AD-A172 851

# PROCEEDINGS

## 27th Annual Conference of the

# MILITARY TESTING ASSOCIATION

## Coordinated by the NAVY PERSONNEL RESEARCH & DEVELOPMENT CENTER

**SAN DIEGO, CALIFORNIA**
**21 - 25 OCTOBER 1985**

DTIC FILE COPY

This document has been approved for public release and sale, its distribution is unlimited

DTIC
ELECTE
OCT 3 1986
S
A

## Volume II

86 10 3 049

# DISCLAIMER NOTICE

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

PROCEEDINGS

27TH ANNUAL CONFERENCE

of the

MILITARY TESTING ASSOCIATION

Coordinated by

the

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

SAN DIEGO, CALIFORNIA

21-25 OCTOBER 1985

VOLUME II

Partial

CONTENTS:

VOLUME II


| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Basic Skills Training**<br>Chairperson: Konoske, P.J.<br>Navy Personnel Research and Development Center | | |
| | Pre-Course Mathematics Skills and Success on Advanced Naval Technical Courses | Amyot, CAPT K.A. | 462 |
| | Causes of Performance in Air Force Initial Skills Training | Harding, F.D.<br>Mumford, M.<br>Weeks, J.L. | 467 |
| PAPER SESSION: | **Methodological and Technological Developments in Survey Research**<br>Chairperson: Morrison, R.F.<br>Navy Personnel Research and Development Center | | |
| | An Examination of the Significance of Non-Participation in Field Studies | Potter, E.H. | 473 |
| | Advancing Survey Design through Technology | Doherty, L.M. | 480 |
| | Implementation of a Civilian Automated Survey System | Ripkin, F.L.<br>Cecil, S. | 486 |
| PAPER SESSION: | **Utilization and Validation of Biodata**<br>Chairperson: Mattson, J.D<br>Navy Personnel Research and Development Center | | |
| | The Utility of Educational and Biographical Information for Predicting Military Attrition | Riegelhaupt, B.J.<br>Bonczar, T P. | 491 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| | Three Variables that May Influence the Validity of Biodata | Walker, C.B. | 497 |
| | Use of Personal History Information to Predict Naval Academy Disenrollment | Mattson, J.D.<br>Abrahams, N.M.<br>Hetter, R.D | 503 |
| | Intercorrelations of Biographical Information, Aptitude Test Scores, and Job Performance Ratings for 108 Occupations | Trattner, M.H. | 509 |
| PAPER SESSION: | **Training Device Evaluation**<br>Chairperson: Konoske, P.J.<br>Navy Personnel Research and Development Center | | |
| | Analytic Prediction of Training Device Effectiveness | Yates, L.G.<br>Macpherson, D. | 515 |
| | Research on Decision Aids for Training Design and Evaluation | Mirabella, A. | 520 |
| PAPER SESSION: | **Attitude Surveys**<br>Chairperson: Wilcove, G.L.<br>Navy Personnel Research and Development Center | | |
| | The Army Experience Survey: Methodological Highlights | Celeste, J.F. | 526 |
| | USAF Spouse Survey. The Final Chapter | Dansby, MAJ M.R.<br>Ibsen, CPT K.A. | 531 |
| | Predictors of Propensity for Continuing Education among Army Chaplains | Pierce, J.E.<br>Goldman, L.A. | 537 |


QUALITY INSPECTED 1

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Selection Methodology**<br>Chairperson: Moreno K.E.<br>Navy Personnel Research and Development Center | | |
| | Composite Ratings as a Performance Criterion | Warren, J.<br>Newton, P.<br>Bondaruk, J. | 543 |
| | Maximizing Criterion Variance in Validation Research: Is This Always Best? | Angus, CAPT R.J.<br>Ellis, MAJ R.T. | 547 |
| SYMPOSIUM: | **Remedial Education in the Navy**<br>Chairperson: Chang, F.R.<br>Navy Personnel Research and Development Center | | |
| | Literacy, Readability and Knowledge | Chang, F.R. | 553 |
| | The XFSP: Reading and Mathematics Project | Sticht, T.<br>Armijo, L.<br>Koffman, N.<br>Roberson, K. | 559 |
| SYMPOSIUM: | **Elements of a Military Occupational Exploration System**<br>Chairperson: Pass, J.J.<br>Navy Personnel Research and Development Center | | |
| | A Conceptual Framework for Occupational Exploration | Pass, J.J. | 565 |
| | Development of the Career Maturity Assessment | Diamond, E.E. | 570 |
| | Navy R&D in Support of the Army JOIN System: Leveraging the Government Research Dollar | Baker, H.G.<br>Ratacz, B.A.<br>Sands, W.A. | 576 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| | Vocational Interests as Predictors of Army Performance | Wing. H.<br>Barge, B.N.<br>Hough, L.M. | 582 |
| | Army Career Vocational Guidance as a Recruiting Tool | Hertzbach, A.<br>Knapp, D.J.<br>Johnson, R.M. | 587 |
| SYMPOSIUM: | **Determining Ability Requirements**<br>Chairperson: Abrabian, J.M.<br>U.S Army Research Institute | | |
| | Developing New Attribute Requirements Scales for Military Jobs | Smith, E.P. | 593 |
| | Methodological Problems in Identifying Ability Requirements Related to Soldier Performance | Miller, C.R. | 599 |
| | Comparison of Weapon Systems Using Ability Requirements Scales | Arabian, J.M. | 603 |
| | Computerized Approaches for Estimating Ability Requirements | Rossmeissl, P.G. | 609 |
| PAPER SESSION: | **Team Performance Measurement**<br>Chairperson: Flaningam, M R<br>Navy Personnel Research and Development Center | | |
| | Chaparral Crew Performance in the Realistic Air Defense Engagement System | Sarli, G G.<br>Johnson, D.M.<br>Lockhart, J.M. | 615 |
| | Command and Control Teams: Techniques for Assessing Team Performance | Cooper, M.<br>Shiflett, S.<br>Korotkin, A.L. | 621 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Analysis of the ASVAB**<br>Chairperson: Larson, G.E.<br>Navy Personnel Research and Development Center | | |
| | Iter Factor Analysis of ASVAB 14 | McCormick, C. | 627 |
| | Alternate Forms Reliability of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10 | Stern, B.M.<br>White, L.A.<br>Wing, H.<br>Sachs, S.A. | 632 |
| PAPER SESSION: | **Personality Assessment**<br>Chairperson: Bjerke, CDR D.G.<br>Navy Personnel Research and Development Center | | |
| | Psychometric Properties of the Safety Locus of Control Scale | Jones, J.W.<br>Wuebker, L.J. | 637 |
| | Health Locus of Control Beliefs among Infantrymen | Grieger, L.W. | 644 |
| | Classifying Military Offenders: Application of the Megargee MMPI Typology | Paris, M.L.<br>Brown, G.E. | 650 |
| PAPER SESSION: | **Team Training**<br>Chairperson: Flaningam, M.R.<br>Navy Personnel Research and Development Center | | |
| | Automated Scheduling of Army Unit Training | Goehring, D J.<br>Hart, R.J. | 655 |
| | Supporting the Changing Role of Army Collective Training Developers | Meliza, L L. | 661 |
| | Application of Model Aircrew Training System (MATS) to B-52 Combat Crew Training | Bills, MAJ C.G.<br>Nullmeyer, R.T. | 666 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| | Vocational Interests as Predictors of Army Performance | Wing, H<br>Barge, B.N.<br>Hough, L.M. | 582 |
| | Army Career Vocational Guidance as a Recruiting Tool | Hertzbach, A<br>Knapp, D.J.<br>Johnson, R.M. | 587 |
| SYMPOSIUM: | Determining Ability Requirements<br>Chairperson: Abrahian, J M.<br>U S Army Research Institute | | |
| | Developing New Attribute Requirements Scales for Military Jobs | Smith, E.P. | 593 |
| | Methodological Problems in Identifying Ability Requirements Related to Soldier Performance | Miller, C.R. | 599 |
| | Comparison of Weapon Systems Using Ability Requirements Scales | Arabian, J M | 603 |
| | Computerized Approaches for Estimating Ability Requirements | Rossmeissl, P.G. | 609 |
| PAPER SESSION: | Team Performance Measurement<br>Chairperson: Flaningam, M.R.<br>Navy Personnel Research and Development Center | | |
| | Chaparral Crew Performance in the Realistic Air Defense Engagement System | Sarli, G.G.<br>Johnson, D M.<br>Lockhart, J.M. | 615 |
| | Command and Control Teams: Techniques for Assessing Team Performance | Cooper, M.<br>Shiflett, S.<br>Korotkin, A.L. | 621 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Analysis of the ASVAB**<br>Chairperson: Larson, G.E.<br>Navy Personnel Research and Development Center | | |
| | Item Factor Analysis of ASVAB 14 | McCormick, C. | 627 |
| | Alternate Forms Reliability of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10 | Stern, B M.<br>White, L.A.<br>Wing, H<br>Sachs, S.A. | 632 |
| PAPER SESSION: | **Personality Assessment**<br>Chairperson: Bjerke, CDR D.G.<br>Navy Personnel Research and Development Center | | |
| | Psychometric Properties of the Safety Locus of Control Scale | Jones, J.W.<br>Wuebker, L.J. | 637 |
| | Health Locus of Control Beliefs among Infantrymen | Grieger, L.W. | 644 |
| | Classifying Military Offenders: Application of the Megargee MMPI Typology | Paris, M.L.<br>Brown, G.E. | 650 |
| PAPER SESSION: | **Team Training**<br>Chairperson: Flaningam, M.R.<br>Navy Personnel Research and Development Center | | |
| | Automated Scheduling of Army Unit Training | Goehring, D J.<br>Hart, R.J. | 655 |
| | Supporting the Changing Role of Army Collective Training Developers | Meliza, L L. | 661 |
| | Application of Model Aircrew Training System (MATS) to B-52 Combat Crew Training | Bills, MAJ C.G.<br>Nullmeyer, R.T. | 666 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Analysis and Interpretation of the ASVAB**<br>Chairperson: Moreno, K.E.<br>Navy Personnel Research and Development Center | | |
| | Sensitivities of Speeded Subtests | Wegner, T.G.<br>Ree, M.J. | 672 |
| | Initial Operational Test and Evaluation of Armed Service Vocational Aptitude Battery (ASVAB) Forms 11, 12, and 13: Data Quality Analysis | Welsh, J.R.<br>Wegner, T.G. | 676 |
| PAPER SESSION: | **Integrated Selection Systems**<br>Chairperson: Bruni, J.R.<br>Navy Personnel Research and Development Center | | |
| | Development of an Integrated Pilot Selection System | Kantor, J.E. | 680 |
| | Selection of Skilled Maintenance Employees in the U.S. Postal Service | Mueller, H.A. | 686 |
| | The Effects of the Flight Screening Program on Attrition in Undergraduate Pilot Training | Quebe, J.C. | 695 |
| PAPER SESSION | **Issues in Test Item Design**<br>Chairperson: Flaningam, M R.<br>Navy Personnel Research and Development Center | | |
| | What Performance Does a Performance Test Test? | Ansbro, T.M. | 701 |
| | How Qualitatively Informative are Test Items?: A Dense Item Analysis | Bart, W.M. | 707 |
| | Test-item Readability: How the Variables Work | Duncan, CPT R.E. | 713 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Entry Standards Based on the ASVAB**<br>Chairperson: Vicino, F.L.<br>Navy Personnel Research and Development Center | | |
| | A Comparison of Service Job Standards for Four Military Specialties | Waters, B.K. | 719 |
| | Exploring a Statistically Viable Assignment Basis Using ASVAB | Schwartz, M.M. | 725 |
| | The Validity of ASVAB for Predicting Training and SQT Performance | Rossemeissl, P.G.<br>McLaughlin, D.H.<br>Wise, L.L.<br>Brandt, D.A. | 730 |
| PAPER SESSION: | **Military Attrition Research**<br>Chairperson: Cory, C.H.<br>Navy Personnel Research and Development Center | | |
| | Gender, Ethnic Group, Aptitude and Personality Determinants of U.S. Coast Guard Attrition | Frey, R.L. | 736 |
| | Study of Wastage from the Territorial Army (UK Army Reserve) | Blyth, D.M. | 741 |
| | Diminishing Returns from an Assessment Centre | Hardy, G.R. | 747 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Officer Training and Testing**<br>Chairperson: Greebler, C.S.<br>Navy Personnel Research and Development Center | | |
| | Increasing Basic Skill Levels of Moderate Risk Officer Candidates Prior to Commissioning in Historically Black Colleges | Sprenger, W.D. | 753 |
| | Direct Mesurement of ROTC Cadets' Writing Skills | Hanlon, J.P. | 759 |
| | Procedures for Refining Written Measurements at USAF OTS | Sako, S.<br>Slaughter, LTC W.J. | 763 |
| PAPER SESSION: | **Job Performance Measurement I**<br>Chairperson: Kroeker, L.P.<br>Navy Personnel Research and Development Center | | |
| | Utility Estimation in Five Enlisted Occupations | Eaton, N.K.<br>Wing, H.<br>Lau, A. | 769 |
| | Relation of Mental and Education Levels to Navy Enlisted Performance | Cory, C.H. | 775 |
| | Selecting Critical Tasks for a Radioman Hands-on Performance Test | Baker, H.G.<br>Laabs, G.J. | 780 |
| | Cognitive Predictors of M1 Tank Gunner Performance | Black, B.A. | 786 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Issues in Instruction I**<br>Chairperson: Windisch, D.J.<br>Navy Personnel Research and Development Center | | |
| | A Retrospective Analysis of Instructional Technology Innovation Using General System Theory | Ekstrom, A.G. | 792 |
| | A Systems Approach to Evaluating High Technology Training | Atwood, N.K. | 798 |
| | An Analysis of Attitudes toward Instruction among Vocational Education Instructors | Usova, G.M. | 804 |
| PAPER SESSION: | **Assessment of English-as-a-Second Language Personnel**<br>Chairperson: Hetter, R.D.<br>Navy Personnel Research and Development Center | | |
| | Relationship of an Experimental Hispanic Enlistment Screening Test to AFQT | Mathews, J.J.<br>French, C.M. | 809 |
| | Attrition and Performance Ratings of ESL Soldiers in BT | Rosenbaum, H. | 815 |
| PAPER SESSION: | **Job Performance Measurement II**<br>Chairperson: Kroeker, L.P.<br>Navy Personnel Research and Development Center | | |
| | Standard Setting Methods for Skills Qualification Tests (SQTs) | Pettie, A.L. | 821 |
| | Assessing Tank Commander and Gunner Proficiency on U-COFT | Graham, S.E.<br>Boldovici, J A. | 826 |
| | Criterion Referenced Testing for the U.S. Navy's Nuclear Submarine Fleet | Cantor, J.A.<br>Walker, L. | 832 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| PAPER SESSION: | **Issues in Instruction II**<br>Chairperson: Windisch, D.J.<br>Navy Personnel Research and Development Center | | |
| | A Study of Suggestopedia at the Defense Language Institute Foreign Language Center (DLIFLC) | Bush, B.J. | 838 |
| | Heat, Chemical Protective Clothing and Sustained Cognitive Performance | Fine, B.J.<br>Kobrick, J.L. | 843 |
| SYMPOSIUM: | **Exceptional Recruits: A Look at High and Low Aptitude Personnel**<br>Chairperson: Lancaster, A.R.<br>Office of the Assistant Secretary of Defense (FM&P) | | |
| | Overview | Lancaster, A.R. | 849 |
| | Using Military High- and Low- Aptitude Recruits Wisely | Sellman, W.S. | 852 |
| | When Low-Aptitude Recruits Succeed | Means, B.<br>Nigam, A.<br>Heisey, J.G. | 855 |
| | Enlistment and Utilization of High-Aptitude Recruits | Laurence, J.H.<br>Schneider, E.F. | 861 |
| SYMPOSIUM: | **Predicting a Broad Variety of Criteria: Elaborating the Predictor Space**<br>Chairperson: Wing, H.<br>U.S. Army Research Institute | | |
| | Mapping Predictors to Criterion Space: Overview | Peterson, N.G. | 867 |
| | Using Microcomputers for Assessment: Practical Problems and Solutions | Rosse, R.L.<br>Peterson, N.G. | 873 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| | Computerized Assessment of Perpetual and Psychomotor Abilities | McHenry, J.J.<br>Toquam, J L. | 879 |
| | Adding to the ASVAB: Cognitive Paper-and-Pencil Measures | Toquam, J.L.<br>Dunnette, M.D.<br>Corpe, V.A.<br>Houston, J. | 885 |
| | Measuring Personnel Attributes: Temperament, Biodata and Interest | Hough, L.M.<br>McGue, M.K.<br>Kamp, J.D.<br>Houston, J.S.<br>Barge, B.N. | 891 |
| SYMPOSIUM: | **The Training and Selection of Army Managers: Quantitative/Qualitative Approaches**<br>Chairperson: Fisher, G P.<br>Human Resources Research Organization | | |
| | Overview | Fisher, G.P.<br>Lilienthal, R. | 897 |
| | Selecting and Training Logistics Professionals and Managers: Qualitative/ Quantitative Approaches | Fisher, G.<br>Lilienthal, R.<br>Hough, L. | 900 |
| | Managerial Competencies Assessment For Army Civilians | King, G. C. | 906 |
| SYMPOSIUM: | **Army Research Institute R&D Program on the National Training Center**<br>Chairperson: Banks, J.H.<br>U S. Army Research Institute | | |
| | Types and Quality of National Training Center Data | Whitmarsh, P.J. | 912 |
| | An Overview of ARI's Research Program on the National Training Center | Banks, J.H. | 916 |

| SESSION | TITLE | AUTHOR | PAGE |
|---|---|---|---|
| | Leader Performance Criteria at the National Training Center (NTC) | Pence, E.C. | 922 |
| | Battalion Performance on the Live Fire Range at the National Training Center (NTC) | Forsythe, T.K.<br>Doherty, W.J. | 928 |
| | Analysis of NTC Force-on-Force Performance | Nichols, J.J.<br>Doherty, W.J. | 931 |
| | Use of Instrumentation to Improve Combat Readiness | Shackelford, W.L. | 937 |
| SYMPOSIUM: | **Two Systems for Computer Analysis of Student Feedback**<br>Chairperson: Ekwall, R.W.<br>U.S. Army Command and General Staff College | | |
| | Collecting and Utilizing Feedback: Combined Arms and Services Staff School | Ekwall, R.W. | 943 |
| | **By-laws of the Military Testing Association** | | 949 |
| | **Minutes of the Steering Committee** | | 953 |
| | **Harry H. Greer Award** | | 956 |
| | **MTA 27th Annual Conference Staff** | | 957 |
| | **Military Testing Association Conference Registrants** | | 958 |
| | **Author Index** | | 970 |

# PRE-COURSE MATHEMATICS SKILLS AND SUCCESS ON ADVANCED NAVAL TECHNICAL COURSES

**Captain Kenneth A. Amyot**

**Canadian Forces Personnel Applied Research Unit**
**Willowdale, Ontario, Canada**

## Background

Training for Naval trades in the Canadian Forces (CF) has become increasingly more expensive and lengthy in recent years. High levels of attrition on technically advanced weapons and electronics courses have become a cause for concern. To place the attrition problem into perspective, it is necessary to briefly describe the selection and training process.

