8 | 4,270.4 |
| TOTAL | | 10,380.6 | 10,014.2 | 1,733.0 | 1,737.1 | 777.6 | 766.6 | 12,891.2 | 12,517.9 | 25,409.1 |

Quality of the Sample. To provide DoD with an assessment of the sample design, development of sample case weights and sampling statistics, an independent panel of sampling experts (Dr. B. F. King, University of Washington; Dr. L. Kish, University of Michigan; Dr. G. E. Hall, U. S. Bureau of Census; and Dr. J. Sedransk, State University of New York) was convened. The panel concluded: (a) the sample design was appropriate for meeting the objectives of the profile study and (b) all of the statistical procedures used in the development of sample case weights and sampling statistics met the professional criteria established for efforts of this nature, both in the public and private sectors. (Frankel & McWilliams 1981).

## TEST ADMINISTRATION

During the period July through October 1980, NORC representatives administered the ASVAB to the 11,914 young people who comprise the profile sample. Testing was generally conducted in groups of five to ten persons. More than 400 test sites, including hotels, community centers, and libraries throughout the United States and abroad were used. The test was administered according to strict guidelines conforming to ASVAB procedures, which assured both accuracy and consistency of results. Great care was also taken to assure confidentiality.

In May 1981, NORC sent to all respondents a copy of their test results, information to interpret the scores, and a brochure containing vocational and educational information. In addition, participants were paid honoraria for completing the test. The decision to pay an honorarium was based on NORC's experience in similar studies which indicated that a powerful incentive would be needed in order to get young people to travel up to an hour to a testing center, spend three hours or more taking a test and then travel home. The honorarium was set at $50.00.

NORC's decision to provide an incentive honorarium was also influenced by the importance of the NLS and an obligation to ensure that the added demands of the profile study on the NLS respondents would do nothing to damage further NLS participation. It was anticipated that the monetary incentive offered for participation in the aptitude profile study would work against attrition of the NLS sample and would even increase the goodwill of its members.

## STUDY QUALITY CONTROL

Quality of Data Files. A DoD team of testing experts and computer programmers verified that ASVAB scores and demographic information had been accurately transcribed from the original source documents (i.e., answer sheets and questionnaires) to the computer tape provided to DoD. A random sample (one percent of the cases) was selected for the data audit. For the sample cases, ASVAB answer sheets were hand-scored and demographic questionnaires were manually reviewed. In every case, the information from the source documents had been correctly recorded (Sellman & Hagan, 1981).

Quality of ASVAB. To evaluate the suitability of the ASVAB for measuring the aptitudes of a national sample of young people, DoD contracted with Dr. R. D. Bock, an authority on educational and psychological testing from the University of Chicago. Dr. Bock evaluated the test to determine its appropriateness for measuring vocational aptitudes and its equity for minorities and females. He concluded that the ASVAB is useful for measuring aptitudes of civilian youth and that cultural test bias was not apparent for minorities and females. Moreover, he indicated that the quality of ASVAB equals or surpasses that of commercial aptitude and achievement tests (Bock & Mislevy, 1981).

## REFERENCES


Bock, R. D. and Mislevy, R. J. Data Quality Analysis of the Armed Services Vocational Aptitude Battery. Chicago, IL: National Opinion Research Center, August 1981.

Frankel, M. R. and McWilliams, H. A. The Profile of American Youth: Technical Sampling Report. Chicago, IL: National Opinion Research Center, March 1981.

Ginzberg, E., Anderson, J. K., Ginsburg, S. W., and Herma, J. L. The Lost Divisions. New York: Columbia University Press, 1959.

McWilliams, H. A. The Profile of American Youth: Field Report. Chicago, IL: National Opinion Research Center, December 1980.

McWilliams, H. A. and Frankel, M. R. The Profile of American Youth: Non-Technical Sampling Report. Chicago, IL: National Opinion Research Center, October 1981.

Sellman, W. S. and Hagan, H. T. The Profile of American Youth: Data Audit. Technical Memorandum 81-1. Washington, D. C. Directorate for Accession Policy, Office of the Secretary of Defense, April 1981.

Sheatsley, P. B. The Profile of American Youth: Pretest Report. Chicago, IL: National Opinion Research Center, September 1980.

Uhlaner, J. E. and Bolanovich, D. J., Development of the Armed Forces Qualification Test and Predecessory Army Screening Tests, 1946-1950. PRS Report 976. Washington, D.C.: Personnel Research Section, Department of the Army, November 7, 1952.

# MILITARY AND CIVILIAN TEST SCORE TRENDS

(1950 - 1980)

AD P001428

By

Brian K. Waters
Mark J. Eitelberg
Janice H. Laurence

HUMAN RESOURCES RESEARCH ORGANIZATION

## ABSTRACT

This paper reviews test score trends over nearly a 40-year period. Since military recruits come from the general youth population, knowledge of aptitude test score trends in civilian environments has direct implications for military enlistment test analyses. The data which track these trends originated from 257 published sources on aptitude and achievement tests. The literature shows a remarkable consistency of aptitude testing trend data over the period. In general, both aptitude and achievement test scores have decreased at a rate of about one to three percent of a standard deviation per year. This trend continues today, although there is evidence that the rate of decline has lessened somewhat in the past three years. Similiar trend data were found for military recruits, particularly in the uppermost ability range. The authors conclude that these trends are real, national in scope, and continuing, though at a decreasing rate of decline since about 1977. They summarize the study by recommending that DoD and service manpower planners monitor national test score trends for both short-term recruiting and mobilization planning purposes.

# MILITARY AND CIVILIAN TEST SCORE TRENDS [1/2/]

(1950 - 1980)

Brian K. Waters
Janice H. Laurence
Mark J. Eitelberg

Human Resources Research Organization

Since 1975, the College Entrance Examination Board (CEEB) has published several reports on Scholastic Aptitude Test (SAT) score decline. The CEEB data were similar to data on the American College Testing (ACT) Program, as well as a number of achievement tests. The subject of declining student aptitudes and achievement has since dominated space in educational and psychological literature, with many reports and books receiving heavy media and public exposure. Numerous symposia, commissions, and studies have also been launched to answer three key questions: 1) Are the test score declines a "real" national phenomenon?; 2) What are the cause(s) for the decline?; and 3) What can be done to reverse the trend of declining scores?

The nature and scope of aptitude and achievement test score changes in the national population are of considerable interest to military manpower and personnel managers. The national population provides the pool from which military applicants are drawn (also draftees during periods of national emergency). The civilian population also provides a baseline upon which to assess current and historical recruit quality. And, in the context of the current major research effort to profile the aptitudes of American youth, a review of the civilian aptitude and achievement test score decline places into better perspective military test score trends over the survey period.

This paper describes test score changes between the early 1950s and 1980 (when the profile study was conducted). The paper begins with a tabular and graphic picture of national scholastic aptitude and achievement test score trends. The paper then describes military accession test score trends on AFQT from 1967 to 1980. AFQT score trends are reported by high, median, and low scores. The paper concludes with a brief summary and interpretation of the trends.

---

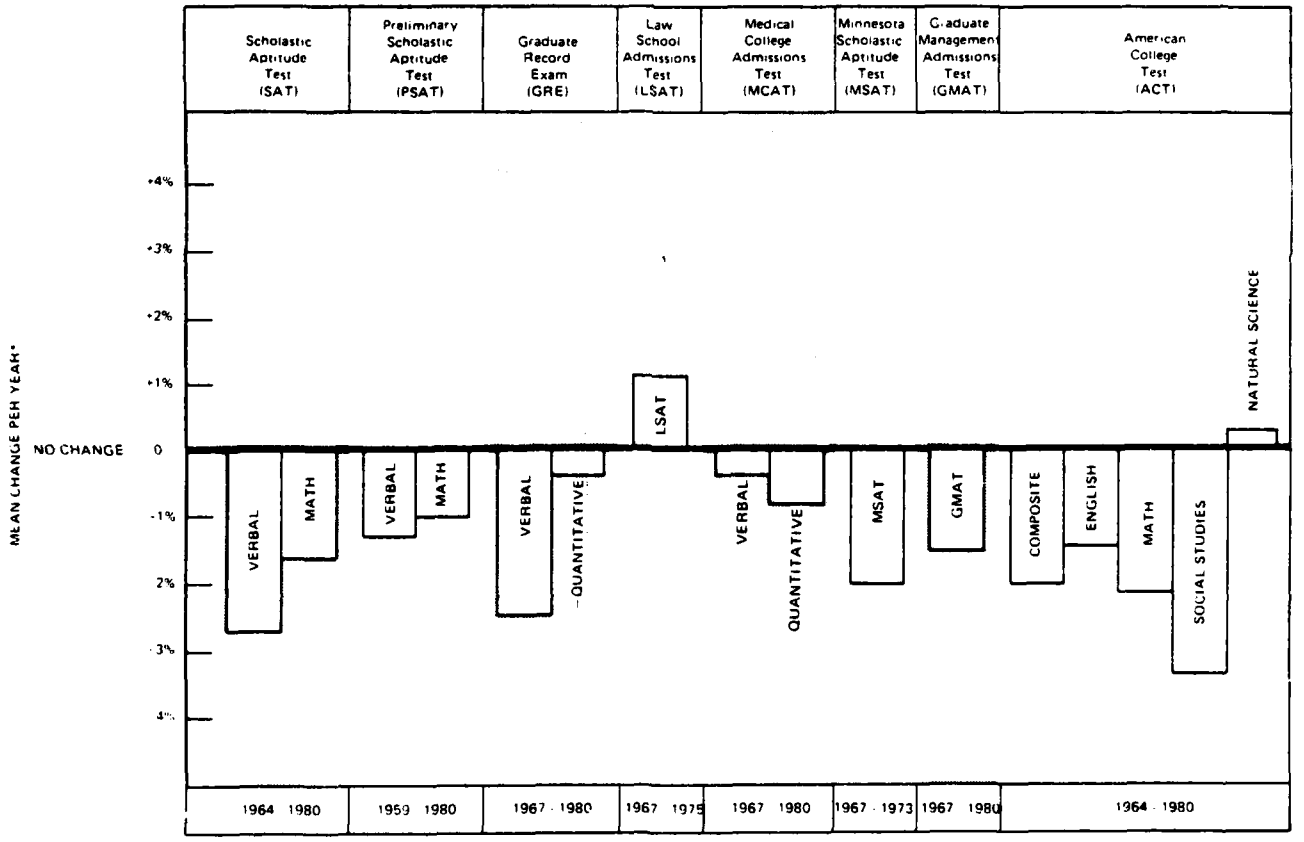
[1/] Paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D.C., October 1981. The views expressed in this paper represent those of the authors and do not necessarily reflect the views of the Department of Defense.

[2/] Much of the material in this paper is based upon Waters, Brian K., The Test Score Decline: A Review and Annotated Bibliography. Technical Memorandum 81-2. Washington, D.C.: Directorate for Accession Policy, Office of the Secretary of Defense, August 1981.

## CIVILIAN SCHOLASTIC APTITUDE TEST SCORE TRENDS

The most clear, consistent, and unambiguous evidence of the decline in the civilian target population comes from the aptitude testing domain. Table 1 and Figure 1 provide a compilation of the trends in this country since the early 1950s.

The aptitude test data show remarkable consistency. With the exception of slight increases on the Medical College Admissions Test-Quantitative Subtest and the Law School Admissions Test, the other measures of scholastic aptitude consistently decreased at a rate of about one to three percent of a standard deviation per year. This trend continues, although there is some evidence that the rate of decline has slowed somewhat through 1980. Other major trends show that verbal scores have tended to decrease faster than quantitative scores; female scores have declined more rapidly than male scores, particularly in the verbal domain; and overall aptitude test scores increased from 1950 through about 1965, and decreased consistently until the late 1970s. There appears to be a lessening of the rate of




*Mean proportion of a standard deviation per year

Source: (Waters, 1981)

**Figure 1. Civilian Aptitude Measures**

Table 1

**Scholastic Aptitude Measures**

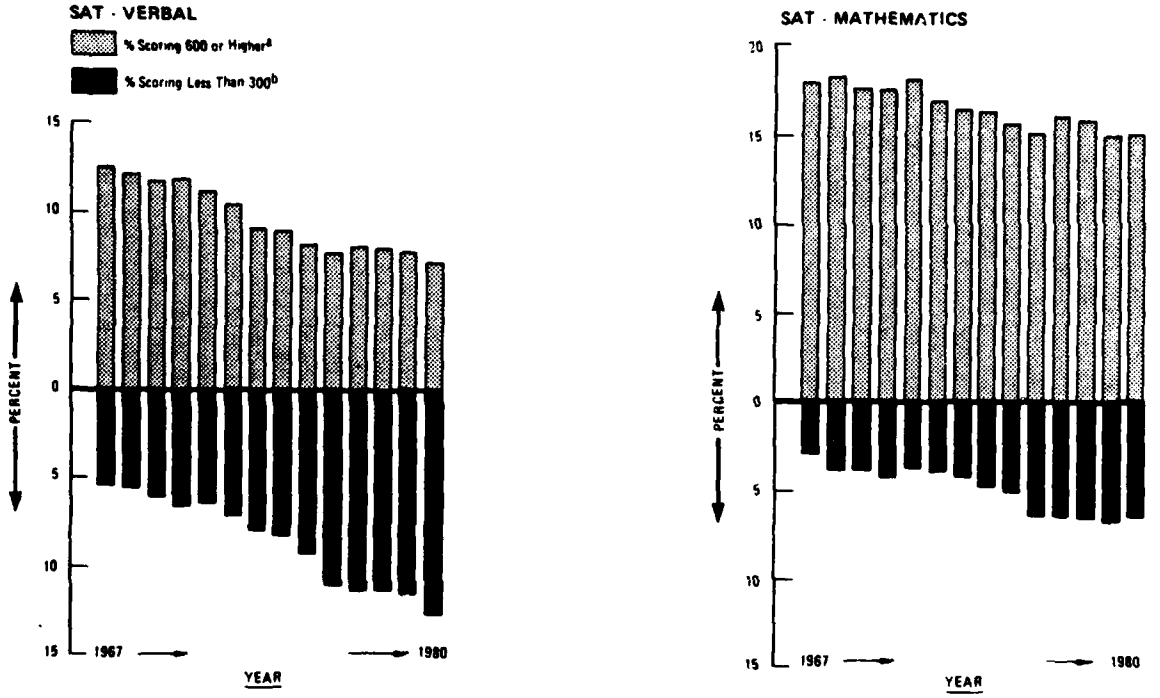
| Instruments | | Time Periods | Grade(s) | Annual N | Areas | Trends %SD/YR |
|---|---|---|---|---|---|---|
| Scholastic Aptitude Test (SAT) | Verbal | 1967-1980 | 11/12 | 1,500,000 | NE, E, EC | Male 2.3 |
| | | 1967-1980 | | | | Female -3.2 |
| | | 1952-1963 | | | | Total +0.2 |
| | | 1964-1980 | | | | Total 2.7 |
| | Mathematics | 1967-1980 | 11/12 | 1,500,000 | NE, E, EC | Male -1.4 |
| | | 1967-1980 | | | | Female 1.6 |
| | | 1952-1963 | | | | Total +0.6 |
| | | 1964-1980 | | | | Total 1.6 |
| American College Test (ACT) | Composite | 1964-1980 | 11/12 | 850,000 | NC, S, W | Male 1.3 |
| | | | | | | Female 2.5 |
| | | | | | | Total -2.0 |
| | English | 1964-1980 | 11/12 | 850,000 | NC, S, W | Male -0.9 |
| | | | | | | Female -2.3 |
| | | | | | | Total -1.4 |
| | Mathematics | 1964-1980 | 11/12 | 850,000 | NC, S, W | Male -2.3 |
| | | | | | | Female -2.0 |
| | | | | | | Total -2.3 |
| | Social Studies | 1964-1980 | 11/12 | 850,000 | NC, S, W | Male -2.4 |
| | | | | | | Female -4.1 |
| | | | | | | Total -3.3 |
| | Natural Science | 1964-1980 | 11/12 | 850,000 | NC, S, W | Male +0.9 |
| | | | | | | Female -0 |
| | | | | | | Total +0.3 |
| Preliminary Scholastic Aptitude Test[1] (PSAT) | Verbal | 1959-1980 | 11 | 1,000,000 | NE, E, EC | Male -0.7 |
| | | | | | | Female -1.8 |
| | | | | | | Total -1.3 |
| | Mathematics | 1959-1980 | 11 | 1,000,000 | NE, E, EC | Male -1.1 |
| | | | | | | Female -0.8 |
| | | | | | | Total -1.0 |
| Minnesota Scholastic Aptitude Test[2] (MSAT) | Form A | 1958-1968 | 11 | 60,000 | Minn. | Total +6.2 |
| | Form C | 1967-1973 | 11 | 65,000 | Minn. | Total -2.0 |
| Graduate Record Exam (GRE) | Verbal | 1967-1980 | 16 | 800,000 | National | Total -1.3 |
| | Quantitative | 1967-1980 | 16 | Not Available | National | Total -0.4 |
| Law School Admissions Test (LSAT) | | 1967-1975 | 16 | Not Available | National | Total +0.6 |
| Medical College Admissions Test (MCAT) | Verbal | 1967-1975 | 16 | 55,000 | National | Total -1.8 |
| | | 1977-1980 | | | | Total -2.8 |
| | Quantitative | 1967-1975 | 16 | 55,000 | National | Total +1.0 |
| | | 1977-1980 | | | | Total -4.3 |
| Graduate Management Admissions Test (GMAT) | | 1967-1975 | 16 | 400,000 | National | Total -1.5 |

[1] PSAT Scores corrected for scale drift, 1967-1980.
[2] MSAT Calculations split in 1986-1987 when current form was introduced. No MSAT after 1973.

Source: (Waters, 1981).

decline between 1977 and 1980, although the decline continued through 1980.[1/] The "causes" of the consistent aptitude test score patterns are not at all clear. Nevertheless, the general conclusion of most authors is that there are multiple factors contributing to the trend.

SAT Score Trends. Figure 2 displays SAT Verbal and Mathematics subtest trends from 1967 to 1980.




[a]Scores of 600 on the SAT - VERBAL and the SAT - MATH equate approximately to scores in the 90th percentile and 92nd percentile, respectively, on the AFQT distrubution

[b]Scores of 300 on the SAT - VERBAL and the SAT - MATH equate approximately to scores in the 50th percentile and 31st percentile, respectively, on the AFQT distribution

SOURCES: College Entrance Examination Board.

**Figure 2. Score Distributions on the Scholastic Aptitude Test (SAT), 1967–1980**

Figure 2 depicts the relatively consistent decline on both subtests at both ends of the scale range. As will be seen in the next section of this paper, these declines are similar to test score patterns of military recruits who score at the higher and lower ends of the Armed Forces Qualification Test (AFQT).

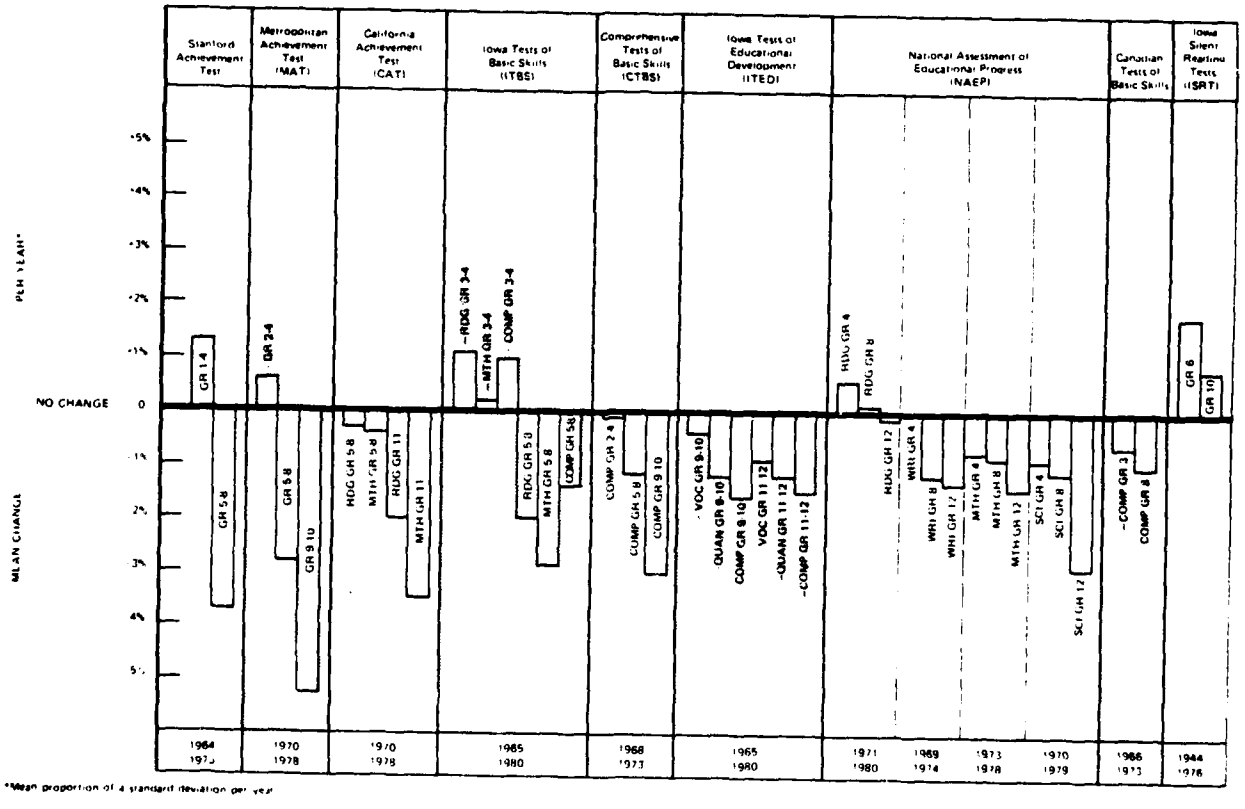
---

[1/] A recent release from ETS reports that 1981 mean SAT scores were identical to 1980 mean scores on both quantitative and verbal subtests.