Although Navy recruits are enrolled directly into electronics and weapons technician trades, their initial or basic training, called Trade Qualification (TQ) 3, is designed to prepare them to become junior operators of the naval combat systems equipment associated with their assigned trades. This is followed by two to three years of employment at sea during which time they qualify to the next higher level, TQ4, through on-the-job training. At this point, technical training to prepare them for employment as maintainers of equipment is provided as a TQ5 formal course.

In most CF trades, TQ5 training is a logical extension of TQ3 training, with success on basic (TQ3) training being a good predictor of success on more advanced (TQ5) training. For this reason, selection standards have been developed and validated on the basis of TQ3 performance. High success rates on Naval TQ3 courses indicate that the entry standards work well for that purpose. However, because of the disjuncture associated with the Naval "user/maintainer" concept, the aptitudes required for success at the higher levels may well be different than for the basic level.

Advanced Naval technician training is substantially academic, emphasizing mathematics. Aptitudes used for selection may include ability to learn, but not necessarily achievement; i.e., whether the material has already been acquired. Training development authorities at the CF Fleet School in Halifax, Nova Scotia, where these courses are conducted, attribute the high TQ5 attrition directly to a lack of sufficient mathematics skills prior to training. To address this deficiency, a three

---

The views and opinions expressed in this paper are those of the author, and not necessarily those of the Department of National Defence.

week Pre-Academic Qualifying Course (PAQC) was introduced to immediately precede the first semester of the TQ5 course. Although part of the purpose of the PAQC was to bring tradesmen up to the required threshold level of mathematics, in reality it serves a screening function. It has been increasingly difficult and expensive to continue to operate PAQC training. The Canadian Forces Personnel Applied Research Unit (CFPARU) was tasked to explore the possibility of a cost-effective substitute for PAQC training.

## Remedial Training

Cognizant of the deficiency in pre-course mathematics skills, the CF Fleet School devised a number of remedial measures. A self-study Programmed Instructional Package (PIP) was found to be ineffective. A Computer Assisted Learning (CAL) Laboratory was established to allow students to upgrade their mathematical knowledge during off-duty time before or during TQ5 training. While effective, the CAL facility (located only on the East coast) was not readily accessible to all students. The scope of this project was, therefore, expanded to include an examination of means to better integrate the CAL facility within the training process.

## Canadian Achievement Tests

Research was initiated to identify standardized, commercially available tests to measure competence in mathematics and to determine their ability to predict success on the PAQC and follow-on TQ5 courses. The tests selected were the two mathematics subtests of the Canadian Achievement Tests (CAT) (Canadian Test Centre, 1981). The California Achievement Tests, a widely used test battery in the United States, provided the initial pool of items for the CAT. The Canadian version met the criterion of being achievement tests for which Canadian national norms were available. The Canadian norms are based on a sample of 76,000 students randomly selected from across the country.

The first subtest, "Computation", relates to the four basic functions of addition, subtraction, multiplication and division. The second subtest, "Concepts and Applications", includes problems covering such areas as number theory, scales, measurement/graphs, geometry, problem solving, fractions, and rounding and estimating. The content of the subtests covers the full range of mathematics taught in grades nine through twelve. The subtest scores, and their combined total, may be used to obtain both norm-referenced and criterion-referenced descriptions of achievement.

## Analysis

The two CAT subtests were administered to 166 Naval Electronics and Naval Weapons Technician students undergoing training at Canadian Forces Base (CFB) Halifax, Nova Scotia and CFB Esquimalt, British Columbia, during the first day of their PAQC training. A supplementary questionnaire, designed to obtain information on the students' mathematics background, was also administered. Pearson product-moment correlations were computed between each CAT subtest score and the total of these two tests, and the

results of the PAQC mathematics examinations and final overall course results. As well, grades obtained on mathematics examinations administered at the end of Semester I of the TQ5 course were correlated with CAT scores and PAQC results. Results in Table 1 indicate that the CAT is a valid predictor of success on PAQC and TQ training. Since the CAT predicts PAQC success, and since the PAQC predicts Semester I success, it was concluded that the PAQC could reasonably be eliminated in favour of CAT-testing.

Table 1

Inter-Correlations of CAT Scores,
PAQC and Semester I Performance Measures

| | CA | Total | PAQC Math | PAQC Grade | Semester I Math |
|---|---|---|---|---|---|
| Comp | .76 | .94 | .75 | .65 | .50 |
| CA | - | .93 | .71 | .61 | .62 |
| Total | | - | .78 | .69 | .61 |
| PAQC Math | | | - | .90 | .70 |
| PAQC Grade | | | | - | .75 |

Notes: Sample sizes vary from 56 to 166 due to missing data.
All correlations are significant at the .001 level.

Reported Mathematics Education Versus CAT Scores

The correlation between grade equivalents measured by the CAT and grades levels reported on the supplementary questionnaire was low (r=.34). This correlation was calculated in several ways, controlling for other variables which might have affected performance such as academic upgrading since enrolment, use of the CAL Laboratory, academic stream (university preparatory or general/terminal), and type of program (business/ commercial, technical/vocational, or academic). This correlation was never higher than .40. Undoubtedly, other factors are involved, such as variability in inter-provincial standards (Ellis & Amyot, 1984), number of mathematics courses taken, and the interval between school leaving and CAT-testing. This points out the unreliability of grade levels reported at Recruiting Centres (even if accurate reporting is assumed) and the problems associated with using these as selection standards. To illustrate, several examinees reported a pre-testing mathematics education of grade nine or less, but their CAT grade

equivalents were in the grade twelve range. Similarly, several examinees who reported post-secondary mathematics at the time of enrolment scored less than grade nine equivalent on the CAT.

Diagnosis Versus Selection

Given the validity of the CAT for predicting success on PAQC and TQ training, the problem became one of when and how to use the CAT, since the primary purpose of the PAQC is screening, not teaching. One proposal was to use the CAT as a means of selecting recruits for Naval technician trades, by administering them during processing at CF Recruiting Centres. However, this was considered problematic for three reasons:

a. knowledge/proficiency measured at the time of enrolment can easily have deteriorated by the time the sailor reaches the point where he will proceed on the advanced training, which may be as long as three years later;

b. opportunities and programs exist for sailors deficient in mathematics to upgrade these skills between enrolment and commencement of advanced training; and,

c. testing time at Recruiting Centres is a major consideration, and the computation subtest alone requires at least twenty-eight minutes to administer.

This still left open the possibility of using the CAT at some point after enrolment to identify those students who are weak in mathematics. It was thought that the CAT might be administered, for example, during the preceding TQ course to identify potential failures on semester training, at which point a compulsory remedial program could be scheduled for these individuals.

Computer Assisted Learning Laboratory

Clearly, the existing Computer Assisted Learning (CAL) Laboratory, could be utilized to deliver the remedial program. The CAL Laboratory uses computer assisted instruction and a course authoring system developed jointly by the Ontario Institute for Studies in Education (OISE) (Gershman & Sakamoto, 1982) and Honeywell Information Systems (Honeywell, 1983) which can be conducted on many kinds of minicomputers. This particular system has been used successfully by 55 licencees in ten countries, including the Training and Development Branch of the Internal Revenue Service, Austin, Texas, who report a 60% reduction in training time compared to previous methods, thereby saving thousands of hours of student time with corresponding cost savings. There is a large collection of fully-validated courseware available including prerequisite, technical and academic mathematics. Most of the courses are tutorial in nature, intended to individualize instruction by diagnosing skill deficiencies and providing lessons as required. Techniques such as random test item generation, immediate answer analysis and feedback, and performance-dependent routing through a hierarchical structure of objectives, are used. Sea trials of the system (Shipboard Mathematical Refresher Training

(SMART)) have been conducted aboard a destroyer using a microcomputer. The CAL Laboratory is also used to provide computer-based instruction on advanced Marine Engineering training. However, as previously noted, participation in the remedial aspect of the CAL Laboratory has been on a voluntary basis.

Integration of CAT-Testing with CAL

CAT-testing is clearly a cost-effective alternative as a screening device, and has several additional advantages. The first is that it provides the potential to identify much more precisely the level and the specific skill/knowledge deficiencies of the TQ5 candidate than is possible with the PAQC. Second, it is simpler to assess a large number of candidates at various times and locations using a one hour test than it is to conduct a three-week course. This means that all TQ5 candidates can be assessed when and where they are available and at a point in time where an effective remedial program can be put in place. There is a considerable investment on the part of the CF and the individual during the years between enrolment and TQ5 training. The aim, therefore, should be to capitalize on this investment by bringing students to a point where they can succeed on this demanding, lengthy and expensive training. CAT-testing cannot, therefore, merely be substituted for the PAQC. There must be provision for remediation. While the CAL Laboratory has the potential to satisfy this requirement, it must be part of a coordinated program to be effective and efficient. The integration of CAT-testing and the CAL Laboratory should prepare students to pass the academic portions of TQ5 courses.

REFERENCES

Canadian Test Centre (1981). Canadian Achievement Tests. Scarborough, Ontario: McGraw-Hill Ryerson Ltd.

CAN-8 Instructional System: Summary description (1983). Willowdale, Ontario: Honeywell Information Systems Inc.

Ellis, R.T., & Amyot, K.A. (1984). Revised selection standards for restructured naval trades (Technical Note 14/84). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Gershman, J.S., & Sakamoto, E.J. (1982). Computer assisted remediation and evaluation: Final report. Toronto, Ontario: Ontario Institute for Studies in Education.

# Causes of Performance in Air Force Initial Skills Training

Francis D. Harding and Michael Mumford
Advanced Research Resources Organization

Joseph L. Weeks
Air Force Human Resources Laboratory

This paper reviews a study in which a system for assessing the impact of aptitude requirement adjustments on Air Force initial skills training was developed (Mumford, Weeks, Harding & Fleishman, 1985). As a result of the modeling process that was carried out, insights were gained into the initial skills training process to show how student attributes and course content variables interact to affect training outcomes.

In a previous but related research effort, the Air Force has developed an index of occupational learning difficulty defined as the time required to learn to satisfactorily perform the tasks required in an occupation. Subsequently, it has been suggested that the Air Force manpower allocation system might be improved if the Occupational Learning Difficulty Index was used as a frame of reference in establishing aptitude minimums. However, an analysis by Weeks (1984) indicated that the realignment of aptitude minimums in accordance with occupational learning difficulty would lead to marked shifts in aptitude minimums for several specialties. Thus, the concern for more definitive information about the possible impact of shifts in aptitude requirements before instituting such changes was expressed by Air Force managers.

## Method

In reviewing the relationship between cognitive abilities and training outcomes, it became apparent that the effects of aptitude on training outcomes could be captured only by considering in addition to aptitude a variety of other student attributes as well as the potential interaction of aptitude and related variables with the training processes associated with particular training courses. Thus, it was decided that a multivariate path analysis approach would be the most appropriate means capable of incorporating a variety of student inputs, course content, and outcome variables.

The first step in the multivariate modeling effort was to define the major outcomes of initial skills training courses along with the student input and course content variables most likely to influence these outcomes. Also, at this time an attempt was made to identify potential measures of these variables and to identify any biases or limitations that might affect the general application of these measures. To obtain information needed to identify the potential variables, a series of structured interviews was conducted at several Air Force technical training centers. At each site variables indicative of the quality of student training performance; student input and traing content variables that might influence training performance; probable relationships between course content and student input variables; potential measures of student input, course content, and training outcome variables; potential sources and biases in these measures. As a result of

these meetings 7 student input, 16 course content, and 7 outcome variables were identified. The variables and their measures are shown in Table 1.

Table 1

Summary of Variables and Measures

| Outcome Variables | Measure |
|---|---|
| Quality of Student Performance | Average of end of block tests |
| SIA Time | Number of hours of special individual assistance (SIA) |
| Academic Counseling | Number of academic counseling sessions |
| Nonacademic Counseling | Number of nonacademic counseling sessions |
| Retraining | Number of hours of retraining |
| Academic Attrition | Elimination for academic reasons |
| Nonacademic Attrition | Elimination for nonacademic reasons |
| **Student Input/Variables** | |
| Aptitude | Scores on selector aptitude index of ASVAB |
| Reading Level | Total score on the Air Force Reading Abilities Test |
| Academic Motivation | Number of difficult courses taken |
| Simple Interest | Received guaranteed specialty |
| Preference Interests | Received preferred specialty |
| Educational Level | Highest educational level attained |
| Educational Preparation | Recommended high school course prerequisites taken |
| Age | Years from birth |
| **Course Content Variables** | |
| Course Length | Total instructional hours in course |
| Day Length | Length of academic day |
| Student-Faculty Ration | Number of students per teacher |
| Instructor Quality | Average of instructor performance evaluations |
| Instructor Experience | Average months of instructor experience |
| Use of Aids | Number of instructional aids divided by course length |
| Hands-on Instruction | Hours of hands-on instruction divided by course length |
| Amount of Feedback | Number of evaluations divided by course length |
| Amount of Practice | Course length divided by number of its units |
| Reenlistment Bonus | Availability of selective reenlistment bonus |
| Yearly Flow | Number of students trained yearly |
| Occupational Difficulty | Overall learning difficulty of occupational tasks |
| Reading Difficulty | Reading grade level of course materials |
| Abstract Knowledge | Ratings of abstract knowledge required in course |
| Diversity | Number of units in course |
| Expected Attrition Rate | Expected attrition rate |

## Sample

Selection of courses to be included in the study was based on several criteria which reflected the pragmatic motivations underlying the study. Among the selection guidelines were that the courses should be representative of all Air Force initial skills courses; the four aptitude areas should be represented, there should be adequate variance in the minimum aptitude levels required for entry into the courses, high cost courses and courses with high student flow were to be chosen. The sample of trainees was chosen to include all who had entered the courses in the most recent six months

with a minimum of 50 students for each course. These procedures resulted in a total of 5,970 students in 48 courses. Thirty-eight of the courses and 5,081 of the students were used in the model development sample with the remaining courses and students held cross-validation sample.

## Procedure

Once data collection and coding had been accomplished, preliminary data analyses were conducted in which it was assumed that all course content variables could be applied to all students in each course and that students with missing data for a variable would be omitted only from analyses which involved that variable. Descriptive statistics and bivariate correlations among all variables were obtained. This information provided a framework for reviewing the initial conceptual model and constructing a hypothetical model interrelating the student input, course content, and training outcome variables. This hypothetical model was used to generate the path specifications to be employed in a LISREL V analysis of the correlations obtained in the model development sample. The goodness of fit test and residual term generated by the LISREL V analysis were used to assess the extent to which the hypothetical model provided an adequate description of the interrelationships among the variables in the model. The multiple Rs generated by the model against each of the dependent variables were obtained along with the standardized path and regression coefficients generated in optimizing the fit between the hypothetical model and the observed correlation matrix. The regression and path coefficients provide a basis for assessing the causal impact of the student input and course content variables on training outcomes.

## Results

Inspection of the bivariate correlations among all the variables which have import for understanding the causes of performance disclosed some interesting findings about Air Force initial skills training. First, the two interest measures produced a weak pattern of relationships which were hard to interpret so these measures were dropped from further analysis. The remainding student input variables, such as aptitude and student motivation, displayed positive relationships of about .50 and correlations with achievement test grades of between .20 and .35. Among the training outcome variables it was found that assessed quality of performance had negative relationships with the other training outcomes as would be expected given the impact of performance on counselling retraining and attrition. It was also found that counselling and attrition had weak relationships with student input and course content variables while assessed quality of performance and SIA time displayed moderate relationships with the input and course variables. This suggests that the effects of student and course attributes on outcomes are moderated through assessed quality of performance and SIA time.

Among the course content variables, instructor quality, instructor experience, amount of feedback, and course length were positively related to assessed quality of performance. Negative relationships with quality of performance were found for length of the academic day, student faculty ratio, student flow, and amount of practice.

Interpretation of the relationships among the course content variables led to the conclusion that occupational difficulty, manpower requirements and course subject matter difficulty were the prime determinants of the

nature of the training course. While occupational difficulty could be measured directly, manpower requirements and course subject matter difficulty represent latent variables that were defined through observed variables. As might be expected, manpower requirements were best defined by yearly flow, while course diversity, expected attrition rate, reading difficulty, and abstract knowledge requirements defined subject matter difficulty.

After the hypothetical model was refined on the basis of the obtained intercorrelations from the model development sample it was analyzed by LISREL V and found to provide an excellent fit to the observed relationships among the student input, course content, and training outcome variables within the model development sample as indicated by a goodness of fit index of .59 and residual of .19 that were obtained. A multiple R, training outcomes, and the predictive power of the model is shown by the .75 was obtained assessed quality of performance. Multiple correlation coefficients of .60 and .50 were obtained against academic and non-academically counselling respectively while retraining time, academic attrition and non-academic attrition produced Rs of .76, .83, and .59 respectively. The weakest prediction of an outcome variable was for SIA time which yielded an R of .35.

The standardized path coefficients generated by the LISREL V program are presented in Figure 1. In interpreting this schematic diagram of the model, the first of the three arrows entering the course parameter variables, e.g., course length, day length, student faculty ratio, etc., represents the effect of subject matter difficulty, the second arrow reprosents the effect of occupational difficulty, while the third arrow represents manpower requirements. Of the two arrows leaving the course parameter boxes, the first shows its impact on assessed quality of performance, the second on SIA time.

Inspection of the schematic diagram of the model indicates that the student input variables produced sizeable causal paths against assessed quality of performance: .16, .16, and .14 for aptitude, reading level, and academic motivation respectively. This suggests that other student input variables in addition to aptitude are prime determinants of training performance. Also, it should be noted that the prime course content variables, e.g., subject matter difficulty, occupational difficulty, and manpower requirements had negative effects on quality of student performance (-.08, -.13, -.10) respectively.

Assessed quality of student performance had substantial effects on all relevant training outcomes. Quality of performance had sizeable paths against academic counseling (-.20) and nonacademic counseling (-.23). Academic counseling in turn had a positive effect on academic attrition (.22) and on retraining (.34). These results show that counseling moderates the relationship between performance in training and subsequent outcomes such as retraining and elimination. This is one of the more interesting findings generated by the model, that is, the manner in which aptitude, or for that matter any student input variable, influences training outcomes. Aptitude and other student inputs have a strong direct causal impact on the assessed quality of student performance. However, the model indicates that student inputs do not have any direct effect on the more distal training outcomes such as academic attrition, nonacademic attrition, and retraining. Rather, student inputs affect the assessed quality of student performance

which in turn effects either academic counseling or nonacademic counseling. The fundamental importance of academic and nonacademic counseling in determining distal training outcomes is an important finding because it indicates that, regardless of the quality of student inputs or any manipulations made to them, whether an individual is eliminated from training or retrained is ultimately in the hands of course instructors. Given the fundamental importance of instructors' counseling decisions in determining retraining and attrition, it appears that more attention should be given to this process in future research efforts, and that both academic and nonacademic counseling should be considered major training outcomes. That retraining is an antecedent of academic attrition is shown by the path of .25. Thus, high retraining time appears to be an antecedent of attrition. Another interesting finding is that subject matter difficulty had a stronger causal effect (.16) on SIA time than did quality of performance (-.07). This suggests that poor performance might have some effect on SIA time, but that the difficulty of the course material appears to be the prime determinant.

While at first glance, the path coefficients obtained for many of the variables may seem low and appear to be lacking in effective predictive power, the sizes of the Rs generated by the model and the stable predictions generated in the cross validation courses shows that the model provides unusually effective predictions of training outcomes. The relatively low magnitude of the individual path coefficients really in effect confirms the initial premise of the study that technical training is a highly complex process which can only be described by a variety of student input, course content, and training outcome variables. While this is not an especially unusual conclusion, it suggests that optimal description and prediction of the training process can never be attained without employing multivariate techniques capable of taking into account this complex web of interrelationships.

## References


Mumford, M.D., Weeks, J.L., Harding, F.D., & Fleishman, E.A. (1985). An Empirical System for Assessing the Impact of Aptitude Requirement Adjustments on Air Force Initial-Skills Training (AFHRL-TR- ). Brooks AFB, TX: Manpower and personnel Division, Air Force Human Resources Laboratory. (In press)

Weeks, J.L. (1984, November). Occupational Learning Difficulty: A standard for determining the order of aptitude requirement minimums. (AFHRL-SR-84-26). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory. (A417 410).

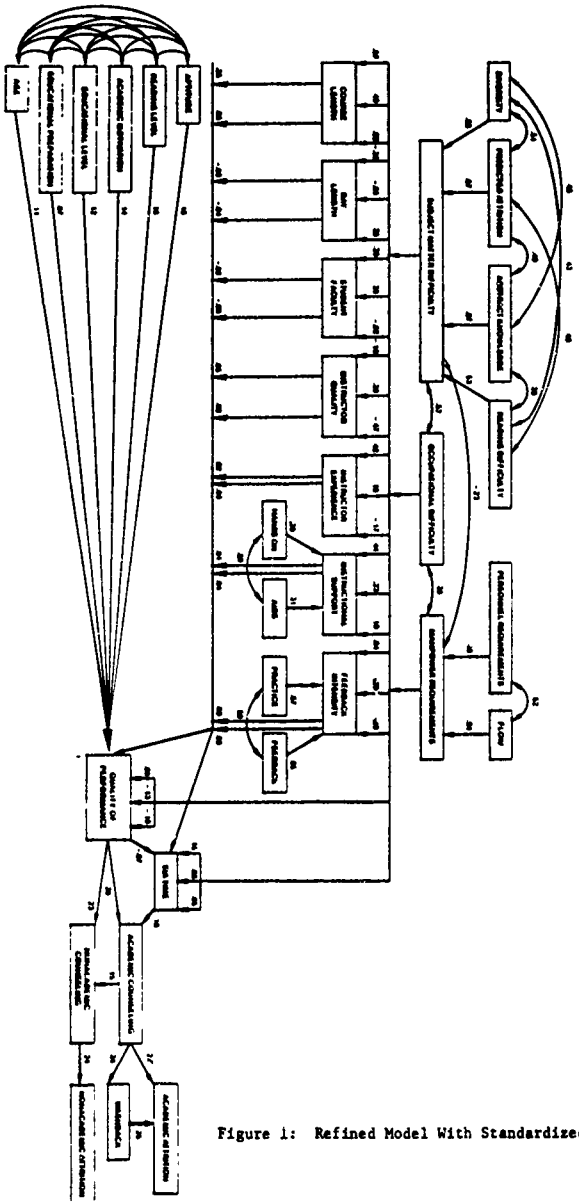



Figure 1: Refined Model With Standardized Path Coefficients.

# AN EXAMINATION OF THE SIGNIFICANCE OF NON-PARTICIPATION IN FIELD STUDIES

Earl H. Potter III
Department of Economics and Management
U. S. Coast Guard Academy

Non-return of survey instruments is a problem that plagues all social scientists. In the analysis of returned surveys researchers generally try to demonstrate that there are no relevant differences between respondants and non-respondants. Most often this determination is made upon the basis of demographic data available to the researchers. Much other data that would be relevant and for which there could be differences is usually not available. This paper examines the personality differences between military participants and non-participants in three voluntary studies which required the keeping of event records over a six-week period.

Over the last several decades researchers have conducted numerous investigations to improve our understanding of the consequences for research validity of non-random attrition in field studies. Such investigations are usually limited by the non-availability of data on non-respondants, but significant differences have been found between subjects and non-respondants who can not be located for telephone interviews (Weaver, Holmes and Glenn, 1975), people who do not return mailed questionnaires (Macek and Miles, 1975) and persons likely to volunteer for experimental studies (Rosenthal and Rosnow, 1969). While the research literature is full of market research studies which examine non-return of questionnaires and the characteristics of volunteers, little work has been done on the characteristics of subjects who drop out of studies or renege on commitments to participate. Silverman (1964) has found that "no-shows" who had agreed to participate in a psychological experiment were higher in self esteem than were participants. Wrightsman (1966) reported that "no-shows" scored lower in social responsibility than did participants. Rosenthal and Rosnow (1969) concluded in a review of volunteerism that "no-shows" are similar to non-volunteers in nature and tend to differ from volunteers in several important respects. Most importantly, they noted that those conditions which support volunteering are exactly those which lead to subjects who had volunteered becoming "no-shows".

This is a significant point in considering research in the military under conditions where subjects are told that participation is voluntary, but hear that announcement from an official spokesperson in the chain of command. As researchers we hope that the suggestion that this is an official duty will encourage participation--dealing with non-participation is troublesome. But... subjects in fact have the right to withdraw. The past research suggests that these conditions may favor the selective withdrawal of certain subjects. This paper examines selective withdrawal in three studies at the U. S. Coast Guard Academy.

## METHOD

SUBJECTS. The U. S. Coast Guard is the smallest of the U. S. Service Academies. Though less well known than West Point or Annapolis, the Coast Guard Academy functions much the same and in fact exchanges students with the other Academies regularly. Cadets enter the Academy at the beginning of the summer preceding their freshman year. The training that takes place during the summer is conducted chiefly by cadets (the Cadre) about to enter their own junior year. This period, called "Swab Summer", is, among other things, intended to be stressful and has changed little since Dornbusch (1955) used the Coast Guard Academy as a model for military socialization. During the remainder of a cadet's freshman year, the cadet is supervised closely and follows a strict academic, athletic and military regimen. While academic grades are given much as they are at any college, cadets are also rated for military career suitability by their peers.

STUDY DESIGN. Study I was planned as a field experiment in which one treatment group would receive a manipulation intended to focus a cadet's attention on his own ability to meet and deal successfully with the challenge of Swab Summer. A second treatment group received an additional manipulation intended to increase support networks within the platoon. It was hypothesized that subjects in the treatment groups would see the Swab Summer experience as being more manageable and less stressful and report themselves to be more ready to meet the challenges of the fall semester than a control group. Furthermore, it was hypothesized that the support network treatment would result in an increase in the quality and number of social supports.

Study II was intended as a replication of Study I. Study III was designed to follow up on the unexpected outcomes of Studies I and II. In Study III the nature of the coping skills treatment was reversed in order to assess the impact of a different instructional set upon perceptions of stress and performance.

PERSONALITY MEASURES. The Edwards Personal Preference Schedule (EPPS) was one of the personality measures. The EPPS is a forced choice inventory with items balanced for social desirability. The fifteen factors include achievement, affiliation, dominance and agression which have fairly clear associations as well as intraception, exhibition and order which generally are less familiar. The second scale was the California Personality Inventory (CPI). The CPI was designed to tap personality constructs with broad social relevance which favor the positive side of human nature. The eighteen subscales fall into four groups--measures of: (a) poise and self-assurance, (b) socialization and responsibility, (c) achievement potential and intellectual efficiency, and (d) interest modes. The third scale was the 16PF which was developed with normal and clinical groups. The sixteen independent factors of the scale are described with bipolar adjective sets in current handbooks, although, early in the development of the 16PF it was noted for the colorful, if obscure, factor labels chosen by the author. These factors include factor A (reserved vs warmhearted), Factor E (humble vs assertive), Factor L (trusting vs suspicious), Factor Q4 (relaxed vs tense). The three scales were administered to the entire entering classes early in their respective "Swab "Summers". In general, their profiles paralleled those of high school or college students with some notable similarities to military reference groups.

PARTICIPATION. The dependent variable is participation defined in Studies I and II as completing a stress diary for three of the six weeks of the study and in Study III as keeping a record of positive events for six of the six weeks of the study.

## STUDY I

SUBJECTS. The subjects of Study I were the 351 cadets sworn into the Class of 1984 during the first week of July, 1980. Of these cadets 311 were men and 40 were women ranging in age from 17 to 21 with the majority (236) being 18 at the time of their entry. The subjects upon entry were randomly assigned by the Coast Guard Academy to one of nine platoons. The one exception to this procedure was that all the cadets who were band members were assigned to the same platoon. Since this fact resulted in the one platoon being systematically different both in composition and with respect to its program for SWAB summer, the band platoon was eliminated from this study as well as Studies II and III. The remaining eight platoons were randomly designated as belonging to Treatment A, Treatment B or Control groups. Platoons 1, 6 and 7 were assigned to Treatment A. During the summer program 12 of the original 117 cadets in Treatment A resigned. Dropping these cadets from the study resulted in Treatment A being comprised of 105 cadets in three separate platoons. Fifteen cadets resigned from the total of 114 cadets who started in Platoons 2, 4 and 8 which were designated as Treatment B leaving 99 cadets in Treatment B. Two platoons, with a total of 75 cadets completed the summer program in the Control group.

RESULTS. In Study I only four subjects in the social support network treatment and no subjects in the non-support treatment completed stress diaries for six weeks. Thirty-nine subjects met the participation criteria of three weeks. Participants were contrasted with non-participants using a T-test on the 16PF, CPI and EPPS. Of all the possible differences, only three emerged. On the 16PF participants appeared to be more conscientious (6.9 vs 6.2, $p<.05$) and more intelligent (6.9 vs 6.3, $p<.05$). On the CPI participants registered a lesser degree of flexibility (7.4 vs 9.0, $p<.01$) which Gough (1975) interprets as meaning that these subjects tend to be industrious, methodical, and responsive to authority.

## STUDY II

Study II was conducted in part as an effort to improve the participation rate. The study was the same in all respects except that the experimenters increased their follow-up efforts by maintaining much closer contact with company officers and cadre than in the previous year. Despite these efforts, support for the study among company officers and cadre varied widely. Some responded to increased contact with more support for this study, some developed a vested interest in their own cadets' participation and supervised more closely than requested or desired and others flatly refused to encourage the participation of their cadet subordinates. What is noteworthy about this variability, and at the same time not unusual, was that command policy was supportive of, although not enthusiastic about, the study. Hence, all the

company officers and members of the cadre were publicly supportive and publicly committed to encouraging but not pressuring cadets to participate. The variability occurred in private statements and in behaviors contrary to public positions. Thus, while the experimenters in their introduction attempted to provide a more supportive and a less pressured introduction to the study, staff follow-up resulted in more pressure which probably contributed to a higher participation rate and, some cadets perceiving that the study was compulsory and not at all voluntary. This feature of Study II probably is common to most "voluntary" participation in officially sanctioned studies, but is clearly undesirable.

SUBJECTS. Study II was conducted during the summer of 1981 utilizing the 396 cadets of the Class of 1985 as subjects. Of these cadets 331 were men and 65 were women. They were divided into nine platoons one of which, the band platoon, was ommitted from the study. Treatment A, the non-meeting group, was comprised of three platoons totalling 133 cadets, 14 of whom resigned before the end of SWAB summer. The non-meeting group included therefore, 119 cadets. Treatment B, the meeting group was comprised of three platoons totalling 141 cadets of whom 15 resigned without completing the summer training program. The meeting group, therefore, included 126 cadets. A control group of two platoons with 85 cadets received no treatment as in Study I. Ages of the cadets were typical of cadet classes ranging from 17 to 21 with 61.6% being 18.

RESULTS. As in Study I, cadets in the meeting group did not appear to keep their committment to meet regularly in small groups. As did cadets in the non-meeting group, they discussed the study with friends of their own choosing. They seemed to form no significant bonds as a result of the randomly generated meeting groups. Given the tremendous pressure on cadet time it turned out to be next to impossible to get cadets to take the time to do something that was not natural when the experimenters could not require or even closely monitor compliance. These further findings support the determination made in Study I to consider participation in both Treatment A and Treatment B as members of one treatment group. In Study II, 91 (out of a possible 245) cadets participated for three or more weeks of the study.

PARTICIPANT VS NON-PARTICIPANT DIFFERENCES. Given the greater participation rate and the perhaps different dynamics of Study II, one would not necessarily expect participants to differ from non-participants in the same fashion as in Study I. Both participants and non-participants rank above the normative means on the CPI's scales for dominance, capacity for status, and social presence. Non-participants, however, are characterized by a greater degree of dominance (3.9 vs 31.6, $p<.001$), greater capacity for status (20.5 vs 19.4, $p<.05$), and greater social presence (39.5 vs 37.7, $p<.05$). Non-participants are, therefore, more independent, self-seeking and less patient and deliberate. On the EPPS, non-participants appear as more dominant (34.0 vs 24.3, $p<.01$) and less verbally aggressive (12.8 vs 14.2, $p<.05$). The overall picture of the non-participants in contrast to the participants is that they are less likely to accept direction from outside, and more likely to invest themselves in demonstrating their independence. As Brehm (1966) would suggest, they are more likely to resist further imposition on their freedom by refusing to participate when the cost to their advancement is low.

## STUDY III

OVERVIEW. Study III was conducted in the same fashion as Studies I and II with the one critical differance being that subjects were instructed to keep daily records of the good things that happened to them instead of stressful events. Since this task was both less demanding and more pleasant, participation for the minimum period was expected to be greater. Also, for the same reasons a greater degree of participation was considered essential. While the recording of one stressful event required writing a paragraph and took some thought, reporting that the cadet received a letter could be done in three words with little thought. By changing the structure of the treatment, therefore, the criterion for participation was also altered. A second consequence of the change in the treatment was that officer and cadet leadership could more easily encourage participation. Nevertheless, variability was still evident ranging from the exertion of too much pressure to less than enthusiastic follow-up.

SUBJECTS. Study III was conducted during the summer of 1982. Owing to a change in the Coast Guard requirements for Academy graduates the Class of 1986 which entered during that summer was considerably smaller than previous classes. A total of 250 cadets including 206 men and 44 women were sworn in during the first week of July. These cadets were randomly assigned to six platoons one of which, the band platoon, was omitted from the study. Two platoons totalling 87 people were designated as controls who received no treatment instructions. One platoon of 41 persons was assigned to Treatment A, the non-meeting group. Of these, 7 persons resigned before the end of the summer leaving 34 in Treatment A. Two platoons totalling 81 persons were assigned to Treatment B, the meeting group. Ages ranged from 17 to 21 with the majority being 18.

RESULTS. Participation in general was much higher as expected. This fact is due most likely to the much greater ease of completing the task. The entry of one event, e.g. "got a letter" would classify a cadet as participant for the week; yet, the real impact of participation might be much less. Therefore, for Study III the participation level was set as six weeks for all subjects. Defined in this manner, 27 of the 107 subjects were classified as participants. This figure places the participation rate of Study III at 25% and between the rate of Study I (19%) and Study II (37%). Once again cadets failed to meet regularly in their small groups. This fact is evident in the finding that social networks did not reflect the impact of assignment to small groups, from discussions with cadets, and from follow-up reports that groups seldom met. The sole virtue in attempting to get groups to meet for each of these studies was comparability. In order to make sense of the negative results associated with random assignment to small groups, however, one only has to recognize that cadets spend 24 hours per day working and living in close association with a peer group. Unlike a large university or even a small liberal arts college, the Coast Guard Academy spends a whole summer before the commencement of the Academic year building peer groups. It is apparent that cadets considered the small group assignments as superfluous. The majority reported a high degree of satisfaction with social supports ($\bar{X}$=3.3 on a 7 point scale with 1= "high satisfaction with support received at the Academy" and 23% reporting some dissatisfaction).

PARTICIPANTS VS NON-PARTICIPANT DIFFERENCES. Participants in Study III were characterized on the 16PF as being more conscientious (6.28 vs 5.24, $p<.01$) and more forthright (4.16 vs 5.12, $p<.05$) than non-participants. On the EPPS participants appeared to be higher in affiliation (18.74 vs 16.32, $p<.05$). This profile, as does the profile of participation in Study I, represents participants as more open and more likely to follow through on commitments.

## DISCUSSION

It is clear that cadets who dropped out of these studies differ from cadets who met the participation criterion in the studies. Participation rates varied with the assigned task and with the energy with which follow-up efforts were made. Participation also varied with personality such that participants like the volunteers studied by Rosenthal and Rosow were more compliant than non-participants. The significance of these findings will vary with the nature of the study. In these studies personality variables on which subjects differ did not correlate with dependent variables. Were these studies of job satisfaction, leadership, motivation or work habits non-participation would not be so benign. In a military setting where subjects might feel compelled to "volunteer" resistance to pressure might well be a systematic bias to research results!

## BIBLIOGRAPHY


BREHM, J.W. - A Theory of Psychological Reactance. New York: Academic Press, 1966.

DORNBUSCH, S.M. - The Military Academy as an Assimilating Institution. Social Forces, 1955, 33, 316-321.

GOUGH, H.G. - Manual for the California Psychological Inventory. Palo Alto, California: Consulting Psychologists Press, 1975.

MACEK, A.J. and G.W. Miles - IQ Score and Mailed Response, Journal of Applied Psychology, 1975, 60, 258-259.

ROSENTHAL, R. and R.L. Rosnow - Artifact in Behavioral Research. New York: Academic Press, 1969.

SILVERMAN, I. - Note on the Relationship of Self-esteem to Subject Self-Selection. Perceptual and Motor Skills, 1964, 19, 769-770.

WEAVER, C.N., S.L. Holmes and N.D. Glenn - Some Characteristics of Inaccessible Respondents in a Telephone Survey, Journal of Applied Psychology, 1975, 60, 260-262.

WRIGHTSMAN, L.S. - Predicting College Students' Participation in Required Psychology Experiments. American Psychologist, 1966, 21, 812-813.

# ADVANCING SURVEY DESIGN THROUGH TECHNOLOGY

Linda M Doherty
Navy Personnel Research and Development Center

*Recent development in computer capabilities and communications are enabling attitudinal surveys to be designed, developed, administered, and analyzed quickly and accurately. The use of computerized survey systems should lead to improvements in survey design, while at the same time providing policy makers with timely and accurate information This paper describes the components one such automated system, and outlines some important survey research issues that may be addressed by using this technology over traditional paper and pencil survey methods*

Computer technology applied to surveys in the past has focused almost exclusively on the statistical processing of data. While this is an important component of the survey research process, other components may now be automated and integrated into a complete system. Recent developments in computer capabilities have enabled the private sector to design, develop, administer and analyze surveys quickly in a broad range of settings that include local and national elections, and consumer product evaluation. Existing computer technology is sufficiently inexpensive so that computers can communicate with remote terminals to efficiently collect and analyze attitudinal information and integrate that information with other data bases Since computer administration of surveys is a relatively new phenomenon, little research has been done comparing this method with other data collection methods Also, survey design research issues more easily studied with this method have not yet been identified.