## CIVILIAN SCHOLASTIC ACHIEVEMENT TEST SCORE TRENDS

Table 2 and Figure 3 depict 1964-1980 mean results for 10 achievement test batteries. Data for the individual batteries have been grouped, when available, into grades 1-4, 5-8, and 9-12 and by subtests that roughly parallel the verbal/quantitative/composite breakouts of the aptitude measures discussed above. As might be expected, long-term trends across achievement content areas are not as consistent as across the more "factorially pure" aptitude areas. Trend data are displayed in Table 2 by percent change in standard deviation per year where both means and variances were provided in the original source. Figure 3 displays Table 2 graphically.

In general, the authors found consistent evidence of achievement test score declines in all areas tested above grade 4, for the 1960s through 1970s. Pre-schoolers and children in the early years of grammar school (1st - 3rd grade) generally scored higher on all measures, while the scores of 4th grade students remained fairly stable. In the opinion of the authors, these trends are real, national in scope, and continuing -- though at a decreasing rate of decline since about 1977.




*Mean proportion of a standard deviation per year

**Source: (Waters, 1981)**

**Figure 3. Civilian Achievement Measures**

Table 2

**Scholastic Achievement Measures**

| Instruments | | Time Period | Grade(s) | Annual N | Areas | Trends %SD/YR |
|---|---|---|---|---|---|---|
| Stanford Achievement Test | | 1964-1973 | 1-4 | 6,000,000 | National | +1.3 |
| | | | 5-8 | | | -3.7 |
| Metropolitan Achievement Test (MAT) | | 1970-1978 | 2-4 | 400,000 | National | +0.6 |
| | | | 5-8 | | | -2.8 |
| | | | 9-10 | | | -5.3 |
| California Achievement Test (CAT) | | 1970-1978 | 2 | Not Available | National | "Slight Gain" |
| | | | 5/8 | | | Rdg -0.3 |
| | | | | | | Mth 0.4 |
| | | | 11 | | | Rdg -2.0 |
| | | | | | | Mth -3.5 |
| Iowa Tests of Basic Skills (ITBS) | | 1965-1980 | 3-4 | 50,000 | Iowa | Rdg +1.1 |
| | | | | | | Mth +0.2 |
| | | | | | | Comp +1.0 |
| | | | 5-8 | | | Rdg 2.0 |
| | | | | | | Mth -2.9 |
| | | | | | | Comp -1.4 |
| Comprehensive Tests of Basic Skills (CTBS) | | 1968-1973 | 2-4 | 200,000 | National | Comp -0.1 |
| | | | 5-8 | | | Comp -1.2 |
| | | | 9-10 | | | Comp. -3.0 |
| Iowa Tests of Educational Development (ITED) | | 1965-1980 | 9-10 | Not Available | Iowa | Voc -0.4 |
| | | | | | | Quan -1.2 |
| | | | | | | Comp -1.6 |
| | | | 11-12 | | | Voc 0.9 |
| | | | | | | Quan -1.2 |
| | | | | | | Comp -1.5 |
| National Assessment of Educational Progress (NAEP) | Rdg | 1971-1980 | 4 | 75,000 to 100,000 | National | Rdg +.06 |
| | | | | | | Wri -0 |
| | Wri | 1969-1974 | | | | Mth -0.7 |
| | | | | | | Sci -0.9 |
| | Mth | 1973-1978 | 8 | | | Rdg +0.1 |
| | | | | | | Wri -1.2 |
| | Sci | 1970-1979 | | | | Mth -0.8 |
| | | | | | | Sci -1.1 |
| | | | 12 | | | Rdg -0.1 |
| | | | | | | Wri -1.3 |
| | | | | | | Mth -1.4 |
| | | | | | | Sci -2.9 |
| Canadian Tests of Basic Skills | | 1968-1973 | 3 | Not Available | Canada | Comp -0.6 |
| | | | 8 | | | Comp -1.0 |
| Iowa Silent Reading Tests[1] (ISRT) | | 1944/1976 | 6 | 15,000/8,000 | Indiana | +1.8 |
| | | | 10 | 11,000/8,000 | | +0.8 |
| General Educational Development (GED) | | 1964-1979 | Mean | 120,000 to | National | -0.86 % |
| | | | 10 | 700,000 | | Met Stds/Yr |
| | | | | | | (73% Vs. 60.1%) |

[1] ISRT scores adjusted for examinee age changes between testing sessions.

Source: (Watters, 1981)

## OTHER INDICATORS OF POPULATION PERFORMANCE CHANGE

The literature on the test score decline includes many references to other indicators of a declining national level of academic competence of youth. These indicators include elementary, high school, and college teachers' opinions, statewide competency-based assessment, measures of curricula content at all levels, analyses of classroom hours attendance per student, analyses of teacher education and practices, and physiological hypotheses about diet, drug, medication, nuclear radiation and other possible correlates of declining test scores. It is beyond the scope of this paper to attempt to analyze the probable or possible causes of the declining scores; however, an excellent review by Rimland and Larson, (1980) is available.

## SUMMARY OF OVERALL CIVILIAN TESTING TRENDS

It is evident that national youth performance on scholastic aptitude and achievement tests has been in a state of decline. Assuming comparability of populations (for current military-age youth, between the ages of 17-24), the scope of the decline would likely represent a decrease of about one-fifth to one-third of a standard deviation on the average from the 1970 pool of AFQT examinees, or about 2-3 percent of a standard deviation per year. This rate would equate to a decline of approximately 4-5 raw score points for the average military enlisted recruit between FY 1971 and FY 1980. The next section of this paper looks at the military data.

## MILITARY RECRUIT APTITUDE TEST SCORE TRENDS

Figure 4 displays the percentages of non-prior service military accessions who scored in AFQT Category I (Sellman & Laurence, 1981) between FY 1967 through FY 1980. Category I scores are roughly equivalent to a score of 600 and above on the SAT (shown in Figure 2 above).

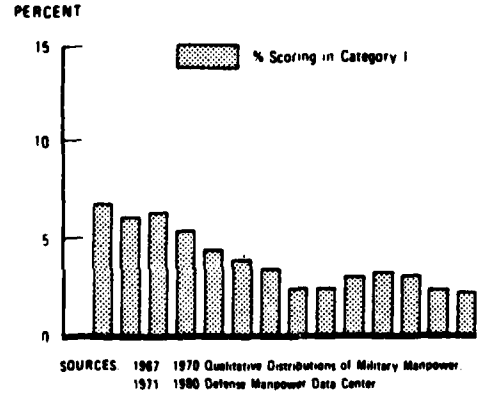



**Figure 4. Military Accessions: AFQT Category I (Top 8%) 1967–1980**

AFQT Category IV military accession statistics are strongly influenced by DoD and Service policies, changes in the recruiting and job market, and other factors independent of the aptitude levels in the national population of enlistment-age youth. Nevertheless, the proportion of AFQT Category IV accessions did increase from 23 percent in FY 1967 - FY 1969 to 26 percent in FY 1979 - FY 1981. The latter figure probably reflects the effects of the miscalibration of AFQT for the first two years of the latter period, (Aptitude Testing of Recruits, July 1980). However, FY 1981 test scores are correct. The three percent increase from the earlier period is roughly comparable to the increased percentages in SAT Verbal and Mathematics scores below 300 over the same time frame (Figure 2).

A median provides a single index of the full distribution of AFQT scores. The medians declined from 79.4 to 72.4 or seven AFQT raw score points over the 14-year period. This decline parallels the two to three percent standard deviation yearly decline observed for civilian aptitude test scores during the same year period. Figure 5 shows median AFQT scores grouped in three-year periods from FY 1967 through FY 1981.

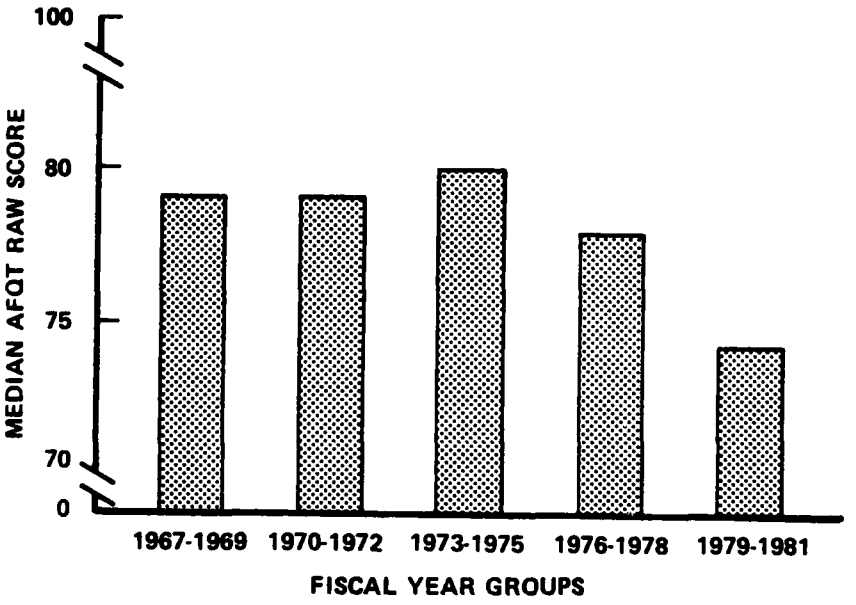



Figure 5. **Median AFQT Raw Scores for Military Non-Prior Service Accessions, 1967-1981, by Three Year Periods**

## IMPLICATIONS OF CIVILIAN AND MILITARY RECRUIT APTITUDE TEST SCORE TRENDS

It appears that civilian aptitude test score trends provide useful information to Defense manpower planners on the aptitude levels of American youth in the pool from which potential recruits are drawn. It is therefore important that the Department of Defense monitor national test score trends for both short-term recruiting and mobilization planning purposes.

## REFERENCES


Aptitude Testing of Recruits. A Report to the House Committee on Armed Services. Washington, D.C.: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics), July 1980.

Rimland, B. and Larson, G. E. The Manpower Quality Decline: An Ecological Perspective. NPRDC Technical Note 81-84. San Diego: Navy Personnel Research and Development Center, November 1980.

Sellman, W. S. and Laurence, J. H. "Aptitude Testing in DoD and the Profile of American Youth Study." A paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D. C., October 1981.

Waters, B. K. The Test Score Decline: A Review and Annotated Bibliography. Technical Memorandum 81-2. Washington, D.C.: Directorate for Accession Policy, Office of the Secretary of Defense, August 1981.

# Course Development and Evaluation Procedures for the Combined Arms and Services Staff School

Stillman, Jon, LTC, Command and General Staff College, Fort Leavenworth, Kansas (Chair); Anderson, Mike, De Frain, Dennis, LTC, Ekwall, Ralph, Command and General Staff College, Fort Leavenworth, Kansas.

This panel describes the course development and evaluation procedures for the Army's new Combined Arms and Services Staff School ($CAS^3$). This is a program to train Army staff officers and was developed from a zero base.

The curriculum consists of two phases, each of which is very different from the other. Phase one provides background information in a self-paced correspondence mode. Phase two is presented in a small group setting by means of learning activities that resemble real world staff work.

The panel will be introduced by the $CAS^3$ supervising author and the three panel members will present information on various aspects of program development. The first presentation will describe overall program development. The second presentation will describe the use of the systems approach in curriculum development. The third presentation will deal with the development of an evaluation system for a two phase program.

PREVIOUS PAGE IS BLANK

AD P001429

# STUDENT AND COURSE EVALUATION AT THE COMBINED ARMS AND SERVICES STAFF SCHOOL

BY

MICHAEL R. ANDERSON, Ph. D.
Education Specialist
U.S. Army Command and General Staff College
Fort Leavenworth, Kansas 66027

This paper focuses on the student and course evaluation that was performed during Phase II of the Combined Arms and Services Staff School (CAS3) in its initial implementation, April through June 1981. Descriptions of the CAS3 training philosophy, goals, and instructional techniques are provided as background. The performance oriented student evaluation philosophy and rating scheme is explained. Results and student acceptance of this student evaluation system are discussed. The information sources and results of the course evaluation are also presented.

## BACKGROUND

The Combined Arms and Services Staff School ($CAS^3$) is designed to train officers of the Active Army and Reserve Components, worldwide, to function as staff officers with the Army in the field. The course consists of two phases: Phase I is a nonresident course which the officer completes prior to attending the resident portion; Phase II is the resident phase which will be attended in a TDY status. The approximate length of the course is 142 academic hours for Phase I and 360 hours for Phase II. The curriculum provides several opportunities, as a staff officer, to think about and analyze situations; formulate courses of action; and recommend and justify a selected course of action to his/her commander. This paper focuses on the student and course evaluation that was performed during Phase II of $CAS^3$ in its initial implementation, April through June 1981.

The Phase II $CAS^3$ educational and training philosophy is that the student learns by doing. The student participates in seven staff exercises each of which serves as the focal point for the staff interactive process. The seven exercises are linked together through a course-long general scenario.

$CAS^3$ believes that successful performance as a staff officer requires: (1) the ability to analyze and solve military problems, (2) effective communication skills, (3) the ability to interact and coordinate as a member of a staff, and (4) an understanding of Army organization, operations, and procedures. The primary function of $CAS^3$ is the development and improvement of these qualities within the individual students assigned. As such, the curriculum of $CAS^3$ has been designed to provide each student with opportunities for personal growth in each of the aforementioned areas. Phase I primarily addresses the fourth requirement while the Phase II instruction addresses the all the requirements. Furthermore, the Phase I cognitive objectives are generally directed at the knowledge and comprehension level of Bloom's taxonomy while the Phase II cognitive objectives are directed at the higher four levels of the taxonomy.

The educational method used during Phase II is markedly different from that employed at other levels of the officer education system. The educational method revolves around the small group participatory process. The students are formed into small groups, or staffs, of 12 individuals. Students fill a variety of roles in these staffs normally based upon their specialities and educational needs. It is within these staff group that the staff processes and products are encountered and developed. Each staff group is under the continuous tutelage of one member of the faculty, the Staff Leader. This individual is the proctor, instructor, monitor, advisor, and evaluator for his staff. He is responsible for the education of his staff group members. He monitors the progress of his students and insures that the students achieve the educational goals and objectives of the course.

## PHASE II STUDENT EVALUATION

The main purpose of the Phase II evaluation system is to provide an indepth assessment of each student's demonstrated capabilities throughout the seven staff exercises as a basis to enhance student growth and development. As such, the evaluation process focuses on the four staff officer requirements stated previously. Secondarily, the evaluation system is used to assess the curriculum, maintain standards for graduation, and insure student accomplishment of the course objectives.

The evaluation philosophy of CAS$^3$ is that the staff leader is in the best position to provide and honest assessment of student capabilities and suggestions for improvement. Hence, there are no standard examinations in Phase II. Instead, the staff leader is constantly evaluating a student's performance and providing feedback both formally and informally. The formal feedback is provided as staff leader ratings on the learning objectives and goals of instruction. The informal feedback consists of the day-to-day written and oral comments provided by the staff leader. A basic principle at CAS$^3$ is that a student's work is at an acceptable level when the course begins and that the work remains acceptable until definitely demonstrated otherwise. The final student assessment is based on the capabilities of the student as he/she exits the course and is not cumulative in nature.

### Informal Evaluation

Staff leaders control the day-to-day feedback and evaluation of students. In this regard, staff leaders use group discussion, the coordination of staff plans, briefings, answers to impromptu questions as well as written assignments in the evaluation of student performance.

Written products are reviewed and critiqued as appropriate, and then returned to students for information or action. At the discretion of the staff leader, a written task may require redoing. Staff leader oral and written comments, as well as consultation, provide adequate opportunities for feedback concerning oral briefings. As with written products, a briefing may require repetition.

### Formal Evaluation

The student receives two interim and a final evaluation during Phase II from his/her staff leader. These evaluations focus on the ability to analyze and solve military problems, communicate effectively, and the ability to interact and coordinate as a member of a staff. The first interim evaluation occurs immediately following the staff techniques exercise (2 weeks after the course begins) and provides an initial assessment of each student's strengths and weaknesses. The second interim evaluation occurs upon completion of the budget exercise (5 weeks). This evaluation summarizes each student's demonstrated performance during the logistics, training, and budget exercises. The final evaluation is similar to the two interim evaluations and occurs at the close of the European scenario exercise. The intent of these evaluations is to let each student know how the staff leader rates the student's performance to date, and the progress toward attaining the terminal course objectives.

Each interim and final report is composed of two main parts. The first part consists of ratings on the specific objectives of the CAS$^3$ curriculum. The second part is a narrative description of the student's performance and may be used to highlight student strengths, weaknesses, and suggestions for further development. Figure 1 provides an example of a completed final report.

Upon the completion of Phase II, a course report is developed for each student. This course report contains the narrative section from the interim reports, ratings on terminal objectives of the CAS$^3$ curriculum, a narrative summary of the student's performance while attending CAS$^3$, an assigned overall rating of performance for the course, as well as the Commandant's remarks, if any. All ratings are assigned using the system defined below. The course report is retained as the official summary of performance at CAS$^3$.

The rating scheme to be used for evaluation the student performance consists of seven categories, as defined below:

1. Exceeded Course Standards--Superior performance; exceeded expectations.

2. Achieved Course Standards Plus--Acceptable performance; somewhat above expectations.

3. Achieved Course Standards--Acceptable performance; met expectations.

4. Achieved Course Standards Minus--Acceptable performance; slightly below expectations.

5. Marginally Achieved Course Standards--Borderline performance; below expectations; improvement required.

6. Failed to Achieve Course Standards--Unacceptable performance; far below expectations.

7. Not Evaluated--Unobserved performance; no opportunity for evaluation.

In addition, an Academic Evaluation Report (AER) is required for each student's official personnel file. Because of the pilot nature of this course, a waiver was obtained for individual reports and all students were issued indentical reports with a narrative description of the course. Eventually CAS$^3$ hopes to gain permission to replace the standard AER with the Course Report which is unique to the outcomes of CAS$^3$.

## Results

One hundred seventeen Captains and Majors graduated from the first CAS$^3$ course. A summary of their final ratings is provided in table 1.