This paper focuses on (1) describing the components of a computerized survey system that should lead to improved survey design, and (2) defining some of the research issues that may be addressed through such automated data collection methods

## Automated Survey System Components

### Automated Survey Administration and Data Collection

The Navy Personnel Research and Development Center is developing an automated survey system to assess the attitudes of the Navy civilian workforce CENSUS (Computerized Executive Networking Survey System) administers questionnaires and collects information using remote terminals linked by phone lines to host computers The attitudinal information is then integrated with existing personnel data bases and statistically analyzed

Besides the efficiency of administering surveys via computer, the technology has other implications for survey design. First, the "demand" characteristics of the computer terminal are such that it is almost impossible to leave items blank, or provide more answers than requested Second, since branching and individually tailored questions are easily presented, it is possible to study different versions of a survey upon individuals, and vary format characteristics, etc. Third, the survey format may be used to provide feedback to participants to increase the

accuracy and reliability of the information provided, and to present results from previous surveys. Feedback could include specific instructions about how to interact with the computer and complete the questionnaire, and provide positive and negative comments contingent on the quality of the responses

Sampling Strategy

The sampling process may be made more efficient by computerizing the strategy used to extract samples from the population, permitting policy makers to select subsamples of interest, yet ensuring an appropriate mix of participants. The computer program will be structured so that an updated population data base may be maintained and monitored, enabling representative replacement participants to be selected efficiently In addition, results from secondary subgroup analyses on the data base should well represent the population of interest While the sampling strategy itself cannot directly influence the number of survey respondents, the computer terminal used to administer the surveys is interesting and involves the participants in the task. This should increase the likelihood that the identified sample will agree to respond to the survey

Data Analysis

The automated survey system allows results to be computed more quickly and in greater depth than by transferring paper and pencil responses to computer, since the questionnaire is already in the appropriate format for data analysis. By automating the data collection and statistical analysis phases, data entry errors will be avoided The faster the results are reported, the more timely and credible they will be, and, hence, more likely to be accepted and utilized by policy makers Also, answers will be integrated into longitudinal data bases, where secondary analyses may be conducted with relative ease. Branching characteristics of the data collection methods will create hierarchically ordered data Statistical methods need to be developed to analyze these type of data, and further advances in application of statistical and computational methods should occur as a result of the need to analyze longitudinal, hierarchically ordered data

Computerized Authoring System

A computerized authoring system for survey development will be an interactive software program that provides a manager untrained in questionnaire design with a prompting, retrieval, and feedback system to produce reliable and useful surveys. As methods for developing questions become more precise, and as data banks of questions and responses are established, it is possible to develop an authoring system that can prompt managers to develop questionnaires independently and interactively. The authoring system will consist of two components - a menu-driven control system that includes an editor and format specifier, and a data base that includes survey data from previous questionnaire administrations and standardized questionnaires The authoring system will improve questionnire design in general, and will enable the development of questionnaires from an **a priori** knowledge and theory building perspective Thus, the quality of surveys should improve, as well as their applicability.

Feedback Systems to Policy Makers

The expected outcomes of automated surveys are results that are improved in quality and timeliness. By automating reports of the results, policy makers should be able to utilize the findings more quickly. Communication should improve by using the technology of presentation and communication methods These feedback methods would include improved information displays such as the enhanced use of graphics (Fienberg, 1979) Also, a true feedback system will be established where policy makers become involved in developing questionnaires through the computerized authoring system and accessing the data base, interactively, through easy to use menu-driven computer programs. By evaluating the needs and questions of policy makers, more effective feedback that provides input to future survey designs is the final component in the automated survey system

## Research Issues

Response Effects

Most sources of variation in the quality of survey responses may be attributed to what Orne (1969) originally referred to as the "demand characteristics of the situation." Bradburn (1983) stated that " the characteristics of the task are the major source of response effects and are, in general, much larger than effects due to interviewer or respondent characteristics." (p 291). Studies of the task characteristics in traditional paper and pencil questionnaires indicate there are few differences in the variability or quality of responses for different methods of presentation for all types of questions (Sudman and Bradburn, 1974) Use of the computer provides the survey developer with the ability to study these effects systematically

Sensitive questions. Collecting responses to sensitive questions may be one area where method of presentation is important. In this area, computerized administration can make a considerable contribution by improving the number and quality of responses obtained. Respondents may perceive that completing a survey on an automated system is more anonymous than other methods, and, hence, they may be more willing to disclose personal and sensitive information One method previously tested to increase anonymity, and believed to increase honesty, is the randomized response technique (Warner, 1965), which requires the respondent to use randomization to determine whether a sensitive question or an innocuous one is to be answered Simple probability calculations then give an estimate of the number in the sample who agreed with the sensitive question, without knowledge of which respondents did so. This technique was found to increase reports of socially undesirable behavior and, perhaps, honesty Some evidence indicates that respondents may answer electronic surveys with more extreme, thus, less socially desirable answers than they answer paper and pencil surveys (Kiesler and Sproull, 1985; Kiesler, et al , 1985).

Questionnaire format. Considerable work has been done to study the effects of task variables such as questionnaire format, wording, and length, with conclusions that these effects are too complex to make definitive statements (Sudman and Bradburn, 1980). Schuman and Presser (1981) found little relation between measures of attitude strength and

propensity to be influenced by question form. However, they did find that the context in which a question is asked - the ordering of questions, the inclusion of other questions and the arrangement of the questionnaire can produce response effects. The implications for studying response effects using an automated system is that it provides the flexibility to vary the form of the questionnaire, its length, wording, level of branching or contingency questions within a single questionnaire or among questionnaires. The Census bureau is testing one automated system, CATI (Computer Assisted Telephone Interviewing) system, where interviewers interact with a computer to make contact with respondents by telephone, and ask them branching, in-depth questions that are recorded immediately into the computer (Nicholls, 1983) With the flexibility in question presentation permitted by automated survey systems and sophisticated sampling plans, it is possible to design experiments that will elicit the specific characteristics of questionnaires to which respondents are most sensitive, and how those characteristics affect their answers

Response sets and memory Characteristics of respondents, or response sets, may influence answers to questions in a way that has nothing to do with the question being asked Some research indicates (Bradburn et al , 1979) that response sets may not be respondents attempting to present a socially desirable image, but may reflect more stable personality characteristics or different life experiences. By exploiting the branching capabilities of the computer, it may be possible to discriminate between response sets and real differences in attitudes

The effects of memory operate on responses because respondents forget what they have previously answered, or they telescope time and report events happening more recently than they actually occurred With a longitudinal panel design for data collection and an automated survey system, the effects of memory may be tested. Previous responses and information about respondents may be resident in the survey software programs and used as memory aids when asking follow-on information Also, updated information may be simply compared to previous responses.

Measurement Issues

One major issue in measurement theory focuses on the number of multiple items or indices required to adequately measure a psychological construct versus the number of questions that are reasonable to include in a survey of limited time (Anderson, et al , 1983) Advantages of a branching, computerized system are that the questions and response scales could be tailored to individuals, probing one issue in depth if appropriate Scales could be utilized that would be sensitive to capture differences in individual ' attitudes and then using the flexibility of the system, adapt questions and scales to more precisely measure individuals attitudes In this way a number of items would be used to assess a construct, but the number and type of questions and scales could vary

Scaling techniques Another implication for the data collection technology is the ability to more easily use complex scaling techniques to assess attitudes While many of these techniques provide researchers with additional information, some of the shortcomings are that they are too cumbersome and time consuming (Anderson et al , 1983) For example, application of nonmetric multidimensional scaling techniques are

advantageous in the level of measurement required (ordinal) and output provided (dimensional plots), but the paired comparison method of data collection takes too much time and is uninteresting to the participant With an interactive computer system, the data collection time could be reduced as additional data is requested only if other information is needed to construct an adequate perceptual map of the items

Measurement models. A second major area in measurement that will be influenced by the applicability of technology to collect and analyze data is the use of more sophisticated measurement models. While methods to combine items to uncover a psychological construct are not new, development of these in the field of attitude measurement is. One technique, latent structure analysis is being developed in attitude measurement, and may provide a way to measure attitude types (Reiser, 1981) The use of the broad set of multivariate techniques (models) should also increase The appropriate use of factor analysis is one such important technique in describing attitudes.

Social Indicators

Automated surveys provide the ability to collect vast amounts of data rapidly and maintain large longitudinal data bases The growth of these data bases should stimulate the growth of statistical and computation methods to analyze these specialized hierarchically ordered data bases, created because of the branching characteristics of automated data collection In particular, the type and quality of the data should stimulate innovative methods of analyses by disciplines other than social science Causal models that combine attitudinal and economic variables to describe behavior for individual subgroups of the samples should be one area that is stimulated by the accumulation of large data bases

The second outcome of developing longitudinal data bases may be the improvement of experimental designs to evaluate the effects of policies By integrating innovative experimental designs with survey research methodology, social indicators of the behavior of the military and civilian workforce may be developed Collecting longitudinal, panel data in an efficient manner permits the establishment of baseline data and the ability to measure the long term effects of policy changes Improvements to the design of experiments testing policy changes and the associated complicated sampling plans for collecting survey data may be the result of the implementation of automated surveys

## Conclusions

Automated survey systems will have a positive impact on survey design, and its methodological problems Through computer technology, surveys will be able to be developed, administered, and analyzed more efficiently, providing high quality results to policy makers. Automated surveys can also substantially increase the amount of available data for subsequent analyses One of the challenges will be to develop the analytic and statistical tools to evaluate the data, and to develop behavioral models using attitudinal and other, traditionally collected demographic variables What is needed is the transfer of methodologies across academic disciplines to actively address the problems encountered in implementating automated survey systems

## References


Anderson, A B , Basilevsky, A., & Hum, D. P J. Measurement. theory and techniques In P. H Rossi, et al (Eds ), Handbook of survey research New York Academic Press, 1983.

Bradburn, N M. Response effects. In P H Rossi, et al. (Eds.), Handbook of survey research New York Academic Press, 1983

Bradburn, N M., Sudman, S , et al Improving interview method and questionnaire design response effects in threatening questions in survey research San Francisco Jossey-Bass, 1979

Fienberg, S E Graphical methods in statistics The American Statistician, 1979, 33, 165-178

Kiesler, S & Sproull, L. S Response effects in the electronic survey (Unpublished paper), 1985

Kiesler, et al Affect in computer-mediated communication. an experiment in synchronous terminal-to-terminal discussion Human-Computer Interaction, 1985, 1, 77-103

Nicholls, W L , II CATI research and development at the Census bureau Sociological Methods & Research, 1983, 12, 191-197

Orne, M T Demand characteristics and the concept of quasi controls In R Rosenthal and R. L Rosnow (eds ), Artifact in behavioral research New York Academic Press, 1969

Reiser, M. Latent trait modelling of attitude items In G W Bohrnstedt and E. F Borgatta (Eds ), Social Measurement Current Issues Beverly Hills Sage Publications, 1981

Shuman, S & Presser, S Questions and answers in attitude surveys experiments on question form, wording and context New York Academic Press, 1981

Sudman, S. and Bradburn, N M. Response Effects in Surveys. A Review and Synthesis Hawthorne, NY Aldine, 1974

Warner, S L Randomized response a survey technique for eliminating evasive answer bias Journal of the American Statistical Society, 1965, 60, 63-69

# IMPLEMENTATION OF A CIVILIAN AUTOMATED SURVEY SYSTEM

FRANK L. RIPKIN
STEVE CECIL

OFFICE OF THE CHIEF OF NAVAL OPERATIONS
CIVILIAN PERSONNEL POLICY DIVISION
ARLINGTON ANNEX
WASHINGTON, D.C. 20350

CENSUS -- The Computerized Executive Networking Survey System is being developed by the Navy Personnel Research and Development Center to address both an immediate operational need as well as to serve as a basis for new technology for automated surveys. This paper will be concerned with the concepts underlying the development of CENSUS and its potential application as a tool for policy development for the Navy.

The Department of the Navy has approximately 350 thousand civil service employees located worldwide, working in some 450 different occupations. With a work force of this magnitude, civilian policy makers require a considerable amount of information and data to formulate policies which range from procedures and regulations concerning the hiring of employees to policies governing their retirement system. The problem facing the policy maker is deriving information of both a factual (work force demographics, trends in retention, etc.) and an attitudinal nature in order to make possible proactive rather than reactive policy decisions. Compounding the problem is the requirement that policy be developed in a timely manner; we often have very little time to assess and respond to policy proposals presented to the Navy by such external agencies as the Office of Management and Budget or the Office of Personnel Management.

Using this as a background we normally have two alternatives for most policy decisions which impact on employees, their opinions and attitudes:

(1) The Quick Fix - - Providing for a gut or knee jerk reaction pertaining to employee attitudes coupled with an estimate of the demographic impact on the work force.

(2) The Steady Course - - Through this method we take the time, six to eighteen months normally, to develop quality survey instruments which address the issues in a straight-forward and accurate manner, identify a proper sampling that reflects the affected population and, conduct a thorough survey and analysis of the work force attitudes toward a given policy or policy proposal. We then couple this assessment of attitude and opinion with information detailing

Copy available to DTIC does not permit fully legible reproduction

work force demographic changes actually observed or modelled.

On reflection these two possible plans both offer a viable means for development of policy. The Quick Fix while fast, may result in a less than desirable outcome because a thorough analysis was not completed. The Steady Course on the other hand, offers a sound basis for policy development decisions, yet by the time the decsison is formulated the issue has often been settled by an outside party (more than likely using the Quick Fix approach)

Thus the genesis of CENSUS - - Utilize advances in the technology of survey development, sampling techniques, telecommunications and computer analysis to reduce the time necessary to ascertain employee attitudes without sacrificing the accuracy of the Steady Course and while still gaining much of the speed of the Quick Fix. Joining these advances with those already being made by Navy in the area of Personnel data collection and analysis through the Navy Civilian Personnel Data System (NCPDS) and Human Resource Modelling, Navy Personnel policy makers will have a strong developmental tool.

The development of CENSUS has three objectives:

(1) General Objective: Develop a quick reaction automated survey system in support of Personnel policy development.

(2) Technological Objective: Develop sophisticated sampling plans.

Investigate the relationship between response effects and questionnaire/survey design

Explore alternative survey presentations.

Develop causal models for application in policy analysis

Develop a computer aided authoring system to assist managers in the development of survey instruments.

(3) Operational Objective: Assess the feasibility of collecting large samples of variables through an automated survey network.

Integration of up-to-date information on the attitudes and opinions of Navy employees with the information being obtained in advanced Navy data and modelling systems.



Copy available to DTIC does not permit fully legible reproduction

Development of longitudinal data bases to assess the long term impact of policies and the changes in employee attitudes over time.

## CENSUS DEVELOPMENT

The development of the Computerized Executive Networking Survey System began in 1983 when the Office of the Chief of Naval Operations, Civilan Personnel Policy Division tasked the Navy Personnel Research and Development Center to develop a quick reaction survey system along the lines of the Harris or Gallup polls. The lead researcher on the CENSUS project is DP Linda Doherty (please see a related paper elsewhere in this document). The project is to be developed in five stages:

(1) Initial systems test

(2) West Coast pilot implementation

(3) East Coast pilot implementation

(4) Nationwide expansion

(5) Implementation and ongoing research

The initial system and pilot tests for CENSUS were conducted in November 1984 and April 1985, in San Diego, California and in Washington, D.C., in August 1985. Approximately 1000 Navy employees have participated representing in excess of 30 Navy activities with usable results applicable to both the San Diego and Washington areas.

## SYSTEM DESIGN

The system components consisted of IBM PC AT & XT's with specially modified software, allowing simultaneous access by participants and presentation of the survey. The surveys were presented to the participants by means of a Northern Telecom DisplayPhone which served as both the telecommunications and terminal device. The micro computers served as host site computers during all tests. A mini computer was utilized as a secondary host during the initial system test. The West and East Coast test sites were established in order to test the reliability, speed, and utility of the host equipment, as well as the logistics of establishing regional survey centers. In order to make the survey test as real as possible actual surveys were developed by the staffs of NPRDC and this Office.

A unique aspect of the CENSUS surveys is that by presenting the survey via the computer and terminal we are able to eliminate two of the major drawbacks of a paper and pencil survey. One being the collection of background or demographic data on participants at the time of survey administration. All demographic data on


Copy available to DTIC does not permit fully legible reproduction

sampled employees is maintained at the host site with the participant's responses linked by use of their Social Security Account Number which is obtained during the sampling process and verified during survey administration. This speeds the survey administration as well as allows analysis of responses keyed to a large amount of demographic data, some of which may be added after the survey has been administered. This is something which is not normally possible in a paper and pencil survey, and can be a strong tool for analysis when twists and turns in policy development require additional or unanticipated information be supplied to the functional manager.

The second benefit of the computer aided survey process is the application of branching in the survey design. By making the process of branching transparent to the participant CENSUS has eliminated the standard and frustrating, "if your answer is yes please proceed to question number 34." We anticipate that the use of branching will make sophisticated survey design a routine matter in CENSUS and also allow immediate feedback of information to participants possible during survey administration

## ASSESSMENT

The issues under review during the system and pilot tests were the equipment and how well it would function, the software operations which allow multiple access to the host computer and stores the survey files, the terminal operations and ease of use, the logistics of establishing the survey network, and the attitudes of the participants. Additional concerns involved the actual survey development process and the utility of survey results for policy development.

These tests have proven to be very satisfactory. The results showed the reliability of the IBM PC AT/XT as a host computer. It was determined, however, that even though the PC proved to be reliable, a back-up computer would be required in future sessions. The mini computer functioned adequately yet due to its age and mechanical condition did not prove to be as reliable as the micro. It was decided that further tests of a mini computer as a host would be delayed until a later date. The software functioned as planned, though results of the later survey pilots have shown improvements in file handling are required before full implementation would be possible.

The main concern from these system and pilot tests was the extensive logistics and up front work required to establish a survey site. The process is labor intensive, requiring personal contacts with activities and participants to be conducted before, during and after the survey. We anticipate the design of the implementation plan can take this into consideration and thus reduce the manpower and logistics involved. The negative aspects of manpower and logistics requirements will continue to be addressed as future pilots are conducted.



Copy available to DTIC does not permit fully legible reproduction

The most encouraging outcome from our prospective was the ease of use for the participants and the utility to policy development of the total CENSUS process. The participants found the survey easy to take, the branching of questions, which is transparent to the user, added significantly to the user acceptance. The general response of the user was one of enthusiasm.

The functional managers have found that the speed of development and the amount of specific data made available to them far exceeds previous survey processes. Following the analyses of the test surveys, data were presented to functional managers for evaluation as a tool for policy development. The results of the survey, while valid only for the area in which the test took place, contained enough specificity and information to rival previous paper and pencil surveys. As stated earlier previous surveys of this nature have taken 6 to 18 months to produce and analyze. Each of these tests took less than 3 months to develop. conduct and analyze. We anticipate that after full implementation of the system, CENSUS will be able to reduce the time for survey development and administration even further.

The functional managers also found the use of electronic mail between NPRDC and this Office during the second West Coast and the East Coast pilots to be extremly beneficial in the development of the survey instrument

From analyses conducted by NPRDC and this Office, the initial system and pilot tests were determined to be a complete success. Further development of the CENSUS program will be continued concentrating now on system improvements, Navy-wide implementation planning, and continuing research in the areas of survey development and analysis

Questions concerning the development and application of CENSUS may be directed to Mr Steve Cecil, CENSUS Project Director. Mr. Cecil may be reached at the above address or by calling (202) 694-5762 or autovan 224-5762.

Copy available to DTIC does not permit fully legible reproduction

# The Utility of Educational and Biographical Information for Predicting Military Attrition

Barry J. Riegelhaupt
Thomas P. Bonczar
Human Resources Research Organization

> An examination of discriminating item responses can tell a great deal about what kinds of employees remain on a job and what kinds do not, what kinds sell much insurance and what kinds sell little, or what kinds are promoted slowly and what kinds are promoted rapidly. Insights obtained in this fashion may serve anyone from the initial interviewer to the manager who formulates personnel policy. (Owens, 1976, p. 612)

Compelling evidence demonstrates that, when appropriate procedures are followed, the accuracy of biographical data as predictors of future work behavior may be superior to any known alternative (Cascio, 1982). Biographical inventories have been developed for a diversity of occupations, from unskilled and clerical workers to research scientists and Army officers. By far, the most commonly used criterion for developing, validating and cross-validating these inventories has been attrition.

The role that biographical information can play in predicting attrition, as well as job performance, both costly problems for the Military Services, has been recognized by the Department of Defense. The Office of the Assistant Secretary of Defense (Force Management and Personnel) as part of a study to validate and improve current education and moral enlistment standards, awarded a contract to the Human Resources Research Organization (HumRRO) to develop the Educational and Biographical Information Survey (EBIS). Analyses relating EBIS responses to military performance are continuing under a contract sponsored by the Office of Naval Research. The long-range objective of this research is to help the Services identify potentially successful nongraduates and refine selection among high school graduates.

## EBIS Composition

EBIS items can be classified into nine constructs, theoretically related to military performance. Demographic items address type of area in which one grew up, family income, and parents' education. Education achievement variables deal with years of education, school grades, and type of education credential. School behavior/attitude items concern school activities, disciplinary incidents, reasons for thinking about quitting school, and so

This research was performed under contract to the Office of the Assistant Secretary of Defense (Force Management and Personnel) and the Office of Naval Research. The views and opinions expressed in this paper are those of the authors and should not be taken as official DoD policy.

on. Family relations items deal with attitude toward parental discipline, presence of each parent in the home, and parent stability (e.g., incidence of alcoholism, parent's arrest, etc.). Length of longest full- and part-time job and all the reasons the respondent ever quit a job are included in the work history category. Status variables are subject characteristics such as sex, year of birth, and kind of high school attended. Arrest-related items deal with traffic, misdemeanor, and felony arrests and convictions. Alcohol/drug use items deal with both frequency and age of onset for use of these substances. Minor Misbehaviors deal with youthful behaviors such as smoking or running away from home, which are not illegal, but might be construed as midly negative.

**EBIS Administration**

The EBIS was administered to approximately 75,000 military applicants and recruits in the Spring of 1983. Applicants took the EBIS along with the Armed Services Vocational Aptitude Battery (ASVAB) at a Military Entrance Processing Station (MEPS) or associated test site. Recruits were given the EBIS during in-processing at their Recruit Training Center (RTC). Individuals who took the EBIS are being tracked through their first terms of service, with an emphasis on identifying military suitability predictors within education groups. Based on simple correlations between EBIS items and attrition status, previous work indentified the 24 best EBIS items for predicting six-month attrition among high school graduates and the best 24 items for nongraduates (Means & Laurence, 1984). Building on that earlier effort, the present work describes the statistical weighting of EBIS items based on 12-month attrition status to produce a total EBIS score, and scores for each of the nine constructs. The utility of the EBIS for predicting 12-month attrition is examined. Of the approximately 75,000 EBIS respondents, only the 3,092 male nongraduates who took the EBIS as recruits were examined in the present effort.

**Research Design**

In addition to nonattrition (stayers) and attrition (leavers) groups for use in weighting biographical items to predict the criterion of 12-month attrition status, nonattrition and attrition hold-out groups from the same population are needed for evaluation of items and weights. Thus, in addition to dividing the sample into stayers and leavers, these two groups were further subdivided into weighting/selection and evaluation (hold-out) groups. For the sample of 3,092, the design is shown below.

| | Criterion Group | | |
|---|---|---|---|
| | Stayers | Leavers | Total |
| Weighting/Selection | 1262 | 592 | 1854 |
| Evaluation | 842 | 396 | 1238 |
| Total | 2104 | 988 | 3092 |

**Figure 1. Allocation of Subjects for Weighting and Evaluation**

## Item Weighting and Selection

While a number of different techniques are available for statistically weighting items (c.f. Owens, 1976), perhaps the simplest technique is the Horizontal Percent Method. Cascio (1982) notes that as long as samples are relatively large, this method yields results comparable to any of the more sophisticated procedures. An example of how items were weighted is shown in Figure 2. The frequency of respondents falling into each criterion group (stayers vs. leavers), response category combination was computed. Then, after adjusting for differences in sample sizes in the two groups, for each response category, (working horizontally) the percentage of cases in the nonattrition (i.e., stayers) group was computed. This percentage became the response category weight.

| EBIS Item | Freq. of Stayers | Freq. of Leavers | Total Number | Percent Stayers |
|---|---|---|---|---|
| **13. Were you ever expelled from school?** | | | | |
| Yes | 140 | 107 | 247 | .38 |
| No | 1086 | 471 | 1557 | .52 |
| | 1226 | 578 | 1804 | |
| **28. Frequency of Physical Fights?** | | | | |
| Never | 600 | 272 | 872 | .52 |
| Once or Twice | 187 | 105 | 292 | .47 |
| Occasionally | 30 | 23 | 53 | .40 |
| Fairly Often | 12 | 15 | 27 | .29 |
| | 829 | 415 | 1244 | |

**Figure 2. Horizontal Percent Method of Weighting EBIS Items**

This approach to weighting has a number of advantages. It can be applied, without modification, to all items regardless of the number of response options that exist for an item. Each option receives a weight. This is of particular importance for the EBIS where the number of response options range from two to thirteen across the 119 items. Since this approach does not require that all items have an equal number of options, it is not necessary to artifically dichotomize, for instance, all items for the sake of consistency in format. Finally, the computation of weights yields information relevant to the determination of item quality (i.e., response frequencies).

Thus, after the weighting algorithm had been applied to the weighting sample, response option weights and frequencies were examined for purposes of eliminating poor items. A poor item was defined as one that either did not discriminate stayers from leavers (i.e., response choices had weights around .50) or had little variance (i.e., greater than 85 percent of the respondents

selected one response option). The distribution of items retained for further analyses is shown in Table 1. As can be seen, the Demographic, Status, and Arrest Related scales were deleted.

Table 1

Distribution of Item Types Represented on the EBIS and Retained for Validation Against Attrition

| EBIS Scale | # Items | # Retained | % Retained |
|---|---|---|---|
| Demographic | 5 | 0 | 0 |
| Education Achievement | 14 | 7 | 50 |
| School Behavior/Attitudes | 32 | 16 | 50 |
| Family Relations | 13 | 2 | 15 |
| Work History | 17 | 5 | 29 |
| Status Variables | 4 | 0 | 0 |
| Arrest Related | 16 | 0 | 0 |
| Alcohol/Drug Use | 12 | 4 | 33 |
| Minor Misbehaviors | 6 | 4 | 67 |
| Total | 119 | 38 | 32 |

## EBIS Scoring

Having selected 38 discriminating items, the next step was to compute an EBIS score for each individual in the item weighting/selection sample. The score an individual received for an item was the weight associated with the response option selected. For example, referring to Figure 2, if the respondent was expelled from school, a .38 would be the score for that item. If the individual fairly often engaged in physical fights, a .29 would be that individual's score for that item. A total scale score was computed for each individual by summing the response weights across the 38 items. In a similar fashion, individual scale scores were also computed.

## Comparison of Stayers and Leavers

Following the computation of EBIS scores for individuals in the item weighting sample, stayers and leavers were compared on the total EBIS scale as well as the six factors that were retained. The results are presented in Table 2.

As can be seen, stayers and leavers had significantly different scores on the total EBIS scale, as well as five of the six factors. Only for Minor Misbehaviors was the mean difference not statistically significant, although it was in the expected direction. These results show that those who remain in the Military Services have different pre-enlistment experiences than those who leave, thus demonstrating the usefulness of a biographical inventory.

Table 2

Comparison of Stayers and Leavers in the Item Weighting/Selection Sample

| | Stayers[a] | | Leavers[b] | | |
|---|---|---|---|---|---|
| EBIS Scale | Mean | SD | Mean | SD | t-value |
| Total | 18.408 | 1.058 | 18.136 | 1.305 | 4.78* |
| Education Achievement | 3.409 | .422 | 3.334 | .454 | 3.29* |
| School Behavior/Attitudes | 7.897 | .353 | 7.812 | .401 | 4.65* |
| Family Relations | .979 | .114 | .965 | .127 | 2.49* |
| Work History | 2.513 | .055 | 2.502 | .056 | 3.97* |
| Alcohol/Drug Use | 1.431 | .242 | 1.398 | .262 | 2.56* |
| Minor Misbehaviors | 1.784 | .309 | 1.763 | .329 | 1.34 |

[a] n = 1262
[b] n = 592
*p <.05.

These results should be interpreted with caution, however. Since these comparisons were performed on the same individuals that were used to weight and select the items, observed differences between stayers and leavers may not reflect true differences, but rather chance fluctuations. The application of the 38 items with their corresponding weights to an independent sample might yield substantially different results from those obtained using the item weighting/selection sample. Thus, the next step was to evaluate the utility of the EBIS in the hold-out sample.

## Comparison of Stayers and Leavers in Hold-Out Sample

As in the previous sample, item weights were applied to individuals in the hold-out sample and EBIS scores were derived. Once again, stayers and leavers were compared on a total EBIS score and on the six factor scores. The results are presented in Table 3. As can be seen, the EBIS total score, as well as the Education Achievement, School Behavior/Attitudes, and

Table 3

Comparison of Stayers and Leavers in the Evaluation (Hold-out) Sample

| | Stayers[a] | | Leavers[b] | | |
|---|---|---|---|---|---|
| EBIS Scale | Mean | SD | Mean | SD | t-value |
| Total | 18.384 | 1.029 | 18.195 | 1.266 | 2.79* |
| Education Achievement | 3.410 | .370 | 3.351 | .474 | 3.36* |
| School Behavior/Attitudes | 7.890 | .328 | 7.821 | .395 | 3.22* |
| Family Relations | .976 | .121 | .962 | .133 | 1.32 |
| Work History | 2.511 | .056 | 2.509 | .057 | <1 |
| Alcohol/Drug Use | 1.440 | .211 | 1.408 | .272 | 2.22* |
| Minor Misbehaviors | 1.768 | .330 | 1.786 | .308 | <1 |

[a] n = 342
[b] n = 396
*p <.05.

Alcohol/Drug Use scales yielded significantly different mean scores for stayers and leavers. The significant differences that were found on the Family Relations and Work History scales in the weighting sample, were not found in the hold-out sample. As in the weighting sample, scores on the Minor Misbehaviors factor did not differ when comparing stayers and leavers in the hold-out sample.

## Conclusions

The data in this report demonstrate the utility of biodata for predicting attrition for nongraduate male recruits. Specifically, items dealing with School Behavior and Attitudes, Education Achievement, and Alcohol and Drug Use all showed significant relationships with attrition. While statistically significant, mean scale differences between stayers and leavers were relatively small. The conclusion, however, that these significant findings are of little practical value is unwarranted. As scaled, item weights can range from 0.0 to 1.0. A value of .50 reflects a nondiscriminating item response. As values move closer to 0.0 or 1.0, the response becomes a better discriminator. It is rare, however, to find response choices with sufficient numbers of respondents with weights approaching 0.0 or 1.0. If such items existed, they would likely be more predictive of attrition than most entire biodata scales. No such items exist on the EBIS. Typically, response choices within items differed by .10 or less. Thus, summing across items for stayers and leavers, produces relatively small mean differences. However, very small standard deviations also resulted, which contributed to the significant relationships found in the hold-out sample. It is the finding of significance in a hold-out sample that is the true test of an instrument's usefulness. The EBIS passed this test. Thus, biodata items, such as those found on the EBIS, may allow the Services to enlist nongraduates who have a better chance of fulfilling their enlistment obligation.

## Next Steps

Future work will look at military applicants' EBIS scores to determine and adjust for the restriction of range that might exist in the sample examined in this paper. Additionally, EBIS cut off scores will be set and the amount of improvement in predicting attrition attributable to biodata will be examined.

## References

Cascio, W. F. (1982). Applied psychology in personnel management (2nd ed.). Reston, VA.

Means, B. & Laurence, J. H. (1984, November). Improving the prediction of military suitability through educational and biographical information. In Proceedings of the 25th Annual Conference of the Military Testing Association, 203-208.

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

# THREE VARIABLES THAT MAY INFLUENCE THE VALIDITY OF BIODATA[1]

Clinton B. Walker
U.S. Army Research Institute for the Behavioral and Social Sciences

This research examines the effect on predictive validity of traditional procedures for developing and implementing suitability screens in the military. For this paper, suitability screens in the form of background questionnaires, or biodata, will be considered. Typically, predictor tryouts have been run on new recruits whose subsequent performance has been tracked for the first six months of service (Atwater & Abrahams, 1983; Walker, 1985). Item selection and keying have then been based on the observed relation between predictor data and the criterion of successful service (versus discharge for bad causes). In the case of the U.S. Army's Military Applicant Profile (MAP), the instruments and keys have been implemented no less than two and a half years after the tryouts.

There is reason to suspect that three aspects of this traditional sequence - viz., testing recruits rather than applicants, tracking the cases for only six months, and implementing long after pilot testing - adversely affect operational validities. Since recruits and applicants are likely to differ in their desire to make themselves look good on self-report measures, applicants could be expected to try more than recruits to earn high scores. As a result, scoring keys that are developed on data from recruits may be less valid for scoring responses of applicants. In support of this hypothesis, Means and Heisey (1985) have found more self-serving responses in data from applicants than from recruits.

The hypothesis that using only a six-month tenure for tracking success/attrition lowers validities is based on the following two premises. First, more than half of attritions occur after the initial six months (Goodstadt & Yedlin, 1980; Hicks, 1981; Walker, 1985). Second, attrition during the first six months may not occur for the same reasons as later attrition. In the first six months recruits make their initial adjustment to military life while undergoing entry-level training; after that they are serving with operational units. Unfortunately, the archival codes for types of attrition are too cryptic (e.g., "Trainee Discharge Program," "Unsuitable Unknown," "In Lieu of Court Martial") to indicate whether earlier and later attrition are qualitatively different phenomena. But if they are, then using a longer than traditional criterion period for developing scoring keys might produce different keys.

A long lag time before implementing scoring keys is suspect because characteristics of the applicant pool change over time. Once the predictor data are collected for developing a biodata instrument, they may obsolesce as the criterion ripens. If the nature of the applicant pool changes much, then a scoring key may lose validity before it is ever used for screening, and continue to lose validity after implementation.

[1]Thanks go to Elizabeth P. Smith for advice on programming and to Winnie Young for creating the dataset on applicants in FY 81/82 from the Project A Longitudinal Research Database. The views in this paper are those of the author and do not necessarily reflect the official positions or policies of the U.S. Army Research Institute or Department of the Army.

The present research uses data from the MAP to test the effects of each of these variables. Responses to a common set of items by contemporaneous applicants and recruits are compared to test the effect of examinees' status. Then, various statistics are examined over the course of years to test the effect of time lapse on the keys' validity. Finally, the effect of duration of the criterion period is tested by comparing the predictor responses of examinees who were discharged within and beyond the first six months of service. Each of those issues is treated in turn below in a separate section.

## Applicants Versus Recruits

### Method

To keep from confounding the effects of examinees' status with those of date of testing (i.e., temporal drift), it is necessary to compare contemporaneous applicants and recruits. Two such comparisons are available in the MAP data. First, MAP scores of 2,374 non-graduate applicants during FY 82 were compared with those of 1,286 non-graduate recruits who were tested in February-June, 1982. These recruits were the non-graduate subset of a sample of 9,603 cases on whom new instruments were being developed (Erwin, 1985). Out of the 240 items in that research, 38 were chosen for use here according to these two criteria: they had to be on the operational form of MAP, so the applicants would have taken them, and they must have shown validity for non-graduates in the developmental research. These 38 were the universe of items that met both criteria. The key for scoring had been developed on all 9,603 cases. Here the comparison was a $t$-test on the total score, 0 to 71 being the possible range.

Data for the second comparison overlap in part with the previous ones. In the developmental work of 1982, the item pool was administered to a sample of applicants at 39 Military Entrance Processing Stations (MEPS) nationwide and to recruits at all seven Army Reception Stations. Out of those groups, a respective 949 and 9,603 examinees of all levels of education, age, and gender were retained for analysis. Retention was based solely on the availability of individuals' criterion data in central personnel files. In the applicant sample, 267 cases retook the instrument later as members of the recruit sample. Presumably the presence of those cases reduces the between-group differences, thus biasing any test against finding differences.

The vehicle for this second comparison was two 101-item forms of MAP which were developed on the 9,603 recruits. These forms each had 78 unique items and 23 items in common, yielding possible scores of 1 to 188 on one and 0 to 194 on the other. Mean MAP scores and validities against the six-month tenure criterion were compared in the applicants and recruits.

### Results

Descriptive statistics for the non-graduate applicants in FY 82 and the recruits in the 1982 development sample are included in Table 1. The observed difference in means of 10.9 points is significant ($t = 52.9$; $p < .001$) and the effect is strong (omega square = .43). Data for applicants and recruits in the 1982 developmental project are summarized in Table 2. Applicants' total scores were higher by 2.81 points on Form 1 and 2.1 points on Form 2. These differences gave $t$'s of 5.62 and 3.96 ($p < .01$ in each case). However, here the strength of effect was less than 1% for each form. For both forms, the

bserved validities were higher for recruits than for applicants. The difference between correlation coefficients for independent groups (Guilford & Fruchter, 1973) was computed on the validities for each form. The observed z's of 2.51 and 1.59 had one-tailed probabilities of .006 and .056, respectively.

Table 1
Descriptive statistics for four samples of non-graduates[a]

| Logical role | Date of predictor data | Status: Applic/ Recruit | n | r | Mean out of 0-71 | SD | % finish 6 mo |
|---|---|---|---|---|---|---|---|
| Develop key | 1-6/82 | Recr | 1,286 | .18 | 33.3 | 6.2 | 79 |
| X-validation & 0 yr drift | 10/81-9/82 | Appl | 2,374 | .02 | 44.2 | 5.5 | 86 |
| 1 yr drift | 10/80-9/81 | Appl | 3,567 | .07 | 44.2 | 5.2 | 86 |
| 2 yr drift | 7/79-6/80 | Appl | 14,771 | .01 | 28.3 | 5.6 | 86 |

[a]The "instrument" for these data was 38 items from MAP 4B which were keyed on the total 1982 development sample of 9,603 cases and were also valid for its non-graduate subsample.