GUIDE TO GRADING
E Exceeded Course Standards
A Achieved Course Standards
M Marginally Achieved Course Standards
F Failed to Achieve Course Standards
N Not Evaluated

COMBINED ARMS AND SERVICES STAFF SCHOOL

THIRD REPORT
(Preparation for Combat Exercise)
(Mobilization/Deployment Exercise)
(European Scenario Exercise)

COURSE 8101

STUDENT NO: 999
RANK: CPT
NAME: John Doe
BRANCH: IN

PART 1 - COURSE PERFORMANCE

| Subject and Level of Performance | Grade | Subject and Level of Performance | Grade |
|---|---|---|---|
| 1. Demonstrated the ability to analyze a corps OPLAN (Mission Analysis). | N | 10. Demonstrated the ability to perform staff duties in the preparations of staff estimates for a division defensive/offensive mission. | A+ |
| 2. Demonstrated the ability to write a division warning order. | N | | |
| 3. Demonstrated the ability to prepare a division staff estimate (G-1/G-2/G-3/G-4) given a division mission and a Corps OPLAN. | A | 11. Demonstrated the ability to perform staff duties in the development of a defensive/offensive OPORD with supporting annexes and overlap. | A+ |
| 4. Demonstrated the ability to write a defensive OPLAN and prepare one or more supporting annexes. | A | 12. Demonstrated the ability to perform staff duties in the execution of a defensive/offensive OPORD during a command post exercise. | A |
| 5. Demonstrated the ability to perform staff duties in the development of a plan of action for the mobilization of a reserve component unit. | A+ | 13. Demonstrated the ability to present ideas orally on military subjects clearly and concisely. | A- |
| 6. Demonstrated the ability to perform staff duties in the development of advance party and arrival plans for a reserve component unit. | N | 14. Demonstrated the ability to employ quantitative decision methods where appropriate. | N |
| 7. Demonstrated the ability to perform staff duties in the development of a closure plan for a mobilized reserve component unit. | A | 15. Demonstrated an attitude of professionalism in the performance of exercise requirements. | A+ |
| 8. Demonstrated the ability to perform staff duties in the preparation of a plan of action to bring the division to readiness condition C-1. | N | 16. Demonstrated the ability to interact with peers to develop/coordinate a staff product. | A+ |
| 9. Demonstrated the ability to perform staff duties in developing a staff estimate for the deployment of a division from mobilization station to the tactical area of operation. | A | 17. Demonstrated the ability to clearly and concisely present written ideas regarding military subjects. | A- |

PART 2 - NARRATIVE

CPT Doe achieved course standards in all areas. He is a totally professional officer and this is evident in his attitude as well as in his work. He is a total team player who interacts well with his peers. He showed a firm grasp of the decision making process in the preparation of estimates/OPLANS & OPOR's at the division level. He needs to continue to work on organizing and clearly presenting oral and written products.

PART 3 - STUDENT'S SIGNATURE

| Date | Name | Rank |
|---|---|---|
| 24 June 81 | John Doe | CPT |

PART 4 - STAFF LEADER'S SIGNATURE

| Date | Name | Rank |
|---|---|---|
| 24 June 81 | John Smith | LTC |

Figure 1. Completed final report.

Table 1. Course ratings for the first CAS3 class.

| Category | Number | Percent |
|---|---|---|
| Exceeded Course Standard | 11 | 9.4 |
| Achieved Course Standard Plus | 41 | 35.0 |
| Achieved Course Standard | 51 | 43.6 |
| Achieved Course Standard Minus | 10 | 8.5 |
| Marginally Achieved Course Standard | 4 | 3.4 |
| Failed to Achieve Course Standard | 0 | 0.0 |

At the end of Phase II the students were queried about the Phase II evaluation system. Remarkably, 60 percent responded favorably. Only 15 percent were dissatisfied enough with the evaluation system to respond in a completely negative manner. Of the remaining 25 percent, 5 percent were indecisive and 20 percent were deemed favorable with reservations. Several students perceived the standards as fuzzy or nonexistent, but regarded the evaluation as fair and equitable. Others shared this opinion for the evaluation taking place within their staff group, but indicated they questioned whether the evaluation was fair and equitable across groups because of different staff leader emphases. One student's thoughts were particularly appropriate: "The evaluation was subjective but far better than a series of objective tests." Another student commented that the varying backgrounds preclude any completely fair and equitable evaluation but the current system is acceptable as long as the course remains essentially pass/fail. In conclusion, it appears the students perceived the present system as acceptable/favorable when compared to the possible alternatives. As several students indicated, positive steps, such as further staff leader training and staff leader group sessions reviewing previously assigned grades, should be undertaken to insure consistent evaluation across groups.

Additionally, the widespread assignment of "achieved course standards" or worse, the detailed and specific recommendations, as well as the frequent requirement to redo papers at the beginning of the course seemed to produce the following results. The fact that everyone was receiving severe criticism and receiving approximately the same grades began to foster cooperation rather than competition within the staff groups. After the staff leader set the standard for excellence, peer pressure to do well reinforced this standard. Because the evaluation process lacked absolute standards, the staff leader was able to raise his expectations for individuals as the course progressed. Thus, the course remained a challenge for each student until completion.

## PHASE II COURSE EVALUATION

Since this was the pilot course of CAS$^3$, naturally the course evaluation focused on formative rather than summative evaluation issues. That is, the purpose of the course evaluation was to assess the strengths and weaknesses of the curriculum to accomplish the stated objectives and provide the basis for curriculum revision as necessary. The evaluation emphasized the course materials, course contents, instructional techniques, and student evaluation system. Primarily, the information gathered from the staff leaders and students themselves formed the basis for the evaluation although comments from the authors, administrative personnel, and student records supplemented the primary sources.

### Information Sources

As each student proceeded through Phase II, the student was asked to maintain a file of comments on an exercise questionnaire. Standard questions regarding Phase I preparation, instructional design, and exercise relevancy were asked. Additionally, space was provided and general questions asked to obtain information regarding unclear assignments, directions, written materials, course contents, etc. The completed questionnaires were forwarded to the exercise authors for review.

As each staff leader proceeded through Phase II, the staff leader was asked to maintain a file of comments on each exercise. These comments reflected teaching difficulties and suggested improvements. Upon course completion the staff leaders met and assembled a master list of comments for each exercise.

To examine the grading standards across staff groups, the top two and bottom two papers of each written assignment were reviewed by the director. Also, the director frequently visited the staff group rooms and monitored oral presentations and subsequently, the staff leader feedback. Reports of all staff leader ratings on the interim, final, and course evaluations were generated for the director as the course progressed.

Upon completion of Phase II, each student was administered a questionnaire of open-ended questions to assess the student's feelings toward the CAS$^3$ experience. These comments were categorized according to their degree of favorableness toward the topic and then tabulated.

### Results

Once the staff leaders had amassed their comments from each exercise and the authors had reviewed the student's exercise comments, the staff leaders and authors met to refine the curricular contents. Principally, the changes focused on the elimination of identified inconsistencies, the elimination of nonessential materials, the inclusion of new materials, the spacing of products requiring staff leader evaluation, and the resequencing of activities.

No major discrepancies in the staff leader grading standards resulted from the examination of this area. Some statistical comparisons did result in statistical significance, but the practical significance was not deemed large enough to require the directors intervention. Additionally, no student complaints were registered at the director level in this area. The high degree of staff leader consistency is hypothesized to result from the common experiences of the staff leaders (all LTCs and all but one former battalion commanders), their staff leader training, and their understanding of the evaluation philosophy.

Based upon the student end-of-course comments, the following general conclusions became evident.

1. Approximately 80 percent of the students stated that the curriculum was relevant for their branch as designed or with minor modifications. Additionally, 85 percent concluded the program is worth the monetary cost associated with implementation.

2. The strengths of CAS$^3$ are the staff group concept, the association of contempories from differing branches, and the experienced staff leaders.

3. The course was particularly effective in providing feedback to students and instilling confidence in their abilities to brief, write, and defend solutions.

4. The performance oriented student evaluation system used during Phase II was well received by a majority of the students.

5. Several students reported that attendance at CAS$^3$ instilled positive attitudes toward the Army and its training system.

6. The after action critiques revealed a tendency for students to focus their attentions on specific curriculum contents rather than on the learning of staff processes and procedures.

7. Phase I is marginally required in its current form and as presented as a prerequisite for Phase II. However, many students indicated it served as a good refresher course and provided valuable professional development information.

8. The learning objectives were not sufficiently integrated into the curriculum in a manner that was useful to students.

9. For the Phase II curriculum, students indicated that the current staff leader-centered instruction is more effective than instruction that could be received primarily from subject matter experts.

10. The pilot CAS$^3$ program did not contain an overriding major flaw that severely restricted learning.

AD P001430

# USING THE SYSTEMS APPROACH TO DEVELOP THE CAS$^3$ CURRICULUM

Dennis A. DeFrain, LTC
Faculty Development Officer/Instructional Technologist
U.S. Army Command and General Staff College

The U.S. Army Command and General Staff College (CGSC) at Fort Leavenworth, Kansas recently implemented an Accountable Instructional System based on the Florida State ISD model. This system was used to develop the curriculum for the new Combined Arms and Services Staff School (CAS$^3$), a part of CGSC. Implementation of Phases II (Design), and III (Develop) of this model are discussed. Specific topics include development and approval of learning objectives, the General Officer subcourse review, and the development of evaluation instruments for the non-resident phase of the CAS$^3$ curriculum. Finally, the advantages of using the systems approach and lessons learned are addressed.

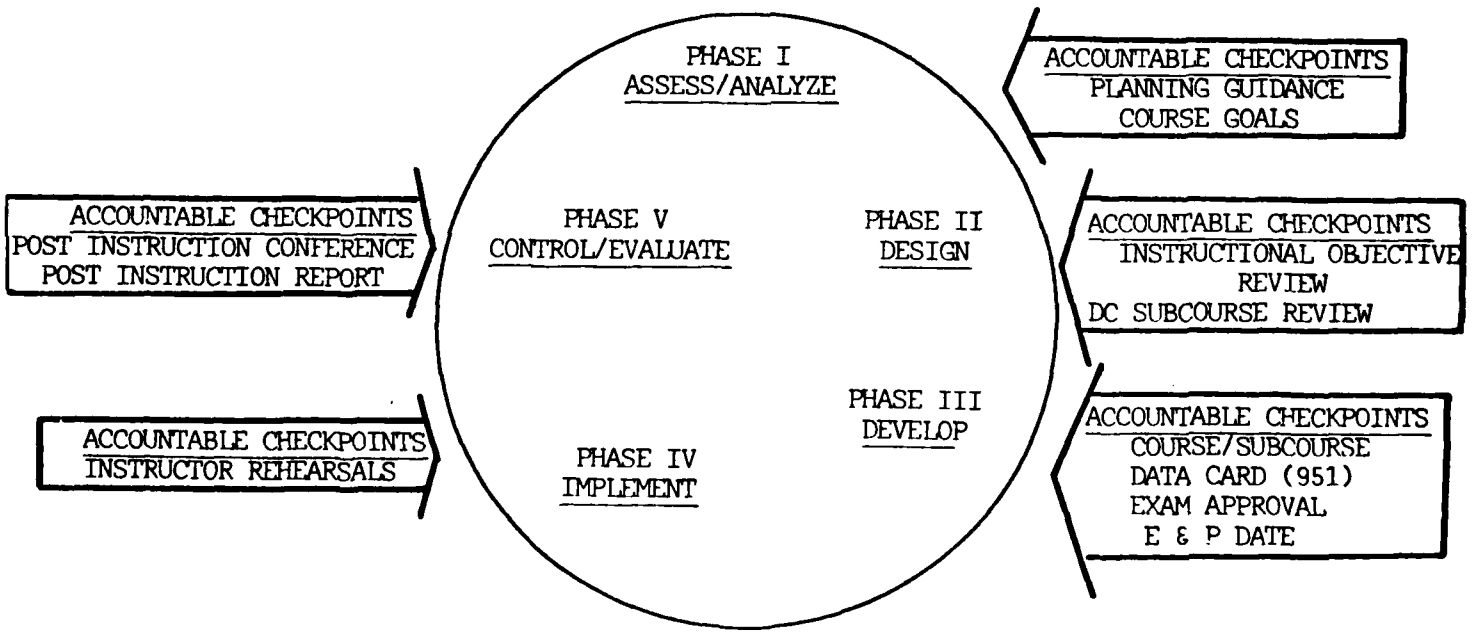
## USING THE SYSTEMS APPROACH TO DEVELOP THE CAS$^3$ CURRICULUM

The U.S. Army Command and General Staff College at Fort Leavenworth, Kansas has implemented within the past two years a five-phased accountable instructional system based upon the Instructional Systems Development (ISD) model as developed by Florida State University for the Department of Defense. This paper will discuss the CGSC system in broad terms and will show how it was used to develop the CAS$^3$ curriculum. Primary emphasis will be placed on the Design Phase (Phase II), and the Develop Phase (Phase III). Other papers will discuss the analysis, implement, and evaluate phases of the system. In discussing the design and develop phases, four major areas will be addressed. They are: Learning objectives, evaluation instruments for the non-resident phase of instruction, the checklist used for evaluating the test instruments, and the subcourse review.

### CGSC Accountable Instructional System (AIS)

The CGSC Accountable Instructional System was implemented in early 1980 because of the perceived need to manage course content in a more systematic manner. It was not designed to make drastic changes in the curriculum processes, in fact many of the steps in the model were already in effect and had been so for many years. The primary changes which the AIS addressed were the regular, periodic checks made during the instructional development process. These checks, called accountable checkpoints appear in each phase of the system and provide feedback to the decision makers at times when changes and improvements can be made. The following chart shows the CGSC model and the accountable checkpoints for each phase of the system.

COMMAND AND GENERAL STAFF COLLEGE ACCOUNTABLE INSTRUCTIONAL SYSTEM



## Learning Objectives

The first steps in Phase II (Design) call for the writing of Terminal Learning Objectives and Enabling Learning Objectives. CGSC used Terminal Learning Objectives as final outcomes that students were expected to master as a result of receiving instruction. Enabling Learning Objectives were intermediate objectives which were to be mastered prior to achieving the Terminal Learning Objective. CAS$^3$ authors wrote Terminal Learning Objectives once they had completed their subject matter research and had determined which knowledges and performances they would teach. Enabling Learning Objectives were then developed. As the learning objectives were being developed, the CAS$^3$ authors were assisted by an instructional technologist from the Office of Curriculum Assistance (OCA), an agency within CGSC, but not a part of the CAS$^3$ organization. This assistance was capped by a final formal review of each objective, the first AIS accountable checkpoint. The review insured that all objectives were in the CGSC standardized format of Task, Conditions, and Standards. The task statements were reviewed for clarity and to insure that they reflected desired learning outcomes and specified only one desired performance. The conditions were checked to determine if they included all necessary descriptions of the environment in which the task was to be performed. The standards were reviewed to insure that they told the student exactly how the task would be evaluated.

Once the formal review was completed, the learning objectives were submitted to the CAS$^3$ director for final approval. Following approval, the learning objectives were subject coded and placed into the CGSC computerized learning objective data base. This data base allowed authors and supervisory personnel to receive computerized listings by subject matter area to check for redundancies. It also provided a readily accessible system for reviewing the learning objectives by those within the CAS$^3$ organization and those outside the organization who had managerial interest in the course, thus providing another form of accountability.

## General Officer Subcourse Review

Once the learning objectives had been developed and the course authors had structured their individual courses, they briefed the course content as planned to a General Officer. He reviewed each course to determine the general direction which the course was taking, to verify consistency with prior guidance, to insure that the proper objectives were being taught, and to review the general evaluation plan. This accountable checkpoint allowed General Officer input early in the development process and precluded changes and revisions at later dates. The subcourse review was followed by detailed subject matter content research, methodology and media selection, and the writing of the evaluation instruments.

## Evaluation Instruments (Non-Resident Phase of Instruction)

Once the learning objectives had been written and approved, the instructional material was developed and the evaluation instruments were written. The material development and the writing of evaluation instruments took place during Phase III (Develop) of the accountable instructional system. For each

of the non-resident modules, a Pre-test and a Post-test was written. As each student began a module, he was required to take the Pre-test and self score it. If he received a score of 90 percent on this test, he was not required to complete that module. If he scored between 75 and 90 percent, it was recommended that he study the material. If he scored below 75 percent, he was required to complete the module. The Post-test was scored in a similar manner and served as a check to insure that the student had mastered the learning objectives. After completing all modules, the students were administered a 180 question qualification examination which tested a sample of all material in the non-resident course.

Evaluation Instrument Checklist

The Pre-tests, Post-tests, and Qualification Examination were submitted to the Office of Curriculum Assistance for approval. This office used a standardized checklist to review each evaluation instrument. This checklist addressed the following: Relationship of learning objectives to test items, weighting, construction of test items, and mechanical features of the test to include such things as item format, scoring arrangements, and distribution of correct responses. The test approval procedures required that a Table of Specifications be filled out by the author to check on the relationship of learning objectives to test items. The following shows a sample Table of Specifications:

TABLE OF SPECIFICATIONS

| Learning Objective | | Level of Achievement and Number of Items | | | | | | Total No. of Items |
|---|---|---|---|---|---|---|---|---|
| No. | Level | KNOW | COMP | APPL | ANAL | SYN | EVAL | |
| A | COMP | | | | | | | |
| A.01 | COMP | 1 | 2 | | | | | 3 |
| A.02 | KNOW | 5 | | | | | | 5 |
| A.03 | COMP | 2 | 2 | | | | | 4 |
| A.04 | COMP | 3 | 3 | | | | | 6 |
| B | COMP | | | | | | | |
| B.01 | COMP | 2 | | | | | | 2 |
| B.02 | COMP | 2 | | | 1 | 1 | | 4 |
| C | COMP | | | | | | | |
| C.01 | COMP | 2 | 2 | | | | | 4 |
| C.02 | COMP | 2 | 2 | | | | | 4 |
| | | | | | | | | |

This sample shows that all test questions are written at cognitive levels similar to those of the learning objectives except for those of learning objective B.02. The two questions which were not similar were required to be rewritten to the proper level.

Across the top of the preceding Table of Specifications chart are Bloom's six cognitive levels of learning, namely: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. The Table of Specifications is used in the following manner: Along the left side in the first column, the cognitive level that the learning objective is written at is indicated. The level that each test item is written at is shown under the appropriate column on the right part of the chart. A review of the chart will quickly show if the test items are written at the same levels as the learning objectives. It if felt that it is proper for a learning objective to be written at a higher cognitive level than a test item, but not vice-versa.

The next step in the test instrument approval process was to insure that the weighting of the test items was in direct proportion to the emphasis placed on the learning objective. This was checked through use of a weighting chart similar to the one shown below:

SAMPLING/WEIGHTING CHART

| TLO/ELO No. | Emphasis Placed on TLO/ELO | SAMPLING-% of items covering this TLO/ELO | WEIGHTING-% of points covering this TLO/ELO |
|---|---|---|---|
| A | | ← No. of questions | |
| A.01 | | 3 | x 2pts = 6 |
| A.02 | 50% | 5 } 18 = 56% | x 2pts = 10 } 30% |
| A.03 | | 4 | x 2pts = 8 |
| A.04 | | 6 | x 1pt = 6 |
| B | | | |
| B.01 | 20% | 2 } 6 = 19% | x 5pts = 10 } 30% |
| B.02 | | 4 | x 5pts = 20 |
| C | | | |
| C.01 | 30% | 4 } 8 = 25% | x 5pts = 20 } 40% |
| C.02 | | 4 | x 5pts = 20 |
| | | 32 | |
| | | | |
| | | | |
| | | | |
| | | | |

In this example, it can be seen that fifty percent of the course dealt with learning objective A. Of the 32 total questions on the examination, 18 of them addressed objective A. If the questions had been weighted equally, the sampling for this objective would have been quite good, however, as seen in the last column, the questions were not weighted equally and the weighting for the examination was not proportional to the emphasis placed on each of the objectives.

In the sample chart on the preceding page, the learning objectives were listed down the column on the left hand side of the form. The emphasis placed on each learning objective as shown in the next column was obtained from a review of the course materials and subjective input from the course author. The sampling column checked the number of items covering each learning objective, and by comparing the sampling column to the data in the second column, test reviewers could readily determine if the weighting was appropriate. If the test items were of unequal weight, the last column would be used to compute the emphasis placed on each objective, and again a quick check of each objective would show whether or not there was appropriate weighting. Each Pre-test and Post-test was reviewed using this process, as was the qualification examination.

In addition to the relationship of learning objectives to test items and weighting, the evaluation instruments were also checked for construction of test items and mechanical features. Each test item stem and answer to include alternatives and distractors when appropriate was checked for form, clarity, grammar, and consistency. Mechanical features which were checked included test format, scoring arrangements, distribution of correct answers, arrangement of items, and directions for answering questions.

## Advantages of the Systems Approach to CAS$^3$ Course Development

There were quite a number of advantages to using the systems approach for development of this course of instruction. Primarily though, the chief advantages included logic, accountability, and standardization. First, this approach used a logical method of course development starting with development of achievable course goals. Learning objectives were written to support the course goals and all course materials were developed around the learning objectives. Each course author knew exactly where he was and where he was going at all times in the development process. Accountability, as has already been discussed, was achieved through regular and periodic checks as the instruction was being developed. Standardization was achieved by the requirement to have learning objectives in a prescribed format and by having evaluation instruments and other course materials constructed in like manners.

## Lessons Learned

1. After the course was administered once, it was determined that some of the subject matter needed to be sequenced differently. Because the subject matter was written around learning objectives, it was easy to move it from one course to another without major revision or rewrite.

2. Course authors at first felt that the systems approach was cumbersome and not needed, however, after the course was given, most, if not all saw the advantages and are now supporters of this approach.