Table 2
Descriptive data for applicants and recruits in 1982 development sample

| | | Form 1 | | | Form 2 | | |
|---|---|---|---|---|---|---|---|
| Status | n | Mean | SD | r | Mean | SD | r |
| Applicants | 949[b] | 125.56 | 14.49 | .24 | 123.72 | 15.54 | .27 |
| Recruits | 9,603[b] | 122.75 | 16.64 | .32 | 121.62 | 17.27 | .32 |

[b]Both samples include 267 cases who took the instrument a second time as members of the recruit sample.

Discussion

Both sets of comparisons support the hypotheses that applicants get significantly higher scores than recruits, even though both samples were selected on the basis of operational MAP. Although the comparison of validities favors the hypothesis, that evidence is weakened by the fact that the recruit sample was also the sample on which the scoring key was developed. Nevertheless, the generalizability of data from recruits to applicants is not supported here.

## Drift in Validity

Method

For examining possible loss of validity over time, a non-operational key was used that had been developed on the 1982 recruit data. The criterion was

successful completion of the first six months of service (vs. discharge for failures to adapt). The "instrument" consisted of the 38 items mentioned earlier. Meeting the criteria of being on the operational form of MAP and being validated on non-graduates, the items could be used to compare results for non-graduates in different year groups who took MAP before entering the service. Three samples of such applicants were available: 2,374 in FY 82, 3,567 in FY 81, and 14,771 in 7/79-6/80. Because the 1982 key was not cross-validated by the developer, the 1982 applicants became a cross-validation sample. Thus, their data were used to see how much validity there was to drift in the first place. Validities in the form of Pearson $r$'s, mean total scores for the 38 items, and success rates (i.e., percent of sample completing the first six months of service) were compared over the four samples.

Results

Table 1 gives descriptive statistics for the recruits in 1982 and for three samples of applicants. In contrast with the original value of .18, validities for applicants in 1982, 1981, and 1979/80 were .02, .07, and .01, in order. The key did not effectively discriminate between examinees who went on to complete the first six months of service and those who did not: mean differences in scores for those two criterion groups reached a maximum of .18 SD in the three samples. Means out of a possible 71 points ranged from 28.3 to 44.2 points in the four groups, while success rates varied from .79 to .86. Using the 1982 applicants as a basis for confidence intervals on the means, we find significant differences ($p < .001$) in both the 1979/80 applicants and in the 1982 development sample. The normal approximation to the binomial found the development sample to have a significantly lower attrition rate than the 1982 applicants ($z = 12.28$; $p < .001$), all of whom had entered the Army.

Discussion

The low validity that was observed in the 1982 applicants amounts to a failure of the (non-operational) 1982 key to cross-validate. Thus, there was little if any original validity that could drift. Absent drift in validity, however, there were significant jumps in both predictor and criterion scores across samples. If changes occur in validity over time, they could be due to gradual trends in the population of applicants, to short range instability in the population, or to both. It is possible that similar variability could be found in subsamples of the 1982 recruits. In order for the 1982 developmental data to have any hope of producing a durable key, they would have to undergo a legitimate cross-validation. Elizabeth P. Smith and I are now working on this problem in-house at the Army Research Institute.

## Six Months Versus Longer Tenure

Method

An operational form of MAP, Form 4B, gave the data for this analysis. Its 60 multiple choice items were validated in 1977 on 2,280 male recruits who had not completed high school (Frank & Erwin, 1978). In content, the questions cover experiences in school, extracurricular activities, work history, and expectations of life in the service. The present examinees were 2,564 17-year old non-graduate males. They all took MAP as a pre-induction screen in Fiscal Years 81/82, entered the Army, and then received adverse discharges in their

first tour. For the first analysis, examinees were split into two groups, those discharged within the first six months of service (n=860) and those with longer tenures ($\bar{x}$=363 days; n=1,704). For each of the 60 items, a chi-square test of association was run on frequencies of response for each alternative by group. Cramer's V for the items was examined as well for estimates of strength of effects. To judge the potential of response choices for keying, differences between groups in rates of endorsing individual choices were examined in items giving a significant groups-by-response choice chi square.

A second similar analysis was done to see whether the sensitivity of biodata items to individual differences in adaptability is masked by lumping successful cases with those who receive bad discharges after six months. For this analysis, chi-square tests were run twice on the total sample of 5,941 non-graduate applicants in FY 81/82. This sample included those who served successfully. The sample was split differently for these runs: once as all dischargees vs. all successful cases, and once as all discharges within six months of entry vs. all other cases. Simple numbers of significant ($p < .05$) chi squares and median p values from the two splits were compared.

Results

In the analysis of dischargees, 10 of the 60 group-by-response choice chi square tests gave probabilities < .05. Of those, three had p's < .01. Cramer's V for the ten items ranged from .056 to .085, while V's for seven items with $.05 < p < .15$ were also above .05. The median level of significance for all 60 items was .30. In each of the ten items with the lowest p values, the single response choice which had the greatest difference between groups in rate of endorsement was tallied. The median of those ten maximal differences was 4.2% (range: 3.19 - 6.78%).

In the second analysis, 13 of the chi-squares on items gave $p < .01$ when the positive criterion group included bad discharges after six months. In contrast, when the criteron groups are pure (i.e., all bad discharges vs. only the successful cases), the significant items rise to 25. Median p values under the two conditions are .31 and .15, in order.

Discussion

The differences in response distributions are small for examinees who were discharged before and after six months. Given that the significance of chi-square is inflated by large sample sizes, and that the probability of Type I errors is great in such a large set of significance tests, a finding of ten items out of sixty with $p < .05$ is not large. Also, given the small values of V for those ten items and the small between-group differences in response frequencies, the data do not support keying the instrument separately for the periods of initial and field service. As for causes of attrition, the very similar distributions of predictor responses for the two groups in this dataset do not imply that the reasons for early and late attrition differ.

Although the usefulness of keying long and short tenures differently is not supported, the value of using a longer criterion tenure for key development is. In the analyses here, almost twice as many items were sensitive to

real differences in success when the positive criterion group was purged of later attritions. The practice of developing keys on six month success seems here to undermine the validity of the predictor.

## Conclusions

We now have evidence that traditional practices in developing biodata may have major flaws. A system for countering these problems is easy to conceive. Starting with a validated instrument, we could continually gather predictor scores of applicants and criterion scores of accessions. Today's selection measures would also be used as the predictor data for a later generation of scoring key, which would be based also on the performance measures. Updating of keys would then be ongoing rather than rare and ad hoc, as it is now. With ongoing updating, keys would be available after a minimal time lag and with appropriate generalizability (i.e., from applicants to applicants). Of course, increasing the criterion tenure would increase the time until new keys were available, but the best tradeoff between lag and quality could be determined empirically. Although problems in operating a biodata screen have been documented here, practical solutions are available.

## References


Atwater, D. C. & Abrahams, N. M. (1983). Adaptability screening: Development and initial validation of the Recruiting Background Questionnaire (RBQ) (NPRDC TR 84-11). San Diego: Navy Personnel Research and Development Center.

Erwin, F. W. (1985). Development of new Military Applicant Profile (MAP) biographical questionnaires for use in predicting early Army attrition. Unpublished manuscript.

Frank, B. A. & Erwin, F. W. (1978). The prediction of early Army Attrition through the use of autobiographical information questionnaires (Technical Report No. TR-78-A11). Alexandria, VA: Army Research Institute.

Guilford, J. P. & Fruchter, B. (1973). Fundamental statistics in psychology and educaton. New York: McGraw-Hill.

Goodstat, B. E. & Yedlin, N. C. (1980). First tour attrition: implications for policy and research (Research Report 1246). Ft. Benjamin Harrison, IN: Army Research Institute.

Hicks, J. M. (1981, March). Trends in first-tour armed services enlisted attrition rates. Paper presented at the annual meetings of the Southeastern Psychological Association. Atlanta, GA.

Means, B. & Heisey, J. (1985, in press). Educational and biographical data as predictors of early attrition. Alexandria, VA: Human Resources Research Organization.

Walker, C. B. (1985). The Army's Military Applicant Profile: Its background and progress. In B. Means (Editor), Recent developments in military suitability research. (In preparation)

# Use of Personal History Information to Predict Naval Academy Disenrollment

Joyce D. Mattson
Norman M. Abrahams
Rebecca D. Hetter

Navy Personnel Research and Development Center
San Diego, California 92152-6800

## Background

The Naval Academy has had great success in recent years in reducing its attrition from an average of about 36% for the Classes of '73 through '78 to the present rate of about 20%. Of the principal types of attrition -- academic and motivational -- motivational represents the largest single cause at 11%.

In an effort to reduce this motivational attrition further, the Naval Academy introduced an experimental background questionnaire initially developed in 1981. This instrument -- the Personal History Questionnaire (PHQ) -- is similar in its structured item format to instruments which have been useful in private industry for predicting such criteria as tenure, job performance, and creativity (Asher, 1972: Chaney and Owens, 1964: Schuh, 1971)

The questionnaire was administered to incoming Academy plebes in 1981 and 1982 (Classes of '85 and '86) and refined after each administration. Refinements were directed at: (1) improving item format, (2) retaining items which predicted attrition at the end of 1 year and (3) adding items suggested by the attrition analysis and a literature review (see Hetter and Neumann, in press). The final version of the questionnaire contains 85 multiple choice items, 25 of which were retained across all questionnaire revisions. The 85 items tap the factor-analytically-derived constructs in Table 1.

Table 1
Factor-analytically-derived PHQ Item Categories

| Category | Number * of Items |
|---|---|
| Time management | 6 |
| Social comfort | 6 |
| Economic achievement orientation | 7 |
| Confidence in ability to handle military aspects of Academy | 5 |
| Degree of skill, participation in athletics | 3 |
| Persistence | 6 |
| Degree of financial self-support | 4 |
| Confidence in ability to handle academic aspects of Academy | 3 |
| Risk-taking propensity | 5 |

| | |
|---|---|
| Firmness of career goals | 3 |
| Degree of enjoyment of adventurous activities | 3 |
| Personal organization, self-discipline | 4 |
| Confidence that will manage life at the Academy well | 3 |
| Sociability | 5 |
| Belief in ability to control own destiny | 7 |
| Amount of parental education | 2 |
| Amount of reading individual does | 2 |
| Style of upbringing | 5 |
| Importance of being in a position of authority | 1 |
| Strength of accomplishment needs | 1 |

* Several items were omitted from the analyses to prevent linear dependency.

Purpose

The purpose of this paper is to present the status of our research to date on the use of the Personal History Questionnaire to predict voluntary motivational attrition. In particular, this includes (1) development and cross-validation of an interim disenrollment scale using some of the plebe-based data collected for questionnaire development and (2) comparison of applicant and plebe administrations of the questionnaire to determine whether this plebe-based scale's validity would be likely to generalize to applicants

Scale Construction and Validation

The interim disenrollment scale was developed using majority male plebes from the Class of '85 who had no prior military service and who completed the PHQ during their first week of plebe summer. This subgroup contains most Academy students but is homogeneous enough that important differences between attritees and non-attritees should be apparent.

Using end of second year attrition information, the item responses of voluntary motivational attritees (N=108) were contrasted with the responses of non-attritees (N=764). Sixteen response alternatives among the 25 items common to all PHQ revisions showed an 8% or greater endorsement difference and were scored on the scale.

The scale was then cross-validated on a dichotomous attrition criterion (voluntary motivational attrition vs. non-attrition) using individuals from the Classes of '86 and '87, who where also administered the PHQ as plebes Both first- and second-year attrition information was available for the Class of '86, but only first year attrition information for the Class of '87. All racial, gender, and prior-service-experience subgroups were combined, since operational predictions would be required for all types of applicants.

Table 2 shows the results of these cross-validity analyses --- the biserial correlations of the disenrollment scale with cumulative voluntary motivational attrition at different tenure points. Biserial correlations were computed under the assumption that the attrition dichotomy actually reflects an underlying normally distributed continuum of motivation (McNemar, 1969).

Table 2
Relationship of the Disenrollment Scale
to Cumulative Voluntary Motivational Attrition

| Academy Class | Attrition status as of the end of: | Sample Sizes: Total | Non-attritees | VM* Attritees | r bis** |
|---|---|---|---|---|---|
| '86 | Plebe summer | 1158 | 1125 | 33 | .35 |
| | Year 1 Semester 1 | 1144 | 1071 | 73 | .26 |
| | Year 1 Semester 2 | 1113 | 1032 | 81 | .28 |
| | Second summer | 1112 | 1028 | 84 | .26 |
| | Year 2 Semester 1 | 1100 | 1004 | 96 | .27 |
| | Year 2 Semester 2 | 1078 | 958 | 120 | .26 |
| '87 | Plebe summer | 1341 | 1281 | 60 | .33 |
| | Year 1 Semester 1 | 1315 | 1208 | 107 | .32 |
| | Year 1 Semester 2 | 1291 | 1166 | 126 | 26 |

*VM = Voluntary motivational.
**Computed by contrasting VM attritees with non-attritees at each successive tenure point.

The correlations in both the '86 and '87 classes are in the .26 to .35 range, indicating reasonably good differentiation between motivational attritees and non-attritees. While correlations computed on the Class of '86 may be inflated by the use of '86 first-year attrition information in revising the PHQ, correlations computed on the Class of '87, which have no such contaminant, are remarkably similar

Assuming for a moment that validities derived on plebes could be generalized to the selection situation, Table 3 suggests the level of improvement in the prediction of voluntary motivational attrition which could result from adding the PHQ scale to the currently used Candidate Multiple.

Table 3
Incremental Validity of the Disenrollment Scale
for Predicting Voluntary Motivational Attrition

| Year Group | Attrition status as of the end of: | N | Candidate Multiple r bis * | Candidate Multiple Plus Disenrollment Scale R bis * |
|---|---|---|---|---|
| '86 | First year | 1109 | .06 | .28 |
| | Second year | 1076 | .12 | .28 |
| '87 | First year | 1288 | .00 | .26 |

Based on these data, the validity against first- or second-year voluntary motivational attrition could be increased substantially by the scale's use. Two caveats are in order, however. First, these increases represent upper limits, since in practice the Naval Academy would be unlikely to weight the disenrollment scale as heavily as is reflected here. Second, even the maximum gain would be slightly less if validities were corrected for range restriction.

Of interest, also, from an operational perspective, is the relationship of the disenrollment scale to other criteria which the Academy attempts to predict -- namely, military quality point ratio (MQPR), academic quality point ratio (AQPR), and academic attrition.

Table 4 shows that the scale has a slight positive relationship to MQPR, and no relationship to AQPR, suggesting that no deterioration in these performance areas would be likely to result from the scale's use. The negative relationship which is apparent with first-year academic attrition becomes slightly positive over the longer perspective of two years, suggesting that the long-term academic attrition rate would not be adversely affected by the scale's use, but that some of the attrition might occur earlier than at present.

Table 4
Relationship of the Disenrollment Scale
to Academic Attrition, AQPR, and MQPR

| Criterion | Academy Class | Attrition status as of the end of: | N | * r |
|---|---|---|---|---|
| Military QPR | '86 | Year 1 Semester 1 | 1092 | .05 |
| | | Year 1 Semester 2 | 1066 | 08 |
| | | Year 2 Semester 1 | 1021 | .08 |
| | | Year 2 Semester 2 | 992 | .08 |
| | '87 | Year 1 Semester 1 | 1235 | .04 |
| | | Year 1 Semester 2 | 1187 | .07 |
| Academic QPR | '86 | Year 1 Semester 1 | 1092 | .0[illegible] |
| | | Year 1 Semester 2 | 1066 | .03 |
| | | Year 2 Semester 1 | 1021 | 01 |
| | | Year 2 Semester 2 | 992 | .01 |
| | '87 | Year 1 Semester 1 | 1237 | .01 |
| | | Year 1 Semester 2 | 1187 | 01 |
| Academic Attrition | '86 | Year 1 Semester 1 | 1083 | - 20 |
| | | Year 1 Semester 2 | 1064 | -.00 |
| | | Year 2 Semester 1 | 1043 | .03 |
| | | Year 2 Semester 2 | 1017 | .05 |
| | '87 | Year 1 Semester 1 | 1226 | -.20 |
| | | Year 1 Semester 2 | 1205 | -.15 |

*Biserial correlations are used for the dichotomous attrition criterion; Pearson product-moment correlations for the other continuous criteria.

Generalizability of Validities to Applicants

To assess the likelihood that the disenrollment scale's promising validities would generalize to applicants, 1111 members of the Class of '88 who completed the PHQ as applicants were retested as plebes. If the two occasions yielded similar disenrollment scale scores, a plebe-based scale would probably retain its validity when applied to applicants; if the two occasions yielded markedly different responses, the scale's validity for selection would be uncertain.

Table 5 shows mean scores and applicant- plebe correlations on the disenrollment scale for the Class of '88.

Table 5
Means, S.D.s, and Applicant-plebe Correlation for the Class of '88 on the PHQ Disenrollment Scale (N=1111)

| Sample | Mean | S.D | r |
|---|---|---|---|
| Applicants | 4 875 | 2.566 | .52 |
| Plebes | 3.319 | 2.892 | |

The applicant means are more than 1/2 standard deviation higher than the plebe means, suggesting that applicants may slant their responses to increase their chances of Academy selection. At the item level, this distortion is apparent in the following representative item:

| Item and Alternatives | Percent Selecting Alternative: Applicants | Plebes |
|---|---|---|
| I find myself putting things off until the last minute. | | |
| A. Almost always | 0 | 3 |
| B. Often | 10 | 25 |
| C Sometimes, but not often | 48 | 50 |
| D. Rarely | 37 | 20 |
| E. Never | 5 | 3 |

In addition to item response distortion, the rank-ordering of individuals on the disenrollment scale also changes, as reflected in an applicant-plebe correlation of only 52. Successively removing items with the lowest applicant-plebe correlations from the scale improves this correlation by only 03.

These results indicate that individuals answer the PHQ differently as applicants than they do later as plebes and that validities based on plebes cannot necessarily be assumed to generalize to a selection situation. Rather,

plebe-derived scales must be validated on applicants or new applicant-based scales must be constructed.

Summary

In summary, results of this research suggest that:

(1) The items of the Personal History Questionnaire, when completed under low incentive to distort, can distinguish between individuals who leave the Academy for voluntary motivational reasons and those who remain.

(2) Individuals complete the PHQ differently as applicants than they do as plebes, in a way suggesting a tendency to slant applicant answers to increase the chances of Academy selection.

(3) The amount of change varies for different individuals so that applicant and plebe administrations yield different rank-orderings on the disenrollment scale and validity results from plebes cannot be assumed to apply to the selection situation.

REFERENCES

Asher, J L. The biographic item: can be improved? Personnel Psychology, 1972, 25, 251-269

Chaney, F B. and Owens, W.A Life history antecedents of sales, research, and general interest. Journal of Applied Psychology, 1964, 48, 101-105.

Hetter, R. and Neumann, I (in press). Preliminary development of a Personal History Questionnaire for predicting Naval Academy voluntary disenrollment (NPRDC in press) San Diego: Navy Personnel Research and Development Center

McNemar, Q. Psychological Statistics. New York: John Wiley & Sons, 1969. 4th Ed.