3. The use of learning objectives was more of an advantage to the authors than to the students. Authors were able to easily develop course materials once they had determined exactly where they were going through the development of learning objectives. Learning objectives, were in some cases over used in that too many of them were developed. Some learning objectives were simply teaching points and actually tended to hinder students who were trying to study and understand all of them.

4. The writing of learning objectives for the resident instruction phase was much easier than for the non-resident instruction. This was because higher cognitive levels of learning could be addressed much more easily in the resident course. The non-resident course dealt primarily with the lower cognitive levels of Knowledge and Comprehension and provided a real challenge to authors to come up with challenging, meaningful objectives.

Summary

The use of a systems approach to the development of CAS$^3$ instruction provided a logical, standardized approach which insured accountability at each developmental step.

AD P001431

# CREATING A NEW CURRICULUM TO TRAIN ARMY STAFF OFFICERS

by

Ralph W. Ekwall Ed. D
Educational Specialist
CGSC, Ft. Leavenworth, Kansas

This paper begins with explanations and definitions needed to under-
tand the remainder of the paper. The major topic is an explanation of
ıe process by which the US Army developed a new program to train US Army
taff officers. A step by step chronology of the curriculum development
rocess is explained. The major emphasis is on the methodology used to devel-
ɔ the curriculum, but a topical summary of the content is included. Follow-
ng the chronology, some specific topics such as strategy for implementation,
lanning activities, personnel, and other topics are discussed. The final
ection is a summary of lessons learned which may have application to other
urriculum developers.

## DEFINITIONS AND EXPLANATIONS

1. BACKGROUND. To begin this presentation several definitions and explanations will be given to enable you to better understand the content. In June of 1978 the U.S. Army completed a study called the RETO study. RETO stands for Review of Education and Training of Officers. This study identified a shortfall in staff officer training and recommended the creation of a new school to train U.S. Army staff officers. The new program was developed by an independent cell of planners within the Command and General Staff College at Fort Leavenworth, Kansas. The name of the new school is $CAS^3$ which stands for the Combined Arms and Services Staff School. This program is designed to train Army staff officers in the common or generic skills rather than in specific skills. Stated another way: $CAS^3$ is not in the business of training logisticians or intelligence specialists: it is in the business of training staff officers.

2. LENGTH AND STRUCTURE. The $CAS^3$ program consists of a 140 hour package of nonresident instruction, a qualifying exam, and a nine week period of resident instruction. The nonresident portion is also called phase I; the resident portion is called phase II.

3. PROGRESS. The first iteration of $CAS^3$ was completed in June, 1981. In 1981, about 120 officers received $CAS^3$ training. In 1982 about 720 officers will be trained; in 1983 and 1984 about 1000 officers will receive training. When fully implemented in 1985 and the years following about 3800 officers per year will receive $CAS^3$ training. At full implementation nearly every Army officer will receive $CAS^3$ training. Most of the $CAS^3$ students will be captains in their 6th to 9th year of service.

The plan of the remainder of this presentation is to go through the major steps in the curriculum development process with a digression on the content of the curriculum and another digression of the characteristics of the two phases. Following the chronology there will be an explanation of some topics you may want to know about and some things that we want to tell you about.

## STEPS IN CURRICULUM DEVELOPMENT

A review of the chronology shows that the curriculum development process was condensed into a relatively short period of time. The major portion of the curriculum development was completed in about 18 months. The amount of time allocated for program development was specified by the TRADOC commander. Intensive efforts were required to complete the job.

A systems approach was used in the curriculum development process. The model used in the initial stages was the general ISD model. The CGSC version of ISD was developed during the $CAS^3$ curriculum development process and in the later stages it served as the systems model for curriculum development.

The first part of the curriculum development process was the Analysis process. At this stage of the process, resources were limited and CAS$^3$ was under some time pressure, even so, staff members were able to do reasonably good job analysis of the work of an Army staff officer. A major effort at this time was the development of a task list. The sources of the task list were sources such as the following:

a. Data from the RETO study
b. Panels of staff officers
c. Command guidance
d. CGSC departmental input
e. Borrowing from other job analysis programs
f. Research and borrowing from other research

The general policy during the task list development stage was to include every suggested task without regard for duplication, level of specificity, scope or language. At one point the list contained about 400 tasks. After collecting task data, the task list was edited and refined. Duplicate tasks were eliminated; tasks judged to be more appropriate for enlisted men were eliminated; tasks written at very high levels of specificity were subsumed under other tasks or eliminated; tasks were edited so that they were stated in similar fashion and stated in behavioral terms. At the conclusion of the editing and refinement process the task list contained 66 tasks and 13 skills and knowledge statements. Examples of tasks: Formulate command operating budget; Develop a plan for employment of electronic warfare assets; Prepare a staff study. Example of skills and knowledge statements: Principles of management; Capabilities and limitations of Soviet weapons systems. The refined task list was used as a basis of a survey of officers attending CGSC who had served as staff officers. Each task was the basis of 4 questions. (1) How much time do you spend performing this task? (2) How often do you do this task? (3) How difficult is it to learn to perform this task? (4) What are the consequences of inadequate performance of this task? About 270 responses were analyzed. Our analysis of the survey data was the basis for our recommendation to the Critical Task Selection Board. The Critical Task Selection Board was a board consisting of 4 generals and the director of CAS$^3$. For the most part the Critical Task Selection Board approved the recommendations.

At this time the first steps in the design process began. Based on the survey data and the philosophical concept developed by the CAS$^3$ task force, a notional resident phase was created. A series of 7 exercises were planned. These exercises were planned so that the recommended tasks would be taught in the exercise. After the phase II exercises were planned, the phase I modules needed to support phase II were designed. This is probably a good time to look at the content of the curriculum.

Some of the material in the nonresident phase serves as a support for the resident phase, but other portions of the material contain general information useful for general professional development.

The phase I module is designed to be done on an individual basis in a self-paced mode. It provides instruction by means of several varieties of programed learning materials. It is designed to be administered as a correspondence

program. Nearly all of the learning objectives are at the knowledge or comprehension level. The intent of phase I is to provide students with a common base of knowledge when they begin phase II. Phase II instruction is presented in 12 person staff groups guided by a senior lieutenant-colonel with battalion command experience. The learning activities in phase II are designed in such a way that interaction and coordination among participants is required. The learning objectives in phase II are those requiring a higher level of mental activity. Students are evaluated on an individual basis for the purpose of improvement; the emphasis is on skills rather than content. Phase II was designed so that the student would perform in much the same way as a real world staff situation. The student would do things that staff officers do; they would prepare staff studies; they would coordinate actions; they would interact, cooperate and exchange information; they would have an opportunity to make real world decisions without paying real world penalties for mistakes.

After phase I and phase II initial planning was completed, authors were assigned and trained. The first task of the authors was to develop TLOs and ELOs. Teams of Authors, with 1-3 members, corresponding to the Phase II exercise were formed. Senior authors for each team were appointed. Each team was responsible for the preparation of the phase II exercise and the phase I modules that supported it. A concurrent task was planning and presenting a detailed briefing of the planned content of each exercise. This briefing was presented to the director and the entire staff for comment. After suggestions for revision had been incorporated, curriculum writing began with phase I modules being done first followed by phase II exercises.

During the time that phase II exercises were being written, staff leaders were trained. Final work on curriculum writing was completed during the first iteration of the instructional program. Following completion of the first iteration, we have done some curriculum revision and we are now in the process of training additional staff leaders. This completes the chronology of curriculum development; your attention is directed to some special topics.

## SPECIAL TOPICS

1. STRATEGY FOR IMPLEMENTATION. One of the design concepts for $CAS^3$ was that phase II instruction would consist of instructional methods and materials that utilized simulations and/or practical exercises, case studies, discussion, and problem solving materials. To implement this concept the concept was presented to newly arrived authors or staff leaders at their in-briefing. The methods were modeled and promoted in author training. Authors were actively encouraged to prepare instructional materials that fitted the $CAS^3$philosophy of instruction. When staff leaders were trained, the methods were modeled and promoted and staff leaders were trained in their use. In this manner a curriculum was developed that conformed to our philosophical concept. The same method was used to implement the phase I-phase II concept and other aspects of the program.

2. SINGLE THREAD CONCEPT. The exercises in phase II are linked together by a single thread. This means that all of the phase II learning activities have

have a single setting. That setting is the 52nd division. In that setting students are trained as staff officers and then do staff work in areas such as budget and training; a roundout brigade is mobilized and the division is deployed to Europe. The final exercise is a combat exercise in Europe. Through all seven exercises there is a single framework, the 52nd division; this provides a common focus even though the situation changes with each exercise.

3. PLANNING ACTIVITIES. At the inception of the CAS$^3$ program, the TRADOC, commander provided direction to CAS$^3$ planners, and having given that direction, then allowed the CAS$^3$ staff to do their job. Planning activities were necessary to accomplish the planned goals.

Throughout the entire program development process, planning activities helped to guide, motivate, and provide priorities for program and curriculum development. The leadership of CAS$^3$ provided overall planning guidance and prepared milestone charts which were revised from time to time. Team leaders and individuals developed their own planning charts within the framework of the milestone charts prepared by CAS$^3$ leadership. At one point a PERT chart was developed, but CAS$^3$ lost the services of the officer who had developed it. Since there was no individual responsible for revising or updating the PERT chart, the program continued to use milestone charts which proved to be adequate.

4. PERSONNEL. You may wish to know more about the people who planned and developed the CAS$^3$ program. There are two full-time civilian employees on the professional staff. One works mostly in the area of staff development and the other in evaluation. Most of the work in planning and curriculum development has been done by a staff that consists of mostly lieutenant colonels and a few majors. These officers were a select group, but had no specific training as authors or in subject matter areas. By means of the CAS$^3$ author training program and intensive self-study they became experts in writing and in their subject matter areas. All of the instruction was provided by lieutenant-colonels who were, for the most part, former battalion commanders. These instructors, or Staff Leaders, were provided with an intensive training program to give the knowledge and skills necessary to function as Staff Leaders. The director of the program is a full colonel.

5. PROMOTING THE IMAGE. The CAS$^3$ staff undertook a series of actions designed to promote the image of the CAS$^3$ program. Since CAS$^3$ functions as a department at the Command and General Staff College at Fort Leavenworth, some of these actions were directed to the Command and General Staff College. CAS$^3$ used the other departments at CGSC in an advisory and consultive capacity. This provided access to a pool of expertise, developed some personal relationships and improved our standing. On a larger basis, CAS$^3$ provided status briefings for the commandant and deputy commandant, for visiting generals and other VIPS, for National Guard and USAR groups, for regular students at CGSC, and to various special groups. The use of general officers on the Critical Task Selection Board helped to promote our image. Our program maintained a constant liaison with TRADOC. A POI was submitted to Army branch schools for review and comment; experts were invited to review and comment and on our work. Not every one of these initiatives were undertaken for the purpose of promoting the image, but all served to improve our standing.

6. NEW PROGRAM. The CAS$^3$ curriculum is unique because it is an entirely new program rather than a revision of an existing program. This meant that the innovative aspects of the program could be implemented rather easily. However, the writing of a new curriculum created some problems. No single text, field manual, or existing curriculum contained the necessary materials; therefore the writing required extensive research and extraordinary creativity. In certain areas, the authors found that a doctrinal base was inadequate or lacking. CAS$^3$ authors managed by use of the following: adopt, adapt, borrow, modernize, and collect advice from proponent agencies.

## LESSONS LEARNED

Finally, some lessons learned that you may find useful.

1. Quality leadership is essential for the success of a new program.

2. Careful planning and adherence to plans is an essential ingredient of success, but plans need to be continuously reviewed and revised.

3. A systems approach to curriculum development provides the needed framework for building a new program.

4. Promotional activities are an important ingredient of a successful program.

5. Control of a program is a necessary ingredient of success. In our case we were able to control things like the philosophy, methodology, author training, staff leader training, most of the content, the structure of the curriculum, and the administration of the program.

6. Defend the program. While we incorporated command guidance, we knew what we wanted to do and refused advice and suggestions that were contrary to our plans.

7. Preparation. Prepare carefully for briefings; make sure that plans are carefully thought out previous to important briefings. Have positive plans or positions prepared. Be ready to say "We plan to do..." or our position is that it should be done in this manner..." Don't be caught in the position of not knowing what you plan to do.

This is only a partial summary of lessons learned; other speakers will add to the lessons learned.

AD P001432

# The Combined Arms and Services Staff School

## From the Perspective of an Author and Instructor

By

Jon C. Stillman, LTC

Author/Instructor

US Army Command and General Staff College

Fort Leavenworth, Kansas 66027

SUMMARY

LTC Stillman was one of the early arrivals to the Combined Arms Services Staff School. He was one of the original course authors as well as one of the original instructors for the pilot course. This paper highlights key factors which assisted LTC Stillman in preparing for and teaching during the pilot course of the Combined Arms and Services Staff School. Initially, author training and experiences are discussed. This is followed by selected highlights of the teaching experience. The paper concludes with a series of lessons learned that could be incorporated in the development of similar courses in the future.

BACKGROUND

Following arrival in July of 1980 LTC Sullivan was assigned as the author of the Training Management Instruction for the Combined Arms Services Staff School. At this time in the curriculum development process looking at the CGSC Accountable Instruction System Model, (Figure 1), Phase I, the analysis, had been completed and the task lists identifying those areas for which instruction had to be prepared had been approved. Course goals had also been specified. The approved tasks had then been assigned to seven writing teams each team responsible for an exercise or block of instruction. Training Management Exercise was one of the seven exercises. The first two weeks after arrival were focused on collecting appropriate resource materials for use later. In late July, author training commenced for a number of newly arrived authors.

AUTHOR TRAINING

Most personnel assigned to write the course had no experience as authors and therefore author training was essential. Should the experience level of authors been greater, it would have still been necessary for an author training program in that the systems approach to writing, in conjunction with the need to develop instruction for use in small group training, required different author skills than those normally applied in Army institutional training systems. In order to limit this paper, selected areas of author training deemed useful will be highlighted.

Accountable Instruction System

The first area to be highlighted is the instruction which provided the authors with an introduction to the CGSC Accountable Instruction System Model (Figure 1). An awareness of this model provided perspective from the beginning.

COMMAND AND GENERAL STAFF COLLEGE ACCOUNTABLE INSTRUCTIONAL SYSTEM

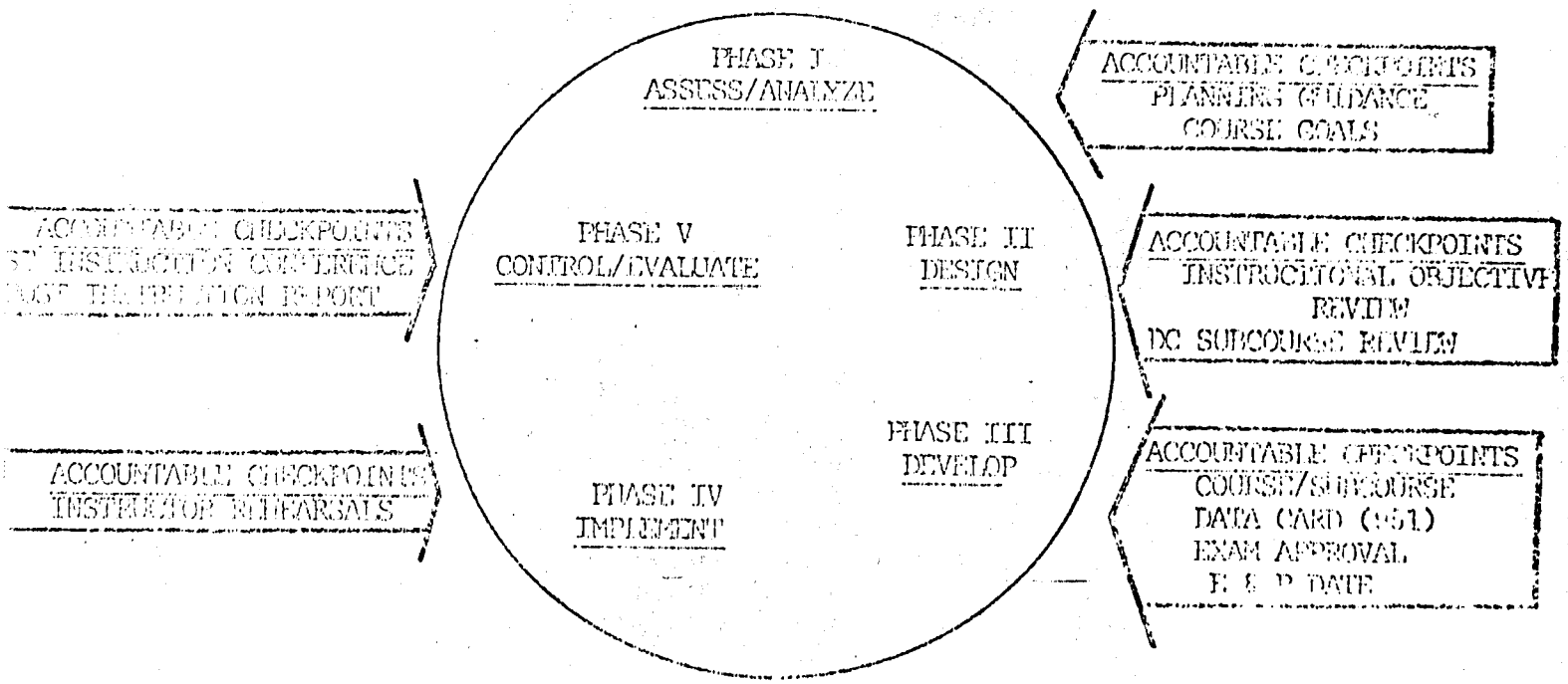



Figure 1



BEST AVAILABLE COPY

The author understood that the instruction he would design (Phase II) and develop (Phase III) would be the precise instruction for which students would be evaluated (Phase V). Further, the author understood that an assessment would be made (Phase I) and redesign (Phase II) would occur as required. An understanding of this system provided great motivation to ensure the course material supported the training needs defined by the task selection board. Further, understanding the model and the interrelationship among each of the phases of the model allowed the author to make writing decisions from the systems perspective. For example, when developing a learning objective, (Phase II) the author could make an initial assessment of how that objective might be evaluated (Phase V) and then use this information in developing the materials (Phase III). The alternative would have been to think one step at a time with the potential for developing course materials which in fact did not support the course objectives and subsequent evaluation instruments which supported course materials that had deviated from the objectives.

Small Group Orientation

Another important aspect of the training was related to small group orientation. This training encompassed information about characteristics of small groups, how groups organize both formally and informally, how they develop norms, the relationship between the size of the group and group interaction along with instruction on other aspects of the group. This was especially useful in orienting the author towards teaching methods that would best serve the group forum. The case study was identified as an example of a method of instruction that is well suited for small groups. During training, authors in fact participated as members of a group, each author having different information with the group, having to work together in order to come up with an acceptable, common solution. This created an awareness in the authors of the effectiveness of the case study approach. It also provided an example of how materials might be developed in order to cause student coordination and interaction.

Blooms Taxonomy

Finally, authors were introduced to Blooms Taxonomy of Learning. This was critical to the writing effort as authors had to develop instructional materials designed to help students acquire basic knowledge initially so they would have the capability of analyzing and synthesizing challenging problems later in the course.

Miscellaneous

Although only three areas of author training were highlighted, several other aspects of the training were useful. The mechanics of processing materials from the authors desk, through editing, through printing and into the classroom was valuable. An introduction to classroom configuration and audio visual equipment available in the classroom was also valuable. The point is that had author training not occurred, the task of writing the course would have been considerably more difficult.

AUTHOR EXPERIENCE

Learning Objectives

Following author training the task of writing commenced. Having had the opportunity to develop example learning objectives during author training, the task did not seem potentially difficult. As a result an early suspense was established to complete development of learning objectives. No author met this suspense. The reason for this is that before meaningful learning objectives can be developed, basic research must be nearly completed. As a result, the writing effort appeared behind schedule from the beginning. The Training Management Objectives were developed between the end of the first week in August and the end of the first week in November. The learning objectives supported 5 hours of non-resident instruction and 39 hours of resident instruction.

In retrospect, early development of learning objectives was useful as it served as an organizer for developing the course materials. In at least one instance, before the author began writing course materials, he placed each learning objective on a separate piece of paper and organized them in a logical sequence of terminal learning objectives and enabling learning objectives on a tack board. He then juggled the objectives to establish a logical flow with easier objectives occurring earlier and more difficult ones later in the flow. Next he developed a story line incorporating the objectives in conjunction with the flow diagram. This resulted in a logical road map for development of course materials and provided logic and direction for the writing effort.