Schuh, A.J. The Predictability of employee tenure· A review of the literature. Personnel Psychology, 1967, 20, 133-152.

# Intercorrelations of Biographical Information, Aptitude Test Scores and Job Performance Ratings for 108 Occupations

Marvin H. Trattner
Office of Personnel Management, Washington, D.C.

In this paper General Aptitude Test Battery (GATB) scores, age, amount of both occupational experience and plant experience, employee sex, education level and job performance ratings are intercorrelated for employees in 108 occupations.

The research was accomplished to serve the following purposes:

1. To calculate and summarize a large number of validity coefficients for biographical variable and aptitude test predictors of job success. Validity coefficients for biographical variables are generally not as plentiful in the general literature as for paper and pencil tests. The Schmidt-Hunter validity generalization procedure is applied to the validity coefficients in order to summarize them and correct them for known artifacts such as criterion unreliability and predictor restriction-in-range.
2. To calculate the unique contribution to valid prediction of the biographical variables.
3. To determine whether predictor validity is related to job complexity level.

## Background

USES GATB testing and validation program. The United States Employment Service (USES) developed the GATB approximately 45 years ago for selecting and counseling applicants for vacancies filled by state employment services. The state employment services generally fill vacancies in low level industrial and clerical occupations. The GATB employs 12 subtests to measure the following nine relatively independent abilities (Anastasi, 1982):

| | |
|---|---|
| G - Intelligence | Q - Clerical Perception |
| V - Verbal Aptitude | K - Motor Coordination |
| N - Numerical Aptitude | F - Finger Dexterity |
| S - Spatial Aptitude | M - Manual Dexterity |
| P - Form Perception | |

Two and one-half hours are required to complete the GATB. The GATB subtests were developed on the basis of factor analysis, the aim was to employ subtests that intercorrelated only minimally except for the Intelligence aptitude which is measured by the sum of three subtests also used in part to measure Verbal, Numerical and Spatial aptitudes. The GATB tests applicants for most occupations in the U.S. economy hence it measures a very wide range of ability. It is similar in this respect to the Armed Forces Qualification Test.

---

The opinions expressed in this paper are those of the author, who is solely responsible for the accuracy of the contents, and does not necessarily reflect the views of the Office of Personnel Management.

The USES has been conducting validity studies on the GATB since its development. The measure of job performance which served as the criterion in these 108 studies were the summed ratings of job performance made by two supervisors. All studies used the same standardized performance rating form. Supervisors rated employees on 6 five-point scales. The scales were used to evaluate employees on the quantity, quality and accuracy of their work, their job knowledge, the variety of duties performed and also on overall job performance.

The data reported here were collected since the early seventies. Subjects were cumulated across geographical locations by occupation. The average number of cases per occupation was 222, a relatively large number when compared to the typical n employed in civilian validity studies. A concurrent validity design was used. That is, current employees in the occupation were used as subjects. The correlations obtained were assumed to apply also to job applicants except there is known to be a restriction-in-range in predictor scores which lowers obtained correlations and requires the use of a correction factor.

Schmidt-Hunter Validity Generalization Procedure (VGP). Schmidt and Hunter (1977) developed a procedure for summarizing validity coefficients for specific predictor/occupation categories. Previously no statistical procedure was available to summarize and correct a group of validity coefficients for known artifacts (such as sampling error, predictor restriction-in-range, and criterion unreliability) in order to estimate both the mean true validity and the standard deviation of the true validity coefficients. They and their associates employed the VGP to demonstrate in a large number of recent studies that predictor validities are generally much more consistent and of higher magnitude than was previously thought.

The VGP is accomplished by performing the following calculations on uncorrected validity coefficients obtained with a concurrent validity strategy.

1. Validity coefficients for a specific predictor/occupation combination are collected.

2. The mean true validity for the predictor/occupation is then calculated by first calculating the n weighted mean of the obtained validity coefficients. This mean is then corrected for attenuation for mean estimated criterion unreliability and also for mean estimated restriction-in-range in the predictor. If no criterion reliability or predictor restriction-in-range data are available for the validity studies at hand then hypothesized distributions are employed.

3. The standard deviation of the true validity coefficients is calculated next. First, the n weighted standard deviation of the obtained validity coefficients is calculated. Then the following extraneous sources of variance are subtracted from the variance of the obtained validity coefficients, the variance due to a) average sampling error b) differences in studies due to specific predictor reliability, criterion unreliability, and

restriction-in-range in the predictor. The remaining variance is then corrected upward by the same correction factor that was used to correct upward the mean obtained validity in order to estimate the mean true validity.

Additional considerations. A modification was required in the VGP for this study because the same specific predictor and the same criterion rating form was employed in every validity study in the data set. It was decided since the only predictor for any predictor/occupation combination was the specific GATB subtest that no predictor unreliability variance would be removed from the obtained validity variance. The criterion rating form was only one of many sources of criterion unreliability, consequently variance across studies in validity due to differences in criterion unreliability was removed from the obtained validity variance. For all predictors the standard hypothesized predictor restriction-in-range distribution was employed to correct the obtained coefficients. The mean selection ratio for this distribution is .45. In the latter two cases, the standard validity generalization procedure was followed.

The occupations included in the study carried Dictionary of Occupational Titles (DOT) codes. The DOT codes contain no job complexity designation. A job complexity score was needed since it has been frequently found that job complexity moderates predictor validity. A job complexity score was obtained for the 108 occupations by summing scaled scores for estimates of required mathematical development, language development and specific vocational preparation (U.S. Department of Labor, 1981). The complexity score sums for the 108 occupations ranged from scores of 4 to 16. Most of the occupations for which data were available were low level clerical and industrial jobs. The two occupations receiving the highest complexity score were Medical Technologist and Electronics Technician. Some examples of occupations receiving the lowest score were Hand Packager, Garment Folder, and Small Product Assembler. The distribution of complexity level was positively skewed, most of the occupations clustered at the lower complexity levels.

## RESULTS AND DISCUSSION

There were individual score data for 23,917 employees in 108 occupations. Complete data were available for all subjects.

Mean true validity for each predictor for all occupations combined. All predictors (GATB test scores and biographical variables) and the criterion score were intercorrelated separately for each occupation. The obtained validity coefficients were then combined for all occupations for each predictor. The mean and the standard deviation of the distribution of obtained validity coefficients was calculated for such predictor. The VGP was then applied to the distribution of 108 obtained validity coefficients for each predictor to calculate the mean true validity, the standard deviation of the true validity coefficients and the "95% credibility value" (the point above which 95% of the true validity coefficients lie). Table 1 contains the obtained and true validity coefficient distribution statistics for each predictor.

Table 1

Mean and Standard Deviation of the Obtained and True Validity Distribution for the Predictors

| | Predictors | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Obtained Validity Distribution | Sex[a] | Age | Educat. Level | Plant Exper. | Occup. Exper. | GATB Aptitudes G | V | N | S | P | Q | K | F | M |
| Mean | -00 | 07 | 02 | 19 | 19 | 20 | 16 | 20 | 13 | 14 | 16 | 08 | 09 | 10 |
| Standard Deviation | 08 | 11 | 10 | 11 | 10 | 09 | 08 | 08 | 08 | 08 | 08 | 08 | 08 | 09 |
| True Validity Distribution | | | | | | | | | | | | | | |
| Mean | -01 | 16 | 05 | 39 | 39 | 43 | 34 | 42 | 28 | 29 | 33 | 18 | 20 | 22 |
| Standard Deviation | 10 | 19 | 15 | 15 | 15 | 06 | 09 | 04 | 10 | 09 | 07 | 12 | 09 | 12 |
| 95% Credibility Value | -18 | -16 | -20 | 14 | 15 | 32 | 21 | 35 | 12 | 14 | 22 | -01 | 04 | 02 |

note: Decimal point omitted for validity coefficients.
a: Scored 1 = male, 2 = female.

Table 1 reveals that when all occupations are grouped together all tests have higher mean validity than the mean validity for sex, age or education level. Both plant experience and occupational experience correlate relatively highly with job performance ratings, only intelligence and numerical ability have higher mean true validity. All the biographical predictors have considerably higher true validity variance than the test predictors.

When the mean predictor intercorrelations were clustered across the 108 occupations the following four groups of predictors were obtained - experience cluster (age, plant experience, occupational experience), cognitive ability (intelligence, verbal, numerical), perceptual ability (spatial, form perception, clerical perception), psychomotor ability (motor coordination, finger dexterity, manual dexterity).

Unique contribution to predictor validity of the experience cluster) for three job complexity categories. The unique contribution to validity of the summed standard scores for age, plant experience and occupational experience when added to the cognitive ability score (sum of G, V, N) was determined separately for high, middle, and low complexity occupations. The calculations were made by first averaging the correlations for the occupations separately for the three complexity categories. The thirteen job complexity categories were compressed to three by combining complexity levels 4-8, 9-11, and 12-16. Employee sex and educational level were eliminated from consideration since they correlated so poorly with the criterion. All variables were unit weighted and standard scores were employed. The calculations were made for the obtained validity coefficients rather than the true validity coefficients. Cognitive ability correlated highest with the criterion and was selected first, except for the low complexity level where it equalled experience.

Table 2

Obtained Correlations of Sum of Unit Weighted Standard Score Cognitive Ability and Experience Predictors for Three Job Complexity levels.

| Job Complexity Level | Predictor Composite Employed G, V, N, Age Plant Exp. Occupa. Exp. | G, V, N | Unique Validity of Experience Cluster |
|---|---|---|---|
| Low | .271 | .180 | .091 |
| Medium | .294 | .245 | .049 |
| High | .270 | .195 | .075 |

Table 2 reveals that the experience cluster adds a considerable amount of unique validity to the validity obtained by the cognitive predictors for all three job complexity levels. The unique validity of the experience cluster is identical to the part correlation of the experience cluster with the criterion. Neither the perceptual or psychomotor ability clusters add unique validity to the cognitive ability cluster. No great differences in validity occurred at the three complexity levels.

Consistency of validity coefficients across job complexity levels. The mean true validity for each predictor was calculated separately for each job complexity level. The product moment correlation of the mean true validity coefficients with the job complexity scale for each predictor confirmed some foreseeable relationships. The following predictors had higher validity in lower complexity occupations: perceptual and psychomotor abilities, age, plant experience, occupational experience. Education level had higher validity in higher level occupations. Women were rated higher than men for lower level occupations and men were rated higher than women for higher level occupations. However, it is important to emphasize that while many of these correlations between validity and complexity level were significant, in general, they would not lead to the use of different predictors for different complexity levels. For instance, although education level is correlated higher with job performance for more complex jobs, the validity for the more complex jobs is not high enough to add unique variance to the cognitive ability score.

I have reported results that showed a lack of important moderation of validity by job complexity in two different contexts. Approximately the same validity was obtained for the predictor clusters for the three different complexity levels and validities for the specific complexity levels show that very rarely would a different predictor be used than the ones that applied to all complexity levels combined. This may be because these data were cumulated over geographical locations which is known to depress the validity coefficients.

## Summary

The following results were obtained:

1. GATB tests in general had higher correlations than biographical variables with supervisory ratings. Two biographical variables, plant experience and occupational experience however had high correlations with supervisory ratings.

2. The experience cluster score, consisting of plant experience, occupational experience and age added unique valid variance to the cognitive ability cluster score. Neither perceptual nor psychomotor ability added valid variance to cognitive ability.

3. The data show some relation between predictor validity and job complexity level but not enough to justify the use of different predictors for different complexity levels. A reason for this lack of relationship was advanced.

## REFERENCES

Anastasi, A. (1982). Psychological Testing. New York: Macmillan, 5th edition.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

U. S. Department of Labor Employment and Training Administration (1981). Selected characteristics of occupations defined in the Dictionary of Occupational Titles. Washington: U. S. Government Printing Office.

# ANALYTIC PREDICTION OF TRAINING DEVICE EFFECTIVENESS

Louise G. Yates
Douglas Macpherson

*U. S. Army Research Institute*

One goal of the U. S. Army is to design cost effective training equipment which prepares troops to do necessary tasks. The design and acquisition of this equipment is usually accomplished during the design and acquisition phases of the parent weapon system. Although the training equipment design is supposed to be based on field tests, often the field tests are eliminated due to logistical problems related to the parent equipment acquisition and design process. When field tests are performed, they often come too late in the acquisition process, due to the same logistical problems, to allow change of the device design based on the results of the field test. Thus, device design often becomes dependent on the analytic procedures of each program manager as no formal analytic procedures for evaluation of training device design have been implemented.

The Army Research Institute has designed the Device Effectiveness Forecasting Procedure (DEFT) to fill the need for a formal analytic procedure that can be used at various stages in the design/acquisition process to evaluate the proposed device design, proposed design changes, and alternate device designs. The instruction and training principles contained in DEFT can also be used as a guide in the initial design process. DEFT is based on a review of relevant literature and on the development of a conceptual approach which takes into consideration theoretical and practical issues of training device design, development, and evaluation. This approach led to a program development model which resulted in a network of hypotheses that relate program inputs to intermediate outcomes.

DEFT is computerized in an interactive menu-driven format for use on the IBM PC or compatible computers. It requires ratings from subject matter experts (SME) on device, operational equipment, trainee, and training system characteristics. It can be used at three levels of complexity, DEFT I, II, and III, varying with the amount of information available to the user. DEFT I is designed for use early in the design process when only general information is available to the user. At this early stage alternate device designs may need to be compared or one device design evaluated. An evaluation made at an early stage is especially valuable as design changes have much less cost and time impact.

DEFT I requires eight global ratings from the SME based on general information about characteristics of the device, the parent equipment, the training needed, the trainees, instructional features of the training, and the physical and functional similarity of the device to the operational equipment.

DEFT II and III are designed to be used later in the design/acquisition cycle when more detailed information is available. This can be the first use of DEFT, or the use of DEFT II or III to check the forecast made by DEFT I, or the use of DEFT II or III to determine the effect of proposed changes in the device. DEFT II ratings are made at the task level. There are thirteen kinds of ratings, some of which are rated for each task trained on the device or the operational equipment.

DEFT III can be used at either the task or subtask level. It requires 35 different types of ratings, some of which are required for each task/subtask level. Added to the information needed to make DEFT I and II ratings are certain features of controls and displays on both the device and the operational equipment.

DEFT II and III provide more diagnostic capability than does DEFT I. With DEFT II and III it is possible to predict the impact of changing one or more specific characteristics of the device or the training system on the effectiveness of the training. DEFT III which requires the most information can provide more diagnostic information than DEFT II.

The information input into DEFT is converted into several numerical indices which are used to estimate device effectiveness. Table I contains the evaluation summary from a DEFT I analysis of a naval training simulator for the SH-3 helicopter.

The indices listed in the summary are the same for DEFT II and III. The numbers in the left column come directly from the ratings (For DEFT II and III these numbers are the ratings summed and averaged over tasks/subtasks and rating scales.). Indices in the other two columns are derived from mathematical combinations of the numbers to their left and above them in the table. The three summary indices (right column) are for acquisition effectiveness, transfer effectiveness and the sum of these two, total effectiveness.

Training Problem defines the deficiency in skills and knowledge that the new trainees have relative to criterion performance on the training device and the difficulty trainees would have in overcoming the deficiency. Training Efficiency is rated on the basis of the instructional features and training principles incorporated in the training device. Acquisition Effectiveness is designed to give a "poor" score to a device that has a large training problem and is inefficient in dealing with it.

TABLE I

DEFT I Evaluation Summary

| | | | |
|---|---|---|---|
| *Performance Deficit* | 50 | | |
| *Learning Difficulty* | 60 | | |
| *Training Problem* | | 30 00 | |
| *Quality of Training Acquisition* | 90 | | |
| *Acquisition Efficiency* | | .94 | |
| *Acquisition Effectiveness* | | | 31 91 |
| *Residual Deficit* | 20 | | |
| *Residual Learning Difficulty* | 30 | | |
| *Physical Similarity* | 80 | | |
| *Functional Similarity* | 80 | | |
| *Transfer Problem* | | 6 00 | |
| *Quality of Training Transfer* | 80 | | |
| *Transfer Efficiency* | | .89 | |
| *Transfer Effectiveness* | | | 6.74 |
| *Total Effectiveness* | | | 38 65 |

The Transfer Problem is defined by the amount of learning required on the operational equipment, or by other means, after training on the device is completed, ratings of the difficulty of this learning, and ratings of the functional and physical similarity of the device controls and displays to those on the operational equipment. The Transfer Efficiency Index is computed from ratings of the instructional and training principles in the device that contribute to transfer of training.

Like the Acquisition Effectiveness Index, the Transfer Effectiveness Index gives a "poor" score to a device that has a large transfer problem and is ineffective in dealing with it.

## Evaluation of DEFT

At the time that DEFT was completed a field evaluation could not be funded. Therefore, an analytic assessment was made to assess certain scalar properties of DEFT and to examine interrater agreement. Based on expected score distributions obtained from Monte Carlo distributions, devices with different DEFT ratings can be evaluated as different or not. Based on certain reasonable assumptions regarding the distribution of expected input values, DEFT outputs are interpretable and meaningful. Sensitivity analysis, which tested the impact on DEFT output of

varying input values, indicated that, with the exception of the two efficiency scales, all input values had the same effect on total score. Efficiency scales had a larger effect than other scales.

Interrater agreement was examined by having four raters use DEFT I, II, and III to evaluate three training devices, the M-60 Gunnery Trainer (VIGS), the Burst-on-Target (BOT) Trainer, and the Maintenance Procedure Simulator (MPS) for the E-3A Navigation Computer System. The seven summary indices shown in the two columns on the right of Table I were computed for each DEFT rater and trainer combination. From an analytic assessment of these results it was concluded that there was substantial interrater agreement for all DEFT indices across the three devices. However, it is necessary for raters to agree on their assumptions regarding the device, trainee population , device utilization, and the meanings of the various DEFT scales prior to making the ratings.

Funds have been allocated for an interservice contract effort, starting in FY86, to field test DEFT. SME ratings of a number of training devices with known performance capabilities will be used. One output expected from this effort is the conversion of pertinent DEFT index scales to scales based on meaningful training concepts such as trials to mastery, training time, etc.

Wide interservice interest has been displayed in DEFT. Personnel from the Training Analysis and Evaluation Group (TAEG) of the Naval Training Evaluation Center have used DEFT I and II to evaluate the 2F64C helicopter trainer, under four different conditions, to determine if DEFT in its present form will be useful to TAEG. Interrater correlations for two raters using DEFT II ranged from .81 to .97 for the seven groups of ratings which result in the seven summary indices of Table I. Correlation of DEFT I and II measures of transfer efficiency with transfer ratios based on trials to mastery in the actual helicopter, after training on the simulator, averaged .54. The Ns were small for all of these results, so the results are not conclusive, but they are encouraging.

In the near future the Marine Corps will be evaluating DEFT for their use by using it to evaluate the LVT Landing Vehicle Trainer. In addition, the Air Force has requested information on DEFT.

The interest of these agencies in DEFT appears to confirm the need for such a technique, and initial results of the evaluation of DEFT, although incomplete, lead us to believe that DEFT can become a very useful tool in designing and evaluating training devices.

## Bibliography

Rose, A. M., Wheaton, G. R., & Yates, L. G. Forecasting Device Effectiveness: Volume I. Issues. Technical Report 680. Alexandria, VA: U. S. Army Research Institute.