Single Thread Concept

The intent of the single thread concept was to tie all seven exercises in the resident portion of the course to a common time line, common organization, and a common setting. From the authors perspective this was difficult during the first iteration for a number of reasons.

Authors were not yet assigned when the initial sequencing of the course was established and as a result did not share in the rationale for establishment of the single thread. There was also some confusion regarding the definition of a single thread. Did single thread mean that each writing team used similar staff products for purposes of evaluation (briefings, letters, memorandums, etc.)? Did it mean a common scenario? Did it mean a common time line? Single thread now means all of these. Additionally, the sole author in one key area retired unexpectedly and for a period no author was available to coordinate single thread in this area. Also there was a feeling of apprehension on the part of authors based on management's initial incorrect assessment of time required to develop learning objectives. These factors resulted in deviation from the single thread in some instances. Feeling time pressure in getting ready for the April course, authors did not coordinate as thoroughly as necessary to insure a common time line. Additionally authors did not assess where the portion of instruction for which they were responsible would best fit in the overall time sequencing until late in the de-



BEST AVAILABLE COPY

velopment of their materials. For example, both the budget formulation and training management cycles require key activity in the spring of the year. This became evident during development of the courses and as a result the respective scenarios had spring activity but there was no common transition between the two.

The notion of writing to support a common thread concept is viable and during the current rewrite of the course deviations from the common thread found in the first iteration, have been corrected. Before the rewrite began all authors sat in conference and discussed settings for respective exercises, identity of the organizations to be used and established a common course time line and scenario. Initial review, subject to trial in the classroom, indicates that the rewrite has successfully incorporated the single thread concept.

Process versus Content

From the authors perspective the notion of concern for process and procedures versus content was challenging. The intent of the course was to have the student gain as much learning as possible from doing. Therefore, instead of telling the student "you must coordinate staff papers" (content) the orientation was on having the student prepare a paper integrating information gained from others and then actually coordinating the paper with outside sources (process and/or procedure). In order to develop situations where process and procedure were incorporated took imagination and creativity on the part of the authors. Frequently, it was necessary for authors to write from several perspectives in order to develop a situation adequately. For example, in developing a situation the author may write the situation from the perspective of the administrative officer. This may be followed by writing about the same situation from the perspective of the operations officer. This would continue until all perspectives necessary were created with each student having partial information with which to solve a problem. Writing in this manner could cause students to interact and to use the coordination process in resolving problems but was difficult from the authors point of view. Instead of providing one document with all the information, the author would provide from three to twelve documents in developing a given situation depending on the group structure the author selected for resolving the problem. He in essence would write a puzzle for the students to solve.

Finally, when the student materials were developed, the author was required to prepare an instructors guide to insure that instructors would satisfy objectives intended by the author. As a result, the writing effort required in developing a course with focus on process and procedure was more difficult and time consuming than it would have been if the focus would have been solely on content.

Printing Procedures

The final action required of the author was to ensure that materials were in fact ready for the classroom. Materials had to be typed and edited

prior to being printed. This was time consuming in that the typing workload became excessive as the course dates neared. Further, the pilot course overburdened during the same period. This resulted in a critical [illegible] and meant that some materials were available for the classroom only one day before the presentation date. This created concern on the part of the instructors. Even though instructors had draft copies of material they would have preferred time to review the final printed materials prior to introducing them in the classroom.

Once the materials were produced, the success of the course was the responsibility of the instructors. Key factors that contributed to the success of the pilot course from the instructors perspective will be discussed next.

## INSTRUCTOR PERSPECTIVE

### Small Group

Staff groups were composed of twelve students and an instructor. The instructor remained with the same twelve students for the nine week course and was responsible for all instruction.

The small group orientation provided a number of advantages. Firstly, the group size was manageable from the instructors perspective. Corrected papers could be returned to the student within a reasonable time, all students could be effectively brought into discussions, and focus could be directed toward voids in experience or learning identified by and within the group. Secondly, group team building occurred that allowed for the group to contribute significantly to the growth and education of individual team members. It was the norm for strong members of a group to pull weak members along even when the strong member was being challenged. Thirdly, the instructor could become thoroughly familiar with each student and thereby provide effective evaluation.

### Branch Diversification

Each group was composed of representatives of several different branches of the Army. Therefore, the group contained a broad base of expertise and experience. The result was that branch representatives became recognized as the authority in their related areas. For example, the quartermaster officer was the group's authority on logistical matters, the adjutant general corps officer the authority on administrative matters, the combat arms officer the authority on tactical matters, etc. This seemed to provide motivation to branch experts to study branch related material so that credibility as an expert would continue to exist. In the final analysis, one of the great strengths of the course was the degree and quality of peer learning that occurred.

### Instructor Credentials

Instructors for the pilot course were lieutenant colonels (two grades senior to the students) and past battalion commanders. This had the effect

of establishing immediate credibility with the student body which is critical to the success of the initial course. A high experience level is required due to the nature of the course. Firstly, the student body is composed of experts in functional areas and secondly the first day the instructor is the teacher for the entire 200 hour curriculum. Given these factors, the course would lose credibility should the instructor be inexperienced.

Intensity

The course was considered a very intense course by the students and the instructors. Students were given problems to solve and were released to solve them within a limited time frame. The standard for solving the problem was the standard encouraged by the instructor but enforced by group norms. All night sessions occurred early in the course and were normally inefficient. Leadership struggles would occur, the student with expertise in an area was not believed by the rest of the group, no milestone schedule was established etc. As the group matured many of these barriers disappeared. Leaders would either be assigned and the group would support the assigned leader or the group would move the best qualified person for the problem into the leadership role. Students with expertise were recognized and assigned responsibility for providing correct information in their area of expertise. Milestone schedules were established. The group became efficient, in fact so efficient that by the end of the nine week course the group was capable of solving what were originally all night problems in a few hours. The intensity of the course contributed to the need for efficiency by the group and seemed to act as a catalyst in the groups development of efficiency.

Evaluation System

The evaluation system employed was key to instructor effectiveness. The instructor was "the boss", would provide guidance to the students and would set the standard much as the boss would in a non-school environment. As a result, the instructor had little difficulty motivating and guiding students.

Subjective. All evaluations during the resident portion of the course were subjective. Certain products were identified as specific products to be evaluated, e.g., tomorrow's briefing, today's letter, next week's decision paper. In reality, every contribution or failure on part of the student was evaluated. Working with a 12 on 1 ratio was possible to "test" all students every day, e.g., "CPT Smith what do you think about GEN Green's article which you were to read last night?" Early on students realized that adequate preparation prevented embarrassment. Also early on, instructors realized the value of establishing a daily record keeping system in order to give students useful feedback.

Timely. Individual feedback was provided students on written products and following student briefings. In most instances, feedback on briefings was provided the day of the briefing and feedback on written products was provided within three days. This would allow the student an opportunity to review comments before his next briefing or before the next paper was due.

School Exercises. There were no "school solutions" only general guidelines on how a paper may be done. This created some difficulty initially as students attempted to "beat the school solution." Emphasis was placed on developing sound, arguable solutions and then developing the skills to present the arguments effectively. Later in the course, students seemed to gain confidence in their own abilities and instead of trying to answer the question "What does the school want?" the question students seemed to answer was, "Is this a sound, logical solution which I am willing to defend?" This appeared achieved the desired result as demonstrated by the fact that students gained confidence in their ability to use logic in solving problems. With the understanding that they were not attempting to develop a solution that would fit a "school mold" students would develop imaginative, creative but logical solutions to problems.

Nature of Evaluation. Feedback provided to students on written products and briefings was highly critical. Initially this created a problem as students were very defensive as they thought their futures in the Army would be destroyed. This provided an initial obstacle to learning which was overcome when students were advised that every report going to official files would indicate only pass or fail, but that reports within the school would remain highly critical. The standard established was essentially a "B" standard - the student receiving the B was considered to be doing very well. Once the standards were established and accepted by the students, growth resulted from the feedback.

Sliding Scale. Another phenomena that occurred was that instructors would grade on a sliding scale. If a student appeared to be getting comfortable while there was still room for improvement, the instructor could up-grade standards, take the comfort away, and in essence keep the student challenged. The 12 to 1 ratio allowed the instructor to sense when more pressure may be needed in order to keep a particular student motivated. This was a definite strength of the subjective evaluation system used.

Lessons Learned

Following is a synopsis of the lessons learned based on the experiences addressed in this paper:

1. Author training contributed to effective writing of course materials in the following ways:

(a) Inexperienced writers were trained in author skills.

(b) Familiarization with the Accountable Instruction System and the use of that system was effective in providing focus to the writing effort.

(c) Awareness of the characteristics of small groups and the types of instruction most suitable to the small group format, assisted authors in developing effective small group instruction.

(d) Understanding of Bloom's taxonomy allowed authors to develop instruction that would provide adequate background knowledge during initial instruction for use at the synthesis and analysis levels during later instruction.

2. Development of learning objectives early in the writing cycle provided a sound foundation for the development of the course materials.

3. The single thread concept has merit but requires extensive coordination early.

4. It is challenging to develop instructional situations where the primary focus is on process and procedures in lieu of but not to the exclusion of content.

5. Printing dates must be well in advance of classroom dates.

6. Instruction in the 12-man staff group was effective.

7. Groups composed of a cross section of experience and expertise resulted in quality peer learning.

8. Highly qualified instructors gained credibility early and contributed to the success of the pilot course.

9. The intensity of the pilot course contributed to the development of team effort and group efficiency.

10. The low threat highly critical evaluation system contributed significantly to student learning.

11. Timely feedback on student products contributed to the learning process.

12. The exclusion of "school solutions" contributed to the student's willingness to develop imaginative logical solutions to problems.

## Decision Aids for Personnel Actions

Ward, Joe H., Jr., Air Force Human Resources Laboratory, Brooks Air Force Base, Texas (Chair); Dumas, Neil S., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia; Kroeker, Leonard P., Navy Personnel R&D Center, San Diego, California.

Military personnel actions are designed to maximize overall system effectiveness. A general framework for viewing personnel action systems will be discussed. The components to be considered are: (1) the measurement of both personnel characteristics and job properties; (2) the estimation of values to the military of each possible person-job assignment; (3) approaches to optimization of the personnel actions, and (4) the importance of constrained choice by members of the system. Attention will be given to a method of Policy Specifying through which pay-off-values are generated. Techniques for ordering lists of proposed personnel actions will be discussed. The concept allows some of these predictor variables to be used to enhance prediction of personnel action outcomes.

DEPARTMENT OF THE ARMY
US ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 EISENHOWER AVENUE
ALEXANDRIA, VIRGINIA 22333

REPLY TO
ATTENTION OF

PERI-RP

# The Army's ABC R&D Program to Build Decision Aids for Personnel Actions

Neil S. Dumas, Ph.D.
US Army Research Institute

While the theory of making "optimal" personnel/manpower decision is easily expressed, the design and implementation of a system which simutaneously provides both the necessary data and mathematical aids has eluded the Services thus far. The Army' R&D program for "Improving the Selection, Classification and Utilization of Enlisted Personnel" comprises three main projects, two of which will be discussed vis-a-vis the selection of "optimal" manpower/personnel decisions in the real world.

> Project B: a "Prototype Computerized Personnel Allocation System" which will use all Project data as well as the latest operations research techniques to produce decision aids that identify "optimal" manpower and personnel choices.
>
> Project C: the National Manpower Inventory - a longitudinal study which: a) assesses the Nation's "stock" of knowledges, skills and aptitudes residing in civilians aged 16-29 and b) provides the basis for a mathematical model that supports policy experimentation and manpower forecasts.

AD P001433

# KALMAN FILTERING TECHNIQUES APPLIED TO ASSIGNMENT SYSTEMS

DR. LEONARD KROEKER
Research Psychologist

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA 92152

Kalman filtering techniques have been successfully applied to time series data by engineers, econometricians and others. Although a standard multiple regression approach can be used under restrictive assumptions, a more dynamic estimation method such as Kalman filtering, can automatically change the estimates with the passage of time. Recent advances in assignment systems employed by the military services have underscored the need for tracking mechanisms that are capable of handling changing process parameters. For example in the Navy's assignment system, CLASP, the 90 entry-level job categories for enlisted personnel yield differing reference population means as the incoming applicant pool changes. The paper focuses on the use of Kalman filtering methodology as a decision aid in generating one step ahead forecasts and discusses the role of such methodology in the broad context of resource management applications.

My topic this afternoon is one that deals with certain aspects of time series prediction. I want to share with you some thoughts on a forecasting methodology and several problems related to applications of the process to military personnel acquisition. In particular, I will concentrate on a problem that involves ninety inter-related time series. In addition to discussing the procedures that generate the time series data and the modeling processes that facilitate explanation, I plan to raise questions concerning solution quality, optimal series selection and the practical use of the prediction mechanism.

State space forecasting involves the determination of models of random processes based on the Markov property that implies independence of the future of the process from its past, given the present state. In other words, the state of a Markov process summarizes all the information from the past that is necessary to predict its future. A general state vector model is typically specified in terms of the following quantities:

(a) vectors of input, output, and internal state variables;

(b) a rule for transforming the state vector from one time point to another;

(c) a relationship among the input, output, and internal state variables;

(d) a description of the initial state of the system; and

(e) the joint statistical relationships among all random variables.

In this paper, each system under investigation is treated as one whose internal functioning is unknown. In other words, each is conceptualized as a black box in which the computer program developed by Raman Mehra and his associates sets parameters according to certain empirical characteristics of the time series data set under study.

A typical time series data set drawn from a military personnel acquisition setting is shown in the following diagram. Notice that the profile is highly irregular in character, that changes in elevations are apparent, and that rapid changes occur before the model or filter is allowed to settle into a stable configuration. In order to understand the purpose of applying state space forecasting methodology to series such as this let us consider the nature of the physical processes generating the series.

The problem originates within the U. S. Navy's recruitment and accession system. The acquisition of manpower occurs in a regular manner despite minor flow perturbations. For example, larger numbers of individuals are inducted during summer months in order to accommodate personal educational objectives. However, the Navy requires a fairly constant mixture of talented persons to enter service throughout the entire year. Since many of these persons may benefit from specialized training, a decision concerning a prospective Navy job must be made at the time of entry. The time series in this study describe fluctuating levels of assignment utility associated with each of the ninety job options.

The assignment of persons to jobs is complicated by certain process and environmental constraints. Specifically, the order of applicant arrival

affects the allocation procedure and necessitates the use of sequential processing. Other influential factors involve individual strengths and weaknesses and organizational quota requirements.

Navy officials have decided to use an automated system to quide the assignment of first term non prior service personnel to jobs. For a given person, the system produces a unique ordered list of job options as a function of personal characteristics, job properties, and system status variables. The way in which the ordered option list is derived depends directly on the time series data points that will be used to develop the forecasting models.

It is important to note that each job option is first evaluated with respect to a number of criteria. Each criterion is represented in the form of a mathematical function that reflects Navy policy concerns and system constraints. Therefore, each aspect of a given person-job match is described by a numerical value. The latter is intended to reflect the merit or utility of a particular match.

In order to derive a single numerical value to describe the quality of any given assignment the utility values for the separate components must be combined. The resulting value is compared to a reference population whose utility values have been similarly calculated, the reference population contains members who may be regarded as potential competitors for the job under consideration. The result of the comparison determines the position of the job option on the unique ordered list.

To illustrate the operation of the list producing process and to develop the argument concerning the role of the reference population in the determination of each time series data set consider a system involving only two job options. Personal and system variables combine to form a composite utility value for each of the two job options. Comparisons with the respective reference populations indicate the relative merit of one job assignment as opposed to another. The mean of a given reference population is the data element of interest for this paper. As the incoming applicant population changes so does the mean of the reference population for a given job. The inter-relationships are complex and are not easily traced.

The mean of any given reference population is called a decision index mean. It is the counterpart to a column mean in a payoff matrix. It fluctuates in value month by month in response to varying personnel quality and other changing system variables.

At this point let me raise a question that bears directly upon the need for a forecasting mechanism, namely, how is the decision index mean obtained for any given month? The means cannot be calculated directly since they depend not only upon personal and system data but also on the configuration of assignments actually made. In short, the pattern of assignments itself influences the values entered in the payoff matrix from which the decision index means are computed.

The calculation of the decision index means, therefore, represents a significant problem and is approached in the following way. An assignment simulation program is run under the assumption that all reference populations

or alternatively, decision index means, are identical. The means associated with the resulting payoff matrix are used as input parameters for the second iteration of the program. A second set of assignments is obtained for all personnel and the payoff matrix changes again. The iterative sequence continues until a stable assignment configuration results. Ironically, when convergence is achieved, the resulting decision index means are those that should have been used for the month that is now a part of the historical record. Clearly, we are always a month behind and what is needed is a predictive mechanism for each time series in order to provide estimates of the decision index means that ought to be used in the assignment system.

The following diagram illustrates the means embedded in the ninety time series. Each column represents the solution for a given month and is independently derived. The rows are the series corresponding to different jobs. The row averages reveal distinct differences between series. The series are typically positively correlated and the distribution of series correlation magnitudes is provided by the following figure.

The following diagram illustrates the behavior of a representative series. The data appear rather noisy indicating that prediction will be relatively difficult. An examination of the data plot for the three-month moving average indicates that the behavior of the series prior to the sixteenth month fluctuates about a value of 4310 and that a temporary shift in elevation occurs at the sixteenth month.

The next diagram displays a plot of the one-step ahead state space prediction for the data just described. The prediction oscillates to a moderate degree in response to the noise component in the data. As a larger number of observations are integrated into the solution the prediction begins to resemble the data profile lagging behind by about one or two time periods near the middle of the series. The filter attempts to track the shift in elevation that occurs about the sixteenth month and appears very sensitive to the noise component as evidenced by the increased oscillation amplitude at time period 21.

Tracking ability can be improved by selecting an appropriate companion series, integrating it into the state vector and deriving the Kalman filter that will yield one step ahead predictions. Effective series selection is difficult to accomplish in practice since the criteria for combination are not well defined.

The effects of adding a second series to the one described earlier can be ascertained from the next slide. The profile resembles the one step ahead prediction generated by the Kalman filter applied to the original single series. The peaks are diminished to a small extent but the shape is highly similar.

A quantitative approach to assessing the degree of improvement resulting from the addition of the second series involves the calculation of the discrepancy sum-of-squares between data points and the corresponding one step ahead predictions. Calculation of the statistic for both the single and double series models allows the desired comparison to be made. The decrease in discrepancy sum-of-squares due to the addition of the second series in this instance is 8.2%, a value that is about average for the data set. Typical reductions range from a high value of 20% to a low one of 0%.

In principle, the reduction of discrepancy sum-of-squares, in order to improve prediction, can be accomplished using more than two series. The data set in this study, limited to 24 observations per series and amply endowed with noise components, has not yielded satisfactory solutions for combinations of three or more series.

The most effective empirical combination procedure involves the selection of series for which one has good quality state space models. For example, when electing to combine two series, one ought to examine the properties of the filter derived for each one separately. A second important consideration in the determination of an effective selection strategy involves the degree of correspondence between the first order differences of the participating series. Preliminary empirical results are encouraging. However, an analytical basis for selection quidelines is desired.

Earlier in this presentation the need for the one step ahead for each series was established and discussed. At this point the question of their use in predicting the decision index means will be addressed.

At present the assignment system is operated under the assumption that the means do not vary with time. The long-term average for each series is used as the measure of location for each respective reference population. Whether or not to employ the model generated predictions in the assignment system as opposed to the long term average for any given series is, therefore, at issue.

The decision procedure to be discussed is intended as a quide to the operational use of the derived state space models in the personnel assignment context. It is designed to determine which of the two options to employ under various conditions.

For a given time point $T_i$ let us consider three quantities, namely, the data point, the predicted point generated by the model, and the long-term average of the series. The discrepancy sum-of-squares term determined from the differences between the data points and the long-term average is defined as $SS_T$. The corresponding sums-of-squares terms due to differences between data points and the predicted values and between predicted values and the long term average are called $SS_D$ and $SS_K$, respectively.

In the following diagrams two conditions that play an important role in the procedure are identified. In the first one, the predicted values tend to be bounded by the data points. In other words, the oscillation amplitude for the predictions is lower than that for the data and $SS_K$ is consequently less than $SS_T$. In the second condition the predicted values display a larger oscillation amplitude than the data points and so the value of $SS_K$ is greater than that of $SS_T$.

During the first stage of the procedure the relative magnitudes of $SS_K$ and $SS_T$ are compared. If the latter is larger, then decision D2 is taken,

namely, that the state space model is used to provide one step ahead forecasts for the series; otherwise the decision is deferred.

During stage two, the relative magnitudes of $SS_D$ and $SS_K$ are the primary determinants of the decision outcome. Small values of $SS_D$, the measure of discrepancy between data points and predicted values, are desired. When $SS_D$ is less than $SS_K$ and when $SS_D$ is smaller than a constant, $SS_T$, the decision is to use the forecasting model. When the two statistics do not meet the above conditions the long-term average value for the series in question is used to anchor the position of the reference population.

The following slide is a flow diagram representing the decision procedure. The outcome designated D2 is achieved when the performance of the state space model meets the criteria outlined above. In effect it governs the conditions under which the forecasting model may be employed within the assignment system.

A minor modification of stage one, as shown in the diagram, represents a slightly more conservative approach with respect to employing the forecasting model.

The results of developing forecasting models for the series in this study are mixed. The models are in the form of time-invariant Kalman filters. In some instances, the prediction of the series are adequate and, in others, unsatisfactory. Rapid shifts in the nature of the series contribute significantly to lack of fit. As a result, the parameters of the state space models change in an unpredictable way and the time-invariant models do not perform well.

Several limitations of the study are noteworthy. The series consisted of only twenty-four data points. Forty points are usually recommended as a minimum number to achieve stable predictions. The data aggregation procedure may also have contributed to the disjointed character of the profiles. Data aggregation over time periods of either one week or two weeks as opposed to one month may result in smoother prediction profiles.

Future research will deal with the investigation of strategies for optimal choice of series to maximize predictability. This appears to be a useful area for further research since sufficient numbers of inter-relationships among series exists and, therefore, better prediction mechanisms can be expected.

In summary, it can be concluded that even under adverse conditions state space forecasting models can make significant contributions. Although time invariant models may not be ideally suited to this kind of data they still outperform other options such as the use of the long-term average. Preliminary results indicate that substantial gains may be made by employing adaptive filtering techniques. It is recommended that the feasibility of employing state space forecasting methodology to a broad range of social science problems be investigated.

# DECISION AIDS FOR PERSONNEL ACTIONS

AD P001434

Joe H. Ward, Jr.

Air Force Human Resources Laboratory
Brooks Air Force Base Texas 78235

## COMPUTER ASSISTED PERSONNEL ACTION CONCEPT

Military personnel actions are designed to maximize overall system effectiveness. In designing personnel systems, attention should be given to evolutionary systems improvement. This allows for future modifications of the system when new policies or new research findings are to be tested and implemented. Evolutionary personnel systems improvement can be helped by using a conceptual model that eases communication among operational and research personnel. This model aids in specifying the research and development required, and in identifying the places in the system where modifications will be made. One description of a personnel system has been described earlier (Ward et al., 1978, Hendrix et al., 1979). A slightly different view is presented below. This representation is called a Computer-Assisted Personnel Action System (COMPAS).

## CONSTRAINED ORDERED LISTS

Consider first the final outputs of the COMPAS model illustrated in Figure 1. The left side of the picture represents a person observing a constrained ordered list of jobs (or other actions) from which to choose. The list is constrained because usually it is not appropriate to offer the entire universe of jobs (or actions) from which a person may choose. The list is ordered in such a way that the choice of a job (or action) near the top of the list will result in better personnel system effectiveness. While constrained ordered lists of jobs are available to personnel of the system, constrained ordered lists of people are available to the job managers -- illustrated on the right side of Figure 1. This list also is constrained and ordered in such a way that the choice of a person near the top will result in better personnel system effectiveness. To produce the constrained ordered lists requires attention to details described below.

## PERSON-JOB INFORMATION

Figure 2 illustrates the major components of the Computer-Assisted Personnel Action System. The basic input required by the System is information about tasks, jobs, or actions required (Job Data Base), and information about personnel available to perform the tasks (Person Data Base). The measurement of personnel characteristics (e.g. aptitudes, interests, achievement, etc) to identify differences among people has been a major research and development activity for many years. In recent years, there has been more attention to the description and measurement of job properties (e.g. mental and physical requirements).

The interaction between people characteristics and job properties is the main concern of personnel action systems. Therefore it is essential to maintain close interdependence of research involving the measurement of people

characteristics and the measurement of job properties. An important example of this interdependence is mentioned below in a discussion of Catalytic Variables.

## PAYOFF VALUE GENERATION

Person and job information must be combined in a way that will provide an estimate of the payoff (or utility) to the personnel system of each possible person-job assignment (or action). One procedure for generating payoff values that has been used operationally by both the Air Force and the Navy is called Policy Specifying (Ward, 1977; Ward et al., 1979). Policy Specifying provides a method of combining many different variables into a single indicator of payoff.

## OPTIMALITY VALUE GENERATION AND CHOICE

It is not always possible to fill each job with that person who has the highest predicted payoff for that job; similarly, every person cannot always occupy that job for which he/she has the highest predicted payoff. The approach used in the COMPAS system is to allocate all persons to jobs in such a way that the sum of the resulting predicted payoff values tends to be as large as possible. When the predicted payoff values are available for all persons on all jobs, then it is feasible to find an allocation of all people to jobs such that the payoff sum is maximized by using an algorithm such as that of Langley et al., (1974). This approach assumes that neither the personnel nor the job managers are allowed any choices in selecting jobs and persons respectively. In many situations it is desirable to allow both personnel and job managers to choose from a constrained ordered list. It is possible that choices from constrained ordered lists by persons or job managers might even lead to closer estimates of the "true" (and unobservable) payoff values. It might be argued that the "choice process" can bring information into the system that "fine tunes" the payoff estimations that are used to produce the ordered lists. Further, if the constrained ordered lists are appropriately developed, then choices from the lists may result in a sum of the predicted values that is very near a maximum value. It is possible that the choice process can yield both a sum that is quite close to the maximum sum of the predicted payoff values and, at the same time, a close estimate of the maximum sum of the "true" (and unobservable) payoff values.

The desire both to optimize and to allow choice from a constrained ordered list leads to generation of an array of optimality values. These values reflect the expected effect on the overall sum of payoff values of assigning each person to each job. Approaches to the generation of Optimality Index values have been described by Ward (1959, 1979). Optimality Index values are computed, put into an Allocation Array and used as a basis for producing the output lists shown in Figure 1. The highest values in the Allocation Array for each person (row) are used to produce the job list. The highest values for each job (column) are used to generate the person list.

## CATALYTIC VARIABLES

The importance of interaction among people characteristics and job properties in personnel classification has been described by Ward (1979). Recognition that interaction among person-job predicted payoff values is the

fundamental prerequisite for differential classification has very important implications. This means that a constant can be added (or subtracted) from any row (or any column) of a person-job match predicted payoff array without changing the optimum allocation of persons to jobs. This fact allows the personnel manager to use selected measures of people characteristics or job properties in a non-interactive way to help estimate accurately the weights associated with the interactive person-job variables. After these selected measures (called Catalytic Variables) have been used to more accurately estimate the interaction component, they are <u>not required</u> for operational use in optimal classification.

To illustrate, assume we are predicting job performance ($Y_{ij}$) of person $i$ on job $j$ from a weighting ($W_1$) of a person's aptitude $X_i$, a weighting ($W_2$) of a job's aptitude requirement $X_j$, and a weighting ($W_3$) of the interaction ($X_i*X_j$)

$$Y_{ij} = W_1 X_i + W_2 X_j + W_3 (X_i*X_j) + E_{ij}$$

Assume that the errors ($E_{ij}$) are large and that the interaction weight $W_3$ is near zero. The objective of adding a Catalytic Variable is to decrease the errors ($E_{ij}$) and increase the size of the interaction weight ($W_3$). This yields an equation with new weights

$$Y_{ij} = (W_1)' X_i + (W_2)' X_j + (W_3)' (X_i*X_j) + W_4X_c + (E_{ij})'$$

where $(W_1)'$, $(W_2)'$, $(W_3)'$ and $(E_{ij})'$ are generally different from $W_1$, $W_2$, $W_3$, and $E_{ij}$. We would like the addition of $W_4X_c$ to the equation to result in a new interaction weight $(W_3)'$ that is larger than $W_3$ and new errors $(E_{ij})'$ that are smaller than $E_{ij}$. Now the predicted values are

$$P_{ij} = (W_1)' X_i + (W_2)' X_j + (W_3)' (X_i*X_j) + W_4X_c$$

If it is desired to classify persons into jobs to maximize the sum of the $P_{ij}$ values, then we can find an optimum classification using only the interaction component

$$(P_{ij})' = (W_3)' (X_i*X_j)$$

Notice that the solution to the classification problem no longer requires $X_c$, the Catalytic Variable, but only the variables $X_i$ and $X_j$. The Catalytic Variable has already completed its required mission of assisting in estimation of the interaction term $(W_3)'$.

Variables that are risky, expensive or controversial may be used in a Catalytic role to enhance interaction among other variables. Then these Catalytic Variables are not required for operational classification.

The application of the Catalytic Variable concept for use in the Armed Service Vocational Aptitude Battery (ASVAB) and the Air Force's Vocational Interest-Career Examination (VOICE) should be investigated. A preliminary study has been completed to illustrate the use of sex and ethnic categories as Catalytic Variables for the Armed Forces Qualifying Test (AFQT).

## SUMMARY

A general framework for viewing personnel action systems was presented. The components considered were: (1) the measurement of both personnel characteristics and job properties; (2) the estimation of values to the military of each possible person-job assignment; (3) approaches to optimization of the personnel actions, and (4) the importance of constrained choice by member of the system. Attention was given to a method of Policy Specifying through which payoff values are generated and the ordering of lists of proposed personnel actions. The concept of catalytic predictor variables for use in differential classification was emphasized.

## REFERENCES


Hendrix, W.H., Ward, J.H., Jr., Pina, M., Jr., & Haney, D.L. Pre-enlistment person-job match system. AFHRL-79-29, AD-A078 427. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, September 1979.

Langley, R.W., Kennington, J., & Shetty, C.M. Efficient computational devices for the capacitated transportation problem. Naval Research Logistics Quarterly, Office of Naval Research, December, 1974, v. 21, No. 4, 637-647.

Ward, J.H., Jr. Use of a decision index in assigning Air Force personnel. WADC-TN-59-38, AD-214 600. Lackland AFB, TX: Personnel Laboratory, Wright Air Development Center, Air Research and Development Command, April 1959.

Ward, J.H., Jr. Creating mathematical models of judgment processes: From policy-capturing to policy-specifying. AFHRL-TR-77-47, AD-A048 983. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, August 1977.

Ward, J.H., Jr. Interaction among people characteristics and job properties in differential classification. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, California, October 1979.

Ward, J.H., Jr., Haney, D.L., Hendrix, W.H., & Pina, M., Jr. Assignment procedures in the Air Force procurement management information system. AFHRL-TR-78-30, AD-A056 531. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, July 1978.

Ward, J.H., Jr., Pina, M., Jr., Fast, J.C., & Roberts, D.K. Policy Specifying with applications to personnel classification and assignment. Proceedings of the 21st Annual Conference of the Military Testing Association, San Diego, California, October 1979.

## MTA STEERING COMMITTEE MEMBERS

US Army Research Institute

US Air Force Human Resources Laboratory

US Air Force Military Occupational Measurement Center

US Coast Guard Institute

US Naval Education and Training Program Development Center

US Navy Personnel Research and Development Center

Belgian Armed Forces Psychological Research Section

Canadian Forces Directorate for Military Occupational Structures

Canadian Forces Applied Research Unit

Federal Republic of Germany Ministry of Defense

Royal Australian Air Force Evaluation Division

# MINUTES

MTA Steering Committee Meeting

Arlington, Virginia, 26 October 1981

## INTRODUCTION

The meeting was opened at 0900 hours at the Quality Inn, Arlington, Virginia by Mr. Arthur Marcus, Conference Coordinator who was representing COL L. Neale Cosby, President. Dr. Arthur C. F. Gilbert, Co-Chairman Program Committee represented Dr. Joyce L. Shields, Secretary. The presence of two representatives from the Federal Republic of Germany was recognized, both from the Ministry of Defense: Fregatten Kapitan R. E. Rolfs and Herr Minrat M. L. Rauch. It was agreed and voted that those individuals would now represent the Federal Republic of Germany instead of the representative from the German Armed Forces Association and the representative from the German Armed Forces Psychological Services Research Institute. The Naval Education and Program Development Center was not represented.

The minutes of the previous year's MTA Conference held in Toronto, Canada was read by the Secretary and unanimously approved by the Steering Committee.

## Publication Review Group

Dr. Waldkoetter reported on the progress of the Publication Review Group. He reported that Praeger Publishers (a subsidiary of Columbia Broadcasting Company), D. C. Heath, and Sage are interested in marketing texts in the area of military psychology. He mentioned the readability of having four books in the area and that the editors and authors were working toward this end.

## Publication of Conference Proceedings

Then followed a discussion of the publication of proceedings of the Association in the future. It was generally agreed that the current practice of publishing the complete text of all papers presented could no longer be entertained because of cost.

Several possibilities were entertained such as placing the complete text of the proceedings in the National Technical Information Service (NTIS) depository and distributing a number of hard-bound copies to the participating agencies. However, it was agreed that the professional publication of any hard-bound copies would be unwise and that the most practical approach would be to place a single photo reproducible copy in NTIS so that either microfiche or hard copies could be made available upon request. The abstracts of the papers would be published as they presently are and these abstracts would serve as a guide to the requestor of papers from the NTIS. Dr. Fischl mentioned the need for a more standardized format for abstracts and papers and this was supported by the members present.

## Suggested Revision of MTA By-Laws

The issue of revision to the current by-laws was raised by the Chairman. He reported that the legality of electing the President of MTA by virtue of his position as Commander of the host organization was questionable as expressed in Article VI B of the current by-laws which states "The President of the Association shall be the Commanding Officer of the armed services agency co-ordinating the annual conference of the Association." The Chairman said that the US Army Research Institute for the Behavioral and Social Sciences (ARI) had obtained a Judge Advocate General's opinion on this and related issues in the current by-laws expressed in Article III that places responsibility on the hosting agency for conducting the annual conference. As a result of the legal opinion the Chairman presented a revised set of by-laws for review by the Steering Committee (Inclosure 1) which principally modifies Article VI B to state "The President, who is elected by a majority of the Steering Committee, serves as the Chief Executive of the Association and as Chairman of the Steering Committee." Then followed a discussion of this issue by the members present. There was some discussion as to the applicability of the change in by-laws to other hosting agencies, but particularly in the case of the foreign agencies represented. The proposed by-laws are to be studied by the Steering Committee who will reach a decision at the next annual conference.

## Greater Participation by Commanders and Staff

The discussion on the proposed changes to the by-laws centered around the legality of the current Military Testing Association procedures in involving command and command staff in the annual conference. It was generally agreed that such participation was meaningful and essential to the successful conduct of MTA conferences.

## Harry Greer Award

There were not any nominations for the Harry Greer award this year. NOTE: It should be incumbent upon members of the Steering Committee to solicit such nominations in the future since the award has not been made since 1979.

## Future Sites of MTA Conferences

Joe Hazel from the AFHRL reported that the Air Force is willing to host the 1982 Annual Conference in San Antonio from 31 October - 5 November 1982 since the Naval Educational and Training Development Center is unable to do so because of problems with facilities. He reported that the Development Center would be able to do so in 1983. The issue was then raised that perhaps the Federal Republic of Germany would like to host the meeting in 1982 since they were scheduled to do so in 1982 and they were next in line after the Navy. The German representatives stated that they did not feel ready to host the conference until 1984 at which time they will be better prepared to do so. The Coast Guard representative stated that the Coast Guard Institute would act as a back-up host for the

1982 and 1983 Conferences and volunteered as the coordinating agency for 1985. Mr. Joe Hazel then suggested that the Air Force would host the conference in San Antonio and that the Conference would be jointly represented by the Air Force Human Resources Laboratory (AFHRL) and the Occupational Measurement Center (OMC). He proposed that COL Terry of AFHRL and COL Ringebach of the OMC serve as Co-Presidents and that the individual who would actually serve as Chairman was not as yet decided. COL Terry and COL Ringebach were then nominated and voted unanimously as the Co-Presidents for the 1982 MTA Conference. In summary, then, the future sites of the MTA Conference are:

1982 Air Force Human Resources Laboratory/Air Force Occupational Measurement Center (San Antonio)

1983 Naval Education and Development Center (Pensacola)

1984 Federal Republic of Germany (Munich)

1985 US Coast Guard Institute (Oklahoma City)

Adjournment

The meeting was adjourned at 1030.

Joyce L. Shields

JOYCE L. SHIELDS, Ph.D.
Secretary

Neale Cosby

L. NEALE COSBY
COL, IN
President

# BY-LAWS OF THE MILITARY TESTING ASSOCIATION*

## Article I - Name

The name of this organization shall be the Military Testing Association.

## Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

## Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

*As approved at the 1978 General Meeting of the Association 2 Nov 78, Oklahoma City, Oklahoma

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

## Article IV - Dues

No annual dues shall be levied against the participants.

## Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. Representation from the majority of the organizations of the Steering Committee shall constitute a quorum.

## Article VI - Officers

A. The Officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to insure notices for conferences. The Secretary shall solicit nominations for the Harry Greer award prior to the annual conference. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

## Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The coordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

C. The annual conference of the Association shall be held at a time and place determined by the coordinating agency. The membership of the association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

## Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in counsultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the association that he cannot accept or continue said duty.

## Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the association.

## Article X - Voting

All members in attendance shall be voting members.

## Article XI - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 2 November 1973).

NAMES AND ADDRESSES OF REGISTRANTS



PREVIOUS PAGE
IS BLANK

MR. WILLIAM ADAMS, JR.
CHIEF, NAVAL EDUCATION AND
TRAINING
N-922
NAS, PENSACOLA FL 32508

HOMER W. ADKINS
N-911
NAS, PENSACOLA FL 32508

DIRECTOR
DEFENSE MAPPING SCHOOL
ATTN: MR. J. ROBERT AINSLEY
DMS-TDE
FORT BELVOIR VA 22060

JUDY AKIN
RAYTHEON SERVICE COMPANY.
2 WAYSIDE ROAD
BURLINGTON MA 01801

DEBORAH ALLEN
ASSESSMENT DIRECTOR
COLLEGE FOR HUMAN SERVICES
345 HUDSON STREET
NEW YORK NY 10014

CMDR NICHOLAS H. ALLEN
US COAST GUARD HDQTRS
G-PO/OPES
2100 2ND STREET SW
WASHINGTON DC 20393

ERNEST J. ANASTASIO, ASST. V.P.
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON NJ 08341

DR. MIKE ANDERSON
CGSC-CAS3
ATZL-SWB
FORT LEAVENWORTH KS 66027

DONNA C. ANGLE
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

MR. THOMAS M. ANSBRO
CNET N-34
BUILDING 624
NAS PENSACOLA FL 32508

LORRAINE APPLETON
COMDT (G-RT) USCG
2100 SECOND STREET SW
WASHINGTON DC 20593

CDR SUSANNE R. ARMSTRONG
HSETC, CODE 21
NAVAL MEDICAL CENTER
BETHESDA MD 20814

EUGENE ARTHUR
OFFICE OF ARMOR FORCE
MANAGEMENT & STANDARDIZATION
ATZK-AM BLDG 641
US ARMY ARMOR CENTER
FT. KNOX KY 40121

EUGENE ARTHUR
OFFICE OF ARMOR FORCE
MANAGEMENT & STANDARDIZATION
ATZK-AM BLDG 641
US ARMY ARMOR CENTER
FT. KNOX KY 40121

LINDA ASLETT
INDUSTRIAL TRAINING ANALYSIS,
RESEARCH, & DEVELOPMENT BRANCH
HSA-TIA
ACADEMY OF HEALTH SCIENCES
FORT SAM HOUSTON TX 78234

NEHAMA BABIN
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

EVA L. BAKER
CENTER FOR THE STUDY OF EVALUATION
UCLA, SCHOOL OF EDUCATION
405 HILGARD AVENUE
LOS ANGELES, CALIFORNIA 90024

ANNE L. BALLARD
NMSETC, CODE 211
NATIONAL NAVAL MEDICAL CENTER
BETHESDA MD 20814

MAJ ROGER D. BALLENTINE, USAF
3400 SWIFT DRIVE
RALEIGH NC 27606

CPT NEIL A. BARRETT
NATIONAL DEFENCE HEADQUARTERS
140 O'CONNOR STREET
OTTAWA, ONTARIO
CANADA K1A 0K2

ALBERT E. BEATON
EDUCATIONAL TESTING SERVICE
PRINCETON NJ 08341

C. DEREK BEEL
R. N. SCHOOL OF EDUCATION AND
TRAINING TECHNOLOGY
HMS NELSON, PORTSMOUTH, ENGLAND

ISAAC I. BEJAR
EDUCATIONAL TESTING SERVICE
PRINCETON NJ 08541

BRUCE BENNETT
SCIENCE APPLICATIONS INC.
1710 GOODRIDGE DRIVE
MCLEAN VA 22102

MR. B. MICHAEL BERGER
DEPUTY MANAGER, ANALYSIS
& EVALUATION DIVISION
NATIONAL HQS
SELECTIVE SERVICE SYSTEM
1023 31ST STREET, N.W.
WASHINGTON DC 20435

DR. MELISSA S. BERKOWITZ
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

SIDNEY BLUM
ATSA-TDI-Q
US ARMY AIR DEFENSE SCHOOL
FT. BLISS TX 79916

DR. DOUGLAS J. BOBKO
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

ARNOLD BOHRER
REKRUTERINGS-EN SELECTIECENRUM
SECTIE PSYCHOLOGISCH ONDERZOEK
KAZERNE KLEIN KASTEELTJE
B 1000 BRUSSEL BELGIUM

DR. STANLEY BOLIN
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

WADE BOSWELL
NODAC, BUILDING 150
WASHINGTON NAVY YARD
WASHINGTON DC 20374

FRANK B. BRAUN
DATA DESIGN LABORATORIES
SUITE 307
1755 S. JEFFERSON DAVIS HIGHWAY
ARLINGTON VA 22202

DAVID B. BREWER
COMMANDANT (G-P-1/2 TP42)
US COAST GUARD HQ
2100 SECOND STREET SW
WASHINGTON DC 20593

MR. CLAUDE BRIDGES
OFFICE OF INSTITUTE RESEARCH
USMA
WEST POINT, NY 10996

SUE T. BRIDGES
HQ USAF/RSXX
RANDOLPH AFB TX 78150

MAJ ROBERT D. BRODY
ACADEMIC INSTRUCTION SCHOOL
USAF AIFOS/EDV
MAXWELL AFB AL 36112

LCDR DWIGHT C. BROBA III
COMDT (G-RT) USCG
2100 SECOND STREET SW
WASHINGTON DC 20593

CORNELIA BRUNNER
COLLEGE FOR HUMAN SERVICES
345 HUDSON STREET
NEW YORK NY 10014

WILLIAM P. BURKE
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

FRED BURKHART
SSC-NCR-MS
200 STOVALL STREET
ALEXANDRIA VA 22332

WILLIAM B. CAMM
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

DAVID W. CAMPBELL
NODAC, BUILDING 150
WASHINGTON NAVY YARD
WASHINGTON DC 20374

RICHARD CAMPBELL
TRANSPORT CANADA, AIR TRAFFIC CONTROL
PLACE DE VILLE
OTTOWA, ONTARIO
CANADA K1A ONB

EDMUND J. CARBERRY
ATZK-DET-FD
USAARMC
FORT KNOX KY 40121

ROBERT R. CARLSON
MDCDEC, EDUCATION PROGRAMS
EDUCATION CENTER
US MARINE CORPS
QUANTICO VA 22134

JAMIE CARLYLE
PECC-FSS, CIVPERCEN
200 STOVALL STREET
ALEXANDRIA VA 22332

RAY S. CARROLL, JR.
US GENERAL ACCOUNTING OFFICE
5705 THURSTON AVENUE
VIRGINIA BEACH VA 23455

ALBERT L. CAVALIERI
FRANKLIN RESEARCH CENTER
20TH & BENJAMIN FRANKLIN PARKWAY
PHILADELPHIA PA 19103

JIM CAVINESS
USA SOLDIER SUPPORT CENTER
FT. BEN HARRISON IN 46216

LARRY STEVEN CECIL
6824
NAVY ANNEX
WASHINGTON DC 20305

RANDALL M. CHAMBERS, PH.D.
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

GERALD CHASIN
US ARMY SOLDIER SUPPORT CENTER
ROOM 3-S-11, HOFFMAN II
200 STOVALL STREET
ALEXANDRIA VA 22332

MAJ YVAN CHATIGNY
MLM DEPT
COLLEGE MILITAIRE ROYAL DE SAINT-JEAN
ST-JEAN, QUEBEC
CANADA J0J 1R0

JOAN T. CHIPPENDALE
SMD-ATSC
BLDG. 2787
FORT EUSTIS VA 23604

1705

ARNOLD T. CHU
CONTROL DATA CORP
P.O. BOX 0
MINNEAPOLIS MN 55440

MR. ROBERT N. CISCO
ATZK-AM(P)
HQ USAARMC
FT. KNOX KY 40121

BERNARD E. CLARK
RESEARCH TRIANGLE INSTITUTE
P.O. BOX 12194
RESEARCH TRIANGLE PARK NC 27709

DR. JOHN CLAUDY
AMERICAN INSTITUTES FOR RESEARCH
P.O. BOX 1113
PALO ALTO CA 94302

MRS. BETTY H. COLLETTI
CODE 0041
NAVAL HEALTH SCIENCES EDUCATION
& TRAINING COMMAND
NATIONAL NAVAL MEDICAL CENTER
BETHESDA MD 20814

CAPT ROBERT S. COLLYER
AUSTRALIAN ARMY
AFHRL/MODS
BROOKS AFB TX 78235

STEPHEN M. CORMIER
PERSONNEL RESEARCH AND
DEVELOPMENT CENTER
OPM
1900 E STREET, N.W.
WASHINGTON DC 20415

BERTHA H. CORY
P.O. BOX 442
CHESTERTOWN MD 21620

COL L. NEALE COSBY
COMMANDER
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

J. LAMARR COX
RESEARCH TRIANGLE INSTITUTE
P.O. BOX 12194
RESEARCH TRIANGLE PARK NC 27709

MR. PAUL R. CROLL
ROOM 3B29
OFFICE OF PERSONNEL MANAGEMENT
1900 E STREET NW
WASHINGTON, DC 20415

RICHARD B. DARLINGTON
DEPT. OF PSYCHOLOGY - URIS HALL
CORNELL UNIVERSITY
CHICAGO IL 60605

MR. D. D. DAVIS
CNET
CODE N-52
MONTEREY CA 93940

LTC DENNIS A. DEFRAIN
COMMAND AND GENERAL STAFF
COLLEGE
OKLAHOMA CITY OK 73169

CPT THOMAS J. DEMBECK
ACADEMY OF HEALTH SCIENCES
HSA-ZTE
WASHINGTON DC 20415

FRANK A. DIBELLO
PEAT, MARWICK, MITCHELL & CO.
1990 K STREET NW
PETALUMA CA 74952

RICHARD W. DICKINSON
OCCUPATIONAL RESEARCH DIVISION
INDUSTRIAL ENGINEERING DEPARTMENT
TEXAS A&M UNIVERSITY
SAN DIEGO CA 92152

CAPT RICHARD DOHERTY
USCG
2100 SECOND STREET SW
WASHINGTON DC 20593

JOHN A. DOMME
ARMY RESEARCH INSTITUTE
PO BOX 476
FT. RUCKER AL 36362

RICHARD D. DOORLEY
MILPERCEN, ROOM 3S07
200 STOVALL STREET
PRINCETON NJ 08541

ANDREW N. DOW
CNET, N-1222
WASHINGTON DC 20008

DR. WALTER E. DRISKILL
USAF OCCUPATIONAL ANALYSIS CENTER
USAFOMC/OMU
RANDOLPH AFB TX 78148

ERIC DUNCAN
USAFOMC/OMDRA
RANDOLPH AFB TX 78128

LTC CHARLES V. DURHAM
AWC/EDV
MAXWELL AFB AL 36112

DR. CAROL DWYER
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON NJ 08534

DR. ROBERT EASTMAN
25 WOODLAND ROAD
POCUDSON VA 23662

NEWELL KENT EATON
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

WILLIAM C. EBELING
NAVAL GUIDED MISSILES SCHOOL
DAM NECK
VIRGINIA BEACH VA

DR. MARK J. EITELBERG
HUMAN RESOURCES RESEARCH ORGANIZATION
300 NORTH WASHINGTON STREET
ALEXANDRIA VA 22314



1706

TIMOTHY W. ELIS
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

DR. JOHN A. ELLIS
NAVY PERSONNEL RESEARCH AND
DEVELOPMENT CENTER
SAN DIEGO CA 92152

MAJ R.T. ELLIS
ATTN: DPSRSC RESEARCH
NATIONAL DEFENCE HEADQUARTERS
101 COLONEL BY DRIVE
OTTAWA, ONTARIO
CANADA K1A 0K2

BARRY M. FARRELL
CORPORATE PERSONNEL RESEARCH
8100 34TH AVENUE SOUTH
MINNEAPOLIS MN 55440

DR. KENN FINSTUEN
AFHRL/MODF
BROOKS AFB TX 78235

DR. MYRON A. FISCHL
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

COMMANDER
SCHO
US ARMY INTELL SCHOOL
ATTN: ATSIE-DE DR. E. B. FLYNN, JR.
FT. DEVENS MA 01433

LT. D.E. FORESTELL
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

TERRY M. FRANUS
MARINE CORPS INSTITUTE
MARINE BARRACKS, BOX 1775
WASHINGTON DC 20013

JON S. FREDA
TRAINING ANALYSIS AND EVALUATION GROUP
DEPARTMENT OF THE NAVY
ORLANDO FL 32813

DR. E. WAYNE FREDERICKSON
P.O. BOX 6057
FT. BLISS TX 79916

ROBERT L. FREY, JR.
HEADQUARTERS, U.S. COAST GUARD
G-P-1/2/42
WASHINGTON DC 20593

CARMELA RAPILLO FRICKERT
NHSETC
T-16 CODE 23 NNMC
BETHESDA MD 20814

DAVID FRIEDMAN
RESEARCH APPLICATIONS INC
SUITE 320 - SOUTH
1776 EAST JEFFERSON STREET
ROCKVILLE MD 20852

DR. PAUL GADE
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

CAPT T.J. GALLAGHER, MSCOSN
NAVAL AIR DEVELOPMENT CENTER (703)
WARMINSTER PA 18974

HELEN BELLETTI GARLAND
ATSA-TDI-Q
COMMANDANT, US ARMY AIR DEFENSE SCHOOL
FT. BLISS TX 79916

DAVID GARNETT
MCI - USMC
USMC BARRACKS
EIGHTH AND "EYE" STREETS SE
WASHINGTON DC 20013

DR. JAMES C. GEDDIE
DRXSY-HEL (DR. GEDDIE)
HUMAN ENGINEERING LABORATORY
ABERDEEN PROVING GROUND
ABERDEEN MD 21005

EDWARD GERAGHTY
US COAST GUARD
TRACEN GOVERNORS ISLAND
NEW YORK NY 10004

DR. ARTHUR C.F. GILBERT
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

JOYCE GIORGIA
AFHRL/MODS
BROOKS AFB TX 78235

J. P. GODBOUT
LSC DETACHMENT
CFB ST-JEAN,
RICHELAIN, P.Q.
CANADA J0J 1R0

DR. DOUG GOODGAME
OCCUPATIONAL RESEARCH PROGRAM
INDUSTRIAL ENGINEERING DEPT
TEXAS A&M UNIVERSITY
COLLEGE STATION TX 77801

CAPT JOHN GOODMAN
CANADIAN ARMED FORCES
14 TRG GROUP
CFB WINNIPEG
CANADA R3R 0T0

BARRY GOODSTADT
WESTAT INC
1650 RESEARCH BLVD.
ROCKVILLE MD 20850

DR. JOHN R. GORAL
FPCD ROOM 4001
US GENERAL ACCOUNTING OFFICE
441 G STREET
WASHINGTON DC 20548

M. MERIWETHER GORDON
AFROTC/XRX
MAXWELL AFB AL 36112

STEVEN GORMAN
806 ASHLAND AVENUE
ST. PAUL MN 55104

1707

DR. ALEXANDER M. GOTTESMAN
HEAD, CURRICULUM BRANCH
NSETC, CODE , BLDG. 1 (T-16)
NATIONAL NAVAL MEDICAL CENTER
BETHESDA MD 20814

DR. R. BRUCE GOULD
AFHRL/HOAP
BROOKS AFB TX 78235

LESLIE M. GREGOR
U.S. GENERAL ACCOUNTING OFFICE
5705 THURSTON AVENUE
VIRGINIA BEACH VA 23455

K. W. GRIERSON
BOX 5986
FLORIDA STATE UNIVERSITY
TALLAHASSEE FL 32303

DENIS GULAKOWSKI
XMCO INC. SUITE 801
8200 GREENSBORO DRIVE
MACLEAN VA 22102

JOHN S. GUTHRIE, JR.
OASA (MRA) ROOM 2E591
DEPARTMENT OF THE ARMY
PENTAGON
WASHINGTON DC

DR. JOSEPH HAGMAN
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

CAPT THOMAS M. HALE, USN
9804 WARD COURT
FAIRFAX VA 22032

EUGENE R. HALL
TRAINING ANALYSIS AND
EVALUATION GROUP
DEPARTMENT OF THE NAVY
ORLANDO FL 32813

U.S. ARMY ENGINEER SCHOOL
ATTN- NANCY J. HAMM
DTD, ITD, SQT BRANCH
BLDG 230
FT. BELVIOR VA 22066

D. L. HANCQ
IBM
P.O. BOX 12195
RESEARCH TRIANGLE PARK NC 27709

KENNETH E. HANSEN
PSYCH SYSTEMS INC.
600 REISTERSTOWN RD, SUITE 404
BALTIMORE MD 21208

DR. JOAN HARMAN
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

KENNETH R. HARMON
DENVER RESEARCH INSTITUTE
UNIVERSITY OF DENVER
DENVER CO

PHILIP R. HARVEY
EDUCATIONAL TESTING SERVICE
1 AMERICAN PLAZA
EVANSTON IL 60202

JOHN E. HASSEN
CNET, BLDG. 679, CODE N9
NAS
PENSACOLA FL 32508

CDR F. J. HAURYSH
ATTN- DMOS
NATIONAL DEFENCE HEADQUARTERS
101 COLONEL BY DRIVE
OTTAWA, ONTARIO
CANADA K1A 0K2

WILLIAM A. HAYES
CNET, N-541, BLDG 624
NAS PENSACOLA FL 32508

DR. JOE T HAZEL
AFHRL/AZ
BROOKS AIR FORCE BASE TX 78235

MISS CAROL HEATON
PERSONNEL PSYCHOLOGY DIVISION
ARMY PERSONNEL RESEARCH ESTABLISHMENT
C/O RAE
FARNBOROUGH, HANTS UK

LTC WM. H. HENDRIX
AFIT/LSB
WRIGHT PATTERSON AFB OH 45433

DR. JACK M. HICKS
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

EDWARD N. HOBSON
DDL SUITE 307
1755 S. JEFFERSON DAVIS HWY
ARLINGTON VA 22202

CHRIS L. HOLM
NODAC BLDG. 150
WASHINGTON NAVY YARD
WASHINGTON DC 20374

DR. LYDIA HOOKE
HUMAN RESOURCES RESEARCH ORGANIZATION
300 NORTH WASHINGTON STREET
ALEXANDRIA VA 22314

MAJOR DAVID S. HORTON
CANADIAN FORCES BASE GAGETOWN
NEW BRUNSWICK CANADA

DR. CLARK L. HOSMER
39 LONGWOOD DR.
SHALIMAR FL 32579

LTC CLIFTON HOUSTON
EXECUTIVE OFFICER
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

BEVERLY HOUTZ
EDUCATION SPECIALIST
NAVY - HOSPITAL CORPS SCHOOL
BLDG 130H
GREAT LAKES IL 60088

1708

R.P HUNDLEY, PH D
VIRGINIA ELECTRIC AND POWER CO.
P.O. BOX 26666
RICHMOND VA 23261

ULYSSES S. JAMES
ARTHUR YOUNG & COMPANY
1025 CONNECTICUT AVE., N.W.
WASHINGTON DC 20036

SQN LDR HANS JANSEN RAAF
AFHRL/MODS
BROOKS AFB TX 78235

E7 R. JENKINS
DEPARTMENT OF DEFENSE
FT. GEO. MEADE MD 21755

JOHN B. JOAQUIN
ONTARIO HYDRO ROOM H2-A17
700 UNIVERSITY AVENUE
TORONTO, ONTARIO
CANADA M5G 1X6

JAMES H. JOHNSON
PSYCH SYSTEMS INC.
600 REISTERSTOWN RD., SUITE 404
BALTIMORE MD 21208

ROBERT N. JOHNSON
DTD-SOLDIER SUPPORT CENTER
FORT BEN HARRISON IN 46216

DANIEL B. JONES
ARI
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

DR. DOUGLAS H. JONES
EDUCATIONAL TESTING SERVICE
T-255 ROSEDALE ROAD
PRINCETON NJ 08541

D. TODD JONES
OFFICE OF RESEARCH AND
DEVELOPMENT (G-DMT)
U.S. COAST GUARD
2100 SECOND STREET SW
WASHINGTON DC 20593

KAREN N. JONES
U.S. COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY OK 73169

TAFT M. JOSEPH, JR
DIRECTORATE TRAINING DEVELOPMENTS
US ARMY FIELD ARTILLERY SCHOOL
FT. SILL OK 73503

JOHN N. JOYNER
HUMRRO
27857 BERWICK DRIVE
CARMEL CA 93950

ADELLE KARMAS
CANADIAN FORCES NDHQ
101 COLONEL BY DRIVE
OTTAWA ONTARIO
CANADA K1A 0K2

RICHARD A. KASS
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE.
ALEXANDRIA VA 22333

JUDAH KATZNELSON
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

CAPT WAYNE E. KEATES
STAFF OFFICER TRAINING DEVELOPMENT 2
14 TRAINING GROUP HEADQUARTERS
WESTWIN MANITOBA
CANADA R2R 0T0

MR. JAMES B. KEETH
USAFOMC/OMYO
RANDOLPH AFB TX 78150

DR. RICHARD P. KERN
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

SUSAN KERNER-HOEG
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

DR. JOHN J. KESSLER
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

MARYANN J. KICINSKI
4444 OPS SQD (OTD)(TAC)
LANGLEY AFB VA 23666

MELVIN J. KIMMEL
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

DEPARTMENT OF THE NAVY
TRAINING ANALYSIS & EVALUATION GROUP
ATTN- DR. J. PETER KINCAID
NTC ORLANDO FL 32813

DR. C. MAZIE KNERR
HUMRRO
300 NORTH WASHINGTON STREET
ALEXANDRIA VA 22314

DEFENSE LANGUAGE INSTITUTE
ATTN: ATFL-TD-T(MAJ A. N. KNOX)
PRESIDIO OF MONTEREY
MONTEREY CA 93940

CHRISTOPHER G. KOCH
HONEYWELL INC.
2600 RIDGWAY PKWY MN17-2318
MINNEAPOLIS MN 55413

CPT JAMES C. KOHLER
HSA-TOD
COMMANDANT
US ARMY ACADEMY OF HEALTH SCIENCES
FT. SAM HOUSTON TX 78234

DR. ARTHUR L. KOROTKIN
INSTITUTE FOR BEHAVIORAL RESEARCH
2429 LINDEN LANE
SILVER SPRING, MD. 20910

TERRY A. KREMER
US GENERAL ACCOUNTING OFFICE
ROOM 4100-A
441 G STREET NW
WASHINGTON DC 20548

1709

DR. SAMUEL E. KRUG, DIRECTOR
TEST SERVICES DIVISION
INST. FOR PERSONALITY &
ABILITY TESTING, INC.
P.O. BOX 188
CHAMPAIGN IL 61820

RICHARD S. LANTERMAN
CHIEF, PSYCHOLOGICAL RESEARCH BRANCH
HQ, US COAST GUARD
G-P-1/2/42
2100 SECOND STREET SW
WASHINGTON DC 20593

MS. JANICE LAURENCE
HUMAN RESOURCES RESEARCH ORGANIZATION
300 NORTH WASHINGTON STREET
ALEXANDRIA, VIRGINIA 22314

THOMAS Y. LAURENCE
G-P-1/42
HQ US COAST GUARD
2100 SECOND STREET SW
WASHINGTON DC 20593

MR. I. GAVIN LIDBERDALE
(RAF EXCHANGE)
AFHRL OPERATIONS TNG DIVISION (OTBO)
WILLIAMS AIR FORCE BASE AZ 85224

SUZANNE LIPSCOMB
AFHRL/MODE
BROOKS AFB TX 78235

JAMES F. LIS
RESERVE TRAINING CENTER
US COAST GUARD
YORKTOWN VA 23690

DR. ALEXANDER A. LONSO
ATTG-DOR
TRAINING DEVELOPMENT INSTITUTE
TRADOC
FT. MONROE VA 23651

DR. ROBERT LOO
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE STREET # 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

DON R. LYON
AFHRL/OT
WILLIAMS AFB AZ 85234

MR. DOUGLAS MACPHERSON
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

EDWARD F. MAGDARZ
IBM CORP. DEPT 812/645
BOX 12195
RESEARCH TRIANGLE PARK NC 27709

ANN MAJCHRZAK
WESTAT
1650 RESEARCH BLVD
ROCKVILLE MD 20850

ARTHUR MARCUS
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE.
ALEXANDRIA, VIRGINIA 22333

OLAN E. MARTIN
MARINE CORPS INSTITUTE
MARINE BARRACKS
EIGHTH & "EYE" STREETS SE
WASHINGTON DC 20390

JOHN J. MATHEWS
AFHRL/MOAM
BROOKS AFB TX 78222

GERARD MAYLAN
DEPARTMENT OF PERSONNEL
301 W. PRESTON ROOM 508A
BALTIMORE MD

PAMELA V. MAYS
US ARMY RESEARCH INSTITUTE
P.O. BOX 2086
FT. BENNING GA 31905

L. KENT MCBEE
NAVAL TECH. TRAINING STATION
CORRY STA.
PENSACOLA FL

SHERYL MCCLELLTAN
NODAC BLDG 150
WASHINGTON NAVY YARD
WASHINGTON DC 20374

JIM MCCUTHEON
CANADIAN ARMED FORCES
NATIONAL DEFENCE HEADQUARTERS
120 LAURIER AVENUE
OTTAWA
CANADA

RICHARD H. MCKILLIP
CHIEF, RESEARCH BRANCH ROOM 3023
OFFICE OF PERSONNEL MANAGEMENT
1900 E STREET NW
WASHINGTON DC 20415

HAROLD A. MCWILLIAMS
NATIONAL OPINION RESEARCH CENTER
6030 S. ELLIS AVENUE
CHICAGO IL 60637

ADELHEID MEISSNER
DEZ. WEHRPSYCHOLOGIE IM
STREITKRAEFTEAMT
POST BOX 205003
D-5300 BONN 2
FEDERAL REPUBLIC OF
GERMANY

DR. WILLIAM H. MELCHING
HUMAN RESOURCES RESEARCH ORG.
P.O. 293
FT. KNOX KY 40121

DR. JOHN B. MEREDITH, JR.
DATA-DESIGN LABS, INC.
BLDG. 15 #140
KOGER EXECUTIVE CENTER
NORFOLK VA

PAUL A. MICHAEL
USAMMCS, ATTN- ATSK-E
REDSTONE ARSENAL, AL 35897

1710

DR. ANGELO MIRABELLA
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

JEAN MITCHELL
DEFENSE LANGUAGE INSTITUTE
LACKLAND AFB TX 78236

AMELIA E. MOBLEY
US COAST GUARD
2100 SECOND STREET SW
WASHINGTON DC 20593

JOHN A. MODRICK
HONEYWELL
2600 RIDGWAY PKWY NE
MINNEAPOLIS MN

KATHLEEN MOONEY
DEPT. OF DEFENSE
FT. MEADE MD

EDNA MORGAN
NAVAL EDUCATION TRAINING PROGRAM
DEVELOPMENT CENTER DETACHMENT
NAVAL TRAINING CENTER BLDG 90
GREAT LAKES IL 60088

HENRY MUELLER
HEADQUARTERS, US POSTAL SERVICE
475 L'ENFANT PLAZA
WASHINGTON DC

JOHN W. MURPHY
TEST DESIGN COORDINATOR
DOTD, RTM
FT. BENJAMIN HARRISON IN 46216

GILLES NADON
CANADIAN ARMED FORCES
LANGUAGE STANDARDS CONTROL DET. ST-JEAN
RICHELAIN QUEBEC
CANADA J0J 1R0

MS. JEAN NEWTON
OFFICE OF PERSONNEL MANAGEMENT (CODE 3609)
1900 E ST NW
WASHINGTON DC 20415

HQ USAREC
ATTN- USARC RO-EP (DR. NANCY A. NIEBOER)
FT. SHERIDAN IL 60037

DR. GLENDA Y. NOGAMI
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

DR. DANIEL NUSSBAUM
US ARMY CONCEPTS ANALYSIS AGENCY
8120 WOODMONT AVENUE
BETHESDA MD 20814

LTJG FRANK X. O'BYRNE, JR.
US COAST GUARD, TRACEN NEW YORK (CTD)
GOVERNORS ISLAND NY 10004

COL JOHN B. O'LEARY
P.O 381 MAYO
UNIV. OF MINNESOTA MEDICAL SCHOOL
MINNEAPOLIS MN 55455

LAUREL W. OLIVER
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA, VA 22333

FRANCIS E. O'MARA
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE.
ALEXANDRIA, VA. 22333

DR. HAROLD F. ONEIL, JR.
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

RICHARD J OREND
HUMAN RESOURCES RESEARCH ORGANIZATION
300 NORTH WASHINGTON STREET
ALEXANDRIA VA 22314

DR. JESSE ORLANSKY
INSTITUTE FOR DEFENSE ANALYSES
400 ARMY-NAVY DRIVE
ARLINGTON VA 22311

WILLIAM C. OSBORN
VICE PRESIDENT AND DIRECTOR
HUMRRO MILITARY RESEARCH DIVISION
633 KNOX BLVD
FT. KNOX, KY 40160

MR. DAVID J. OWEN
NATIONAL DEFENSE HQ
ATTN- DMOS 3-4
101 COLONEL BY DRIVE
OTTAWA, ONTARIO
CANADA K1A 0K2

TANYA L. PAGE
STATE OF MD. DEPT OF PERSONNEL
301 PRESTON STREET
BALTIMORE MD

DR. ROBERT P. PALESE
HEADQUARTERS (G-P-1/2/42)
US COAST GUARD
WASHINGTON DC 20593

DR. RICHARD L. PALMER
US ARMY RESEARCH INSTITUTE
HQ TCATA
FORT HOOD TX 76544

MS. LINDA PAPPAS
MAY ASSOCIATES
1110 VERMONT AVENUE NW
WASHINGTON DC 20005

MR. CORTEZ PARKS
CHIEF DLIELC-LEE
DEFENSE LANGUAGE INSTITUTE
ENGLISH LANGUAGE CENTER
LACKLAND AFB TX 78236

DR. JOHN J. PASS
NODAC
BLDG 150
WASHINGTON NAVY YARD
WASHINGTON DC 20374

RAY STEPHEN PEREZ
INTERAMERICA RESEARCH AND ASSOCIATE
1555 WILSON BLVD SUITE 600
ROSSLYN VA 22209

WILLIAM J. PHALEN
AFHRL/MOMA
BROOKS AFB TX 78235

JAMES ERIC PIERCE
USACMCS, ATSC-EV
FT. MONMOUTH NJ

E. RICHARD PIGEON
DEPARTMENT NATIONAL DEFENCE - CANADA
BERGER BLDG., 7TH FLOOR
OTTAWA, ONTARIO
CANADA K1A 0K2

DR. ROBERT PLEBAN
U.S. ARMY RESEARCH INSTITUTE FIELD UNIT
P.O. BOX 2086
FT. BENNING GA 31905

WILLIAM POPYUK
LANGUAGE STANDARDS CONTROL DETACHMENT
CFB ST. JEAN, RICHELAIN,
QUEBEC, CANADA J0J 1R0

MG JACK I. POSNER (RET)
ASSOCIATE DIRECTOR MANAGEMENT AND ORGANIZA
GENERAL RESEARCH CORPORATION
7655 OLD SPRINGHOUSE ROAD
WESTGATE RESEARCH PARK
MCLEAN VA 22102

DR. EARL H. POTTER III
DEPARTMENT OF HUMANITIES
US COAST GUARD ACADEMY
NEW LONDON CT. 06320

DR. FRANK C. PRATZNER
CENTER FOR VOCATIONAL EDUC.
OHIO STATE UNIVERSITY
1960 KENNY RD.
COLUMBUS OH 43210

MAJOR TERRY J. PROCIUK
MLM DEPT
ROYAL MILITARY COLLEGE
KINGSTON, ONTARIO
CANADA K7L 2W3

JOHN PYECHA
RESEARCH TRIANGLE INSTITUTE
P.O. BOX 12194
RESEARCH TRIANGLE PARK NC 27709

EDYS QUELLMALZ
CENTER FOR THE STUDY OF EVALUATION
UCLA, SCHOOL OF EDUCATION
405 HILGARD
LOS ANGELES CA 90266

GLENN M. RAMPTON
CANADIAN FORCES
101 COLONEL BY DRIVE
OTTAWA, CANADA K1A 0K2

DR. WM. C. RANKIN
TRAINING ANALYSIS & EVALUATION GROUP
BLDG 2047
NAVY TRAINING CENTER
ORLANDO, FL 32813

MARTIN L. RAUCH
CHIEF PSYCHOLOGIST-MOD-BONN
MINISTERY OF DEFENCE
POST BOX 1328
53 BONN 1,
FEDERAL REPUBLIC OF GERMANY

LOUIS REEVES
EMPLOYMENT AND IMMIGRATION CANADA
PLACE DU PORTAGE PHASE 4
OTTAWA, CANADA K1A 0G9

JOHN A. RICCOBONO
P.O. BOX 12194
RESEARCH TRIANGLE INSTITUTE
RESEARCH TRIANGLE PARK NC 27709

MR. FRANK RIPKIN
DEPARTMENT OF THE NAVY OP-14
CODE OP-140F2
ROOM 6824
ARLINGTON ANNEX
WASHINGTON DC 20350

MR. W.H. RITCHIE
NATIONAL DEFENCE HEADQUARTERS DGPRD
OTTAWA, ONTARIO
CANADA K1A 0K2

DOROTHY A. RIVERA
DEPT. OF DEFENSE
FT. MEADE MD

DR. DONALD ROCK
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON, NEW JERSEY 08541

LARRY ROGERS
I. G. BROWN PROFESSIONAL MILITARY
EDUCATION CENTER
BOX 10
ALCOA TN

ECKART ROLFS
MOD, 5300 BONN 1
FUE S IV SDB/FAPS
P. B. 13 28
FEDERAL REPUBLIC OF
GERMANY

CHARLES R. ROLL JR.
POLICY AND MANAGEMENT PLANNING GROUP
SCIENCE APPLICATIONS INC.
1710 GOODRIDGE DRIVE
MCLEAN VA 22102

ALAN P. ROMANCZUK
WESTAT, INC.
1650 RESEARCH BLVD.
ROCKVILLE MD 20850

KENDALL L. ROOSE
ACADEMIC TRAINING
NAS WHITING
MILTON FL 32570

ALEXANDER ROSE
NODAC BLDG 150
WASHINGTON NAVY YARD
WASHINGTON DC 20374

MR. ROBERT ROSS
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

1712

MARJORIE M. ROYLE
NPRDC CODE 15
SAN DIEGO CA 92152

DR. HENDRICK W. RUCK
AFHRL/MODS
BROOKS AFB TX 78235

M. RUMSEY
US ARMY RESEARCH INST
5001 EISENHOWER AVENUE
ALEXANDRIA VA. 22333

ROCCO P. RUSSO
INTERAMERICA RESEARCH ASSOCIATES INC.
1555 WILSON BLVD SUITE 600
ROSSLYN VA 22209

MR. SYDNEY SAKO
CHIEF, MEASUREMENT BRANCH
OFFICER TRAINING SCHOOL (OTS/MTDM)
LACKLAND AFB TX 78236

DR. LENORE E. SALTMAN
NSETC CODE 21
NATIONAL NAVAL MEDICAL CENTER
BETHESDA MD 20817

DR. JOHN A. SANDERSON
EDUCATIONAL ADVISOR
DEPT OF THE ARMY
JUDGE ADVOCATE GENERAL'S SCHOOL
CHARLOTTESVILLE, VA 22901

WILLIAM A. SANDS
NAVY PERSONNEL RESEARCH AND
DEVELOPMENT CENTER (CODE P310)
PT. LOMA
SAN DIEGO CA 92152

MILDRED E. SARGENT
NETPDC
NAS SAUFLEY
PENSACOLA FL 32509

GARY G. SARLI
U.S. ARMY RESEARCH INSTITUTE
P.O. BOX 6057
FT. BLISS TX 79916

WILLIAM E. SCHLENGER
RESEARCH TRIANGLE INSTITUTE
P. O. BOX 12194
RESEARCH TRIANGLE PARK NC 27709

JAMES E. SCHROEDER
ARMY RESEARCH INSTITUTE
P. O. BOX 2086
COLUMBUS GA 31905

DR. MADY WECHSLER SEGAL
DEPT. OF MILITARY PSYCHIATRY
WALTER REED ARMY INSTITUTE
OF RESEARCH
WALTER REED ARMY MEDICAL CENTER
WASHINGTON DC 20012

MR. BRADFORD P. SHARP
US COAST GUARD HDQTRS
2100 2ND ST. SW
WASHINGTON DC. 20593

JUDITH S. SHELLNUT
GENERAL RESEARCH CORP.
7655 OLD SPRINHOUSE ROAD
MCLEAN VA 22102

DR. JOYCE SHIELDS
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333

LCDR W.S. SHIELDS
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
4900 YONGE ST. 0600
TORONTO, ONTARIO
CANADA M2N 6B7

WILLIAM R. SHOEN
SENIOR ED ADVISOR
CODE 02A SERVICE SCHOOL COMMAND
NAVAL TRAINING CENTER
ORLANDO, FL 32813

MAJ LAWRENCE O. SHORT, USAF
LMDC/ANC
MAXWELL AFB AL 36112

GUY L. SIEBOLD
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

SUZANNE P. SIMPSON
6 CLOVER HILL LANE
COLTS NECK NJ 07722

WILLIAM H. SIMS
CENTER FOR NAVAL ANALYSES
200 N. BEAUREGARD ST.
ALEXANDRIA, VA 22311

MARY J. SKINNER
AFHRL/MODE
AIR FORCE HUMAN RESOURCES LABORATORY
BROOKS AFB TX 78235

LCOL D.J. SLINMAN
DIRECTOR PERSONNEL DEVELOPMENT STUDIES
NATIONAL DEFENCE HEADQUARTERS
101 COLONEL BY DRIVE
OTTAWA, ONTARIO
CANADA K1A 0K2

BRANDON B. SMITH
NRDC UNIV OF MINN
B-12 FRASER HALL
106 PLEASANT ST S.E.
MINNEAPOLIS, MN 55455

COMMANDER
USAIS
ATTN: ATSH-I-V-ED
DR. ROBERT M. SMITH
FORT BENNING, GA 31905

LARRY E. SOUTH
DEPT. OF DEFENSE
9800 SAVAGE ROAD
FT. MEADE MD

DR. DANIEL E. SPECTOR
US ARMY CHEMICAL SCHOOL
ATTN - ATZN-CM-DI
FT. MCCLELLAN AL 36205

1713

PAUL P. STANLEY II
USAF OCCUPATIONAL MEASUREMENT
CENTER/OMD
RANDOLPH AFB TX 78150

DR. THOMAS G. STICHT
HUMAN RESOURCES RESEARCH ORGANIZATION
300 NORTH WASHINGTON STREET
ALEXANDRIA VA 22314

JAMES R. STOKES
US COAST GUARD HQS
WASHINGTON, D.C. 20593

DENIS J. SULLIVAN, JR.
HQ USAISD
ATTN: ATSIE-TD-AB-D
FORT DEVENS MA 01433

L. DOUGLAS SWANSON
MARYLAND DEPT. OF PERSONNEL
301 W. PRESTON STREET
BALTIMORE MD

COL JOHN E. SWINDELLS
DIRECTOR
OFFICE OF ARMOR FORCE MANAGEMENT
AND STANDARDIZATION
FORT KNOX KY

CPT L. A. TALMAGE
OMU, BLDG 2009
QUANTICO VA 22134

CPT RON TARR
DEPARTMENT OF THE ARMY
TRAINING DEVELOPMENT INSTITUTE
ATTN - ATTG-DOR
FT. MONROE VA 23651

MAJOR JOHN F. TAYLOR
RAEC CENTRE, WILTON PARK
BEACONSFIELD ENGLAND

SPENCER C. THOMASON
ESSEX CORPORATION
P.O. BOX 147
WHITE SANDS MISSILE RANGE, N.M. 88002

NANCY A. THOMPSON
AFHRL/MODS
BROOKS AFB TX 78235

MR. HARVEY L. THORSTAD
CNET N5, US NAVY
NAS, PENSACOLA, FL. 32508

JAMES A. TIPPENS
THE MARINE CORPS INSTITUTE
P.O. BOX 1775
WASHINGTON DC 20013

CDR JEANNIE K. TODARO
NODAC BLDG 150
WASHINGTON NAVY YARD
WASHINGTON DC 20374

ROGER TOURANGEAU
NATIONAL OPINION RESEARCH CENTER
461 8TH AVENUE
NEW YORK NY 10001

DR. MARVIN H. TRATTNER
U.S. OFFICE OF PERSONNEL MANAGEMENT
1900 E STREET NW
WASHINGTON, D.C. 20415

JOHN D. TUBBS
USA TRASANA ATAA-THC
WSMR, N.M. 88002

NANCY E. TULLOH
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

MR. DUANE E. TYERMAN
TS HQ, CFB TRENTON
MPO 303
ASTRA, ONTARIO
CANADA K0K 1B0

HARRY M. VALENTINE
MARYLAND DEPT. OF PERSONNEL
ROOM 508A
301 W. PRESTON STREET
BALTIMORE MD

ROBERT VINEBERG
HUMRRO
27857 BERWICK DRIVE
CARMEL CA 93923

DR. LLOYD W. WADE
TEST AND EVALUATION SECTION
MARINE CORPS INSTITUTE
P.O. BOX 1775
WASHINGTON DC 20013

MICHAEL P. WAGNER
MCFANN, GRAY & ASSOCIATES, INC.
2020 NORTH 14TH STREET SUITE 409
ARLINGTON VA 22201

DR. HOWARD WAINER
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON, NEW JERSEY 08541

DR. RAYMOND O. WALDKOETTER
ARI FIELD UNIT
P.O. BOX 16117
FT. HARRISON IN 46216

NORMAN K. WALKER
NORMAN K. WALKER ASSOC., INC.
SUITE 121
4 PROFESSIONAL DRIVE
GAITHERSBURG MD 20879

DR. JOE H. WARD, JR.
AFHRL/MPMD
BROOKS AFB TX 78235

CDR J. B. WASHBUSH
CNET - N-122,
N.A.S.
PENSACOLA FL 32508

DR. BRIAN WATERS
HUMRRO
300 N. WASHINGTON STREET
ALEXANDRIA VA 22314

MIRIAN J. WATT
ATSH-I-V-ED
U.S. ARMY INFANTRY SCHOOL
FORT BENNING GA 31905

1714

MAJ JOHN R. WELSH, JR.
AIR FORCE MANPOWER AND
PERSONNEL CENTER, MPCYPT
RANDOLPH AFB TX 78148

CHARLES M. WEST
CNET CODE N525
NAS PENSACOLA, FL 32508

LEONARD WEVRICK
STATISTICS DIVISION
MINISTRY OF THE SOLICITOR GENERAL
340 LAURIER AVE. W.
OTTAWA, ONTARIO
CANADA K1A 0P8

CPT E. WILLIAMS
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333

RAYBURN A. WILLIAMS
CHIEF OF NAVAL EDUCATION AND TRAINING
CODE N51
PENSACOLA FL 32508

LUCY D. WILSON
BOOZ-ALLEN
400 MARKET STREET
PHILADELPHIA PA 19106

MR. CASIMER WINIEWICZ
522 N. WHITE STATION RD.
MEMPHIS, TN 38117

LCDR FRANK J. WINN JR.
OUTPATIENT CLINIC USCG
GOVERNORS ISLAND
NEW YORK NY 10004

MARIA BELLANTONI WINSTON
US ARMY SOLDIER SUPPORT CENTER
200 STOVALL STREET
ALEXANDRIA VA 22332

DEPARTMENT OF THE NAVY
ATTN- DR. ROBERT WISHER
NPRDC, CODE 13
SAN DIEGO CA 92152

DR. MARTIN F. WISKOFF
NAVY PERSONNEL RESEARCH & DEVELOPMENT CENT
POINT LOMA
SAN DIEGO, CA 92115

DR. BOB G. WITMER
ARI FIELD UNIT
STEELE HALL
FORT KNOX KY 40121

GILBERT F. YARD
C/O METHODOLGY & EVALUATION OF
TRAINING UNIT
RCM POLICE
250 TREMBLAY ROAD
OTTAWA, ONTARIO
CANADA K1A 0R2

DR. JOSEPH ZEIDNER
TECHNICAL DIRECTOR
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA VA 22333




EOF:786