Rose, A. M., Wheaton, G. R., & Yates, L. G. Forecasting Device Effectiveness: Volume II. Procedures. Research Product 85-25. Alexandria, VA: U. S. Army Research Institute.

Rose, A. M., Wheaton, G. R., & Yates, L. G. Forecasting Device Effectiveness: Volume III. Analytic Assessment of Device Effectiveness Forecasting Technique. Technical Report 681. Alexandria, VA: U. S. Army Research Institute.

# Research on Decision Aids for Training Design and Evaluation

Angelo Mirabella
Army Research Institute
for the
Behavioral and Social Sciences

A key mission of the Army Research Institute (ARI) is to produce decision aids which will help training developers design and evaluate media, methods, and programs of instruction. This paper summarizes a program of research (focusing on maintenance training) to achieve that mission. The summary will address some research products which have been completed, and some which are still being developed. It will describe cooperative efforts among the services.

The programs long-range objective is to combine a number of job aids into a decision-support system (DSS) which will unburden the training developer. The DSS should provide an audit trail for his or her decisions, instill confidence that cost-effective training is being developed, improve communication among developers, and yet be easy to understand and use. Yet, products which are easy to use often require very complicated science and technology to develop as well as substantial investments in time and money. Therefore, we have set an ambitious, high risk goal. Why should we pursue it and what obstacles do we face? First, why do we need job aided, perhaps high technology, approaches to training development?

We need them to support increasing requests for accountability from the Congress. Congress has asked the Services to do a better job of planning, explaining, and defending how they train their respective forces. OSD and the Services have responded with new policies. For example, recent directives from OSD on training device acquisition and computer-based instruction, require that the choice of a training medium, its specific design and use in programs of instruction be justified. This requires data on the expected and/or measured effectiveness of the medium. At the very least, training managers and weapon developers will have to point, increasingly, to authoritative bases for decisions on training system design.

At the grass roots level, we need decision aids to support the Service schools and their training directorates. The schools face a very difficult task in developing training not only for current weapon systems, but for those on the drawing board. Their difficulty is exacerbated by shortages of training development specialists and turnover among those that are available. One result is that the schools develop training partly on the basis of intuitive judgement, which may be difficult to explain or justify. Therefore, the need for decision aids, if not a decision support system, is evident. Why haven't we met this need before? The answer lies in chronic gaps in the science and technology of training design, particularly for maintenance.

1. Lack of a critical mass of scientific and expert data to support training design decisions, and

2. Lack of methodology for translating scientific and expert data into design technology. These will continue to be chronic until a concerted and systematic effort is made to deal with them. Working towards a decision support system forces the issue for reasons to be explained shortly.

## The Program

Our program at ARI is designed to help move the Army from a paper-based, approach to developing training to a higher technology, more objective, more easily documented, and more defensible approach (Figure 1). We have a systematic, institutionalized approach in Instructional Systems Design (ISD). But ISD has not been very successful, because it places a heavy information overload on the training developer and requires a many subjective judgements that are difficult to make. A decision support system could shift much of this burden from human to machine.

**JOB AIDS FOR TRAINING SYSTEM DESIGN, DEVELOPMENT, ACQUISITION**

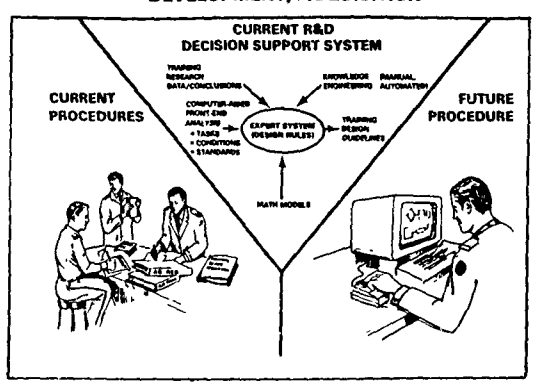



Figure 1

The middle of Fig. 1 illustrates how we can bring about this kind of magic. This section shows one kind of DSS, based upon artificial intelligence in the form of an expert system. Here, the training developer would communicate to the computer the maintenance tasks to be trained, conditions of training, and standards of achievemett, and receive a set of recommendations for designing a maintenance simulator. The designer would have to provide the front-end analysis, i.e., tasks, conditions, and standards. But once this was done the design decisions would be made automatically by a set of built-in "if..then rules." Moreover, job aids can and have been developed to assist with front end analysis, so even this human function could be machine assisted.

Having decided to bring about this kind of magic, what implications does the decision have for R&D strategy? The answer is provided by outlining the steps required to construct a DSS. These provide the skeleton for the research program which I am addressing in this paper.

Step One is to accumulate a relevant set of scientific and expert data. In our case, that means data which show how various training variables influence the acquisition and transfer of maintenance skills. It also means some form of accumulated wisdom from training experts with much experience. We are in trouble on both counts, and we're only on Step One. For many reasons, we lack adequate amounts and kinds of the first set. But we're also in trouble on the second set because the training business does not possess the kinds of expertise that one finds in more mature disciplines like engineering, or medicine.

Step Two is to organize the scientific data so that it is easy access and process. The traditional approach has been to store scientific data in journals which are often difficult to obtain and understand. Data Base Management System technology offers an alternative approach.

Step Three is to develop and assess a variety of technologies for processing the data in Step Two so that we can generate useful conclusions and principles of training design. Many technologies are available. But while they work well in physical science, engineering, and medicine, they have not yet been exploited for training problems. Examples of these technologies include the methods of operations research, mathematical modeling, pattern recognition techniques, decision analysis, and artificial intelligence. In Step Four, the results of the first three steps have to be packaged for easy use by the ultimate customer.

At ARI we have been or plan to do some work under each of the four headings above. For example, under Step One we have set up a maintenance training research facility at George Mason University which is somewhat unique. We're using Army simulators and studying Army maintenance tasks (diesel and Hawk radar repair), but under the controlled conditions of a laboratory. Over the next three years, we hope to accumulate, a systematic collection of training effectiveness data which can be converted into

specifications for designing simulation and computer-based instruction of maintenance skills. These in turn would provide part of the basis for an expert system we are developing.

We have completed the first two of approximately 10 major experiments. We will be looking at the effects of alternative media, ( e.g. computer-aided instruction vs simulator training), training strategies, and individual differences. We want to identify the most potent training system variables and translate them into design principles. Initial results have been described by Swezey (1985). Under Step Two we have constructed a preliminary training simulation research data base. It contains of 150 research papers which have been analyzed into 20 categories of information. The data exists in hard copy (Ayres, et al, 1984) and exists in computerized form. We plan to conduct a meta-analysis on an expanded and revised version of the data base and use the results as inputs to our expert system. This brings us to Step Three: developing techniques for converting data into useful decisions.

As part of Step Three we have developed procedures to help design and evaluation of training. The evaluation aids are finished or nearly finished products ready to implement, so I will describe them first. One of these products, a method called Comparison-Based Prediction of Training Device Effectiveness, is based upon Comparability Analysis (CA) which is currently in use by the military -particularly the Air Force - for estimating the reliability of new weapon systems. We have adapted CA for use in training effectiveness analysis. The analyst uses his or her knowledge (or performance data if available) about existing training devices to estimate the effectiveness of new devices. A handbook, explaining how to conduct the analysis has recently been produced (Klein et al, 1985) and we are currently exploring ways to field test and implement it.

We have also developed an alternative - and in this case computerized - evaluation procedure called the DEFT - Device Evaluation Forecasting Technique (Rose et al, 1985a). DEFT is a decision analytic method - as is CBP, but one which is based on a mathematical model. The analyst is led through questions about the device being evaluated. His or her answers (judgements on a scale of 0 to 100). provide the inputs for formulas which yield various measures of device effectiveness. These, in turn, are combined mathematically into an measure of device effectiveness. The method is being used on a trial basis by the Naval Training Equipment Center and will be extensively demonstrated and field tested during the current fiscal year under a new program called ASTAR (Automated Simulator Test and Assessment).

Lastly, we have developed a draft guidebook on how to empirically test simulators and training devices for transfer of training effectiveness. This document is designed to help action officers who are responsible for evaluating training devices but who may not be experts in testing. It aims to teach the reader enough about the basic concepts, language and methods of testing so that he or she can communicate with

evaluation specialists and assess their recommendations. It is also designed to help that action officer argue for adequate testing resources. We have made preliminary plans to hold a workshop for the Army Training Device Community.

Currently, our Step Three work is devoted to building expert system for design of training simulators. So far we have built a demonstration system containing 60 rules. We'd like to triple the size of the demonstration and then take it to a TRADOC School to develop it further with the help of maintenance instructors.

Recently we began a joint program with NTEC and AFHRL to develop automated techniques for producing training courseware. Our objective is to construct models of students, instructors, and subject matter experts and then convert these using artificial intelligence methodology into automated procedures for doing training front end analysis and building diagnostic routines. We plan to test the results of this re search on a training problem that holds common interest for the three services. We anticipate combining training evaluation methods, the expert system for training device design, and the automated procedures for courseware development into a comprehensive decision support system for training design.

Under Step Four we have written some concept papers identifying and defining issues in the human engineering and design of a decision support system (Macpherson & Mirabella, 1985; Singer & Hays, 1985). We are also developing a joint program with the Ordnance School so that we can conduct research on user requirements and test approaches to meeting those requirements, including specific human engineering designs for the decision aiding technology that evolves from our various lines of research.

## Future Plans and Research Pay-offs

Where do we go from here? We plan to make a concerted effort to field test, demonstrate, and implement the training evaluation tools already completed. We will also continue research on artificial intelligence approaches to training development (expert system and automated courseware procedures). Finally, we will be developing a test site where we can study user requirements and match them with the technologies evolving from our basic research program. We are targeting FY87 to complete the system components, FY88 to integrate them, and FY89 to demonstrate the resulting DSS at the test bed site.

Suppose we succeeded beyond our wildest dreams and did, in fact, evolve a DSS which could significantly unburden the training developer. What benefits would accrue? How could we use such a product to help the Army meet its future maintenance training requirements? First, it would contribute significantly towards meeting Congressional and OSD/Service directives for accountability in developing, executing, and managing training. Secondly, it could help document future training requirements.

A projection of future population characteristics could be used as input data for the DSS, and if it were sufficiently comprehensive, it could return specifications for appropriate future training system designs. If, on the other hand, we discovered that a particular future projection could not be represented in the DSS, we could then document this state of ignorance and use that documentation as a guideline to new research. With such a system, we should also be able to significantly reduce the problem of orienting future action officers who will be faced with the responsibility for training development. Finally, we could cure the chronic ailment of shortages in training and educational specialists.

Whatever the future holds in store because of changes in Army missions, technology, resources, or population characteristics we will be better able to meet the future if we close critical gaps in the science and technology of training design. The concept of the DSS provides a useful way to organize the research and development which will be required to close those gaps.

## References


Ayres, A., Hays, R.T., Singer, M.J., & Heinicke, M. An Annotated bibliography on the use of simulators in Technical Training. ARI Research Product 84-21, October, 1984.

Klein, G.A., John, P.G., Perez, R., & Mirabella, A. Comparison-Based Prediction of Cost and Effectiveness of Training Devices - A Guidebook. ARI Research Product, In press, 1985.

Macpherson, D.G. & Mirabella, A. APPLIED META-ANALYSIS: A Procedure for Speeding Innovation by Transferring Scientific Knowledge More Quickly. In "Artificial Intelligence and Simulation", William M. Holmses, Ed. San Diego: Simulation Councils Press, 1985a.

Rose, A.M., Wheaton, G.R., & Yates, L. Forecasting Device Effectiveness: Procedures. ARI Research Product 85-25, June, 1985.

Singer, M., and Hays, R. New Directions in Simulation Research: Generating, Handling, and Delivering Empirically Based Guidance. ARI Technical Report RR 1389, January, 1985

Swezey, R.W., Criswell, E.L., Huggins, R.S., Hays, R.T., & Allen, J.A. Training Effects of Hands on Practice and Three Instructional Methods on a Simple Procedural Task. ARI Technical Report, 1985.

# THE ARMY EXPERIENCE SURVEY: METHODOLOGICAL HIGHLIGHTS

Jeanna F. Celeste
Westat, Inc.
1650 Research Boulevard
Rockville, Maryland 20850

Forecasting future enlistment and reenlistment trends, and developing effective recruitment and reenlistment policies and procedures has become a very complex task for the all-volunteer forces. Recent upturns in the state of the U.S. economy have placed military recruiters in competition with one another and with the civilian workplace for the limited labor resources. Military policy makers regularly need current information about the characteristics of people enlisting as well as those separating in order to make informed decisions. More importantly, they need to understand the reasons for enlistment and separation.

The Army Research Institute for the Behavioral and Social Sciences,* recently awarded a contract to Westat, Inc. to design and administer a large random survey of Army veterans. A major impetus for performing the survey was the interest expressed by the Secretary of the Army in assessing the post-service experiences, plans for reenlistment or reserve duty, and attitudes toward military experience of recently separated soldiers.

It is hoped that the responses of recently separated Army veterans can provide important data to illuminate their expectations and values concerning civilian life and to detail gaps between their expectations versus experiences in their recent period of active duty. In addition, when compared to data from Army subgroups who elected to reenlist, these data can illuminate differences in the expectations/values across various population subgroups.

This paper reviews the approach taken to obtain these data. Sampling and survey administration procedures are described along with a brief discussion of survey outcomes.

## Survey Administration Procedures

The Army Experience Survey was comprised by a broad variety of tasks. A sample of recently separated Army veterans was designed and selected, and a multi-method approach to survey administration was implemented to include multi-wave survey mailings, telephone followup of mail nonrespondents, and conducting a limited respondent tracing effort. Each of these survey components are described below.

---

*The views expressed in this paper are those of the author and do not necessarily reflect the official policy or position of the Department of the Army, the Army Research Institute for the Behavioral and Social Sciences, or the U.S. government.

Sample Design. An extensive sample design effort was undertaken to select a representative sample of recent Army veterans. The population of interest was defined as enlisted soldiers who separated from service between October 1981 and September 1984.* The sample was drawn using Army personnel records maintained on the FY82, FY83, and FY84 versions of the Enlisted Master File. The targeted population was further defined on the basis of their separation statuses. The survey sample was selected drawing most heavily upon young soldiers separating after completion of a successful first term of enlistment. Smaller samples were also obtained on four other groups of separatees including: first-term attritees, soldiers separating at two different points in mid-career (i.e., soldiers serving more than one term but less than ten years, and soldiers serving more than ten years but not retired), and enlisted retirees.

The distribution of sample members by separation status is presented in Table 1. As the table suggests, the group of primary interest was the sample of first-term separatees. A related priority was to be able to perform comparisons of first-term separatee and attritee responses. The samples of separated mid-careerists and retiree groups were of secondary interest in the survey effort.

Table 1

Separation Status of Army Experience Survey Sample

| Separation Status | n |
|---|---|
| First-term Separatees | 5,413 |
| First-term Attritees | 1,616 |
| Soldiers with more than one term of service, but less than 10 years | 601 |
| Soldiers with 10 or more years of service, but not retired | 500 |
| Retirees | 500 |
| Unknown separation status | 123 |
| TOTAL SAMPLE | 8,753 |

The sample design used provided a representative sample of the desired size for each of the six separation groups identified in Table 1. The individual categories of five major variables

*Excluded from the sample frame were: soldiers who separated to enter officer programs, medical retirees, soldiers who died while in service and separating soldiers granted early release for reasons other than insufficient retainability.

were used to determine sample representativeness. These variables included race, gender, level of AFQT (Armed Forces Qualification Test), initial term of enlistment, and time elapsed since separation. Table 2 presents the 15 categories which were considered in designing the sample.

Table 2

Sampling Categories (Within Separation Groups)

| Race | Gender | Initial Term of Enlistment | AFQT Level | Time Elapsed Since Separation |
|---|---|---|---|---|
| White | Male | 2 years | Categories I and II | 1 year |
| Black | Female | 3 years | Category IIIA | 2 years |
| | | | Category IIIB | |
| Hispanic | | 4 or more years | Category IV and below | 3 years |

A representative systematic random sample was drawn from the population frame. Then the basic sample was supplemented with additional systematic random samples to achieve the desired sample size across the sampling categories. This design resulted in a self-weighting sample.

Survey Administration. The survey administration took a multi-method approach. In general, survey procedures consisted of multiple survey mailings, respondent tracing, and telephone followup of nonrespondents. Address information was provided by the Defense Manpower Data Center (DMDC) on computer tapes. Respondent tracing procedures were employed for cases in which: (1) no address was available from DMDC, (2) survey mail was returned with no forwarding address, and (3) telephone followup was required. Each of these AES administration procedures is briefly reviewed below.

DMDC provided the complete set of Army personnel records from the Enlisted Master Files for FY82, FY83, and FY84. In addition, they performed record searches of several different military personnel files to locate address information for this population.

A sample of 8,753 veterans separating from the Army during FY82-FY84 was selected from the EMF files. DMDC was able to provide at least one address for 83% of the sample (n=7,232). In some instances, multiple addresses were provided. There were no initial working addresses for 17% of the sample (n=1,521). These cases underwent sample tracing.

A variety of sources were employed to assist in locating addresses and/or telephone numbers for sample members. Sources consulted include: NPRC/RCPAC (National Personnel Record Center/Reserve Component Personnel Administration Center), the U.S. Postal Service, the Johns Holding Company, Telematch, and Directory Assistance.

The sample was separated into three groups distinguishable by the type of tracing required: Group I had addresses provided by DMDC which were presumed valid;* Group II sample members did not have addresses provided by DMDC and required manual searches of personnel records at NPRC/RCPAC; addresses for the third group of sample members could not be located either on DMDC tapes nor at NPRC/RCPAC and were, therefore, forwarded to the Johns Holding Company (the credit bureau) for address searching. Sample members who apparently had good addresses but who never responded to the survey were forwarded to the Telephone Research Center for followup. In order to obtain telephone numbers, sample members' names and addresses were sent to Telematch and/or Directory Assistance was contacted in the city of their last known address.

Workable addresses were never located for 2% (n=166) of the AES sample. However, 14.5% (n=1,270) of the sample members were located by NPRC/RCPAC, and 5% (n=418) of the sample were located by the Johns Holding Company. Telematch confirmed addresses and provided phone numbers for 28.6% (n=2,501) of sample members. Directory Assistance was contacted when cases referred for telephone followup either had no telephone number provided by Telematch or the number provided did not reach the sample member.

Upon receipt of an address, the sequence of survey administration proceeded as follows:

- Mailing of prenotification letter;
- First-wave mailing of survey;
- Reminder/thank you postcard;
- Second-wave** mailing of survey and prenotification letter; and
- Telephone followup interviewing.

---

*Valid addresses for Group I sample members met one of the following criteria: (1) the mail was not returned by the Postal Service, or (2) mail was returned with an address correction and mail sent to the new address was not subsequently returned by the Postal Service. Mail which was not returned as undeliverable was, for the most part, considered to have reached the intended party. This assumption was tested through two mail experiments.

**In cases in which no response was received to the first survey mailing and multiple addresses were available, either from DMDC or NPRC/RCPAC, surveys were sent to all known addresses.

Survey Results. Table 3 presents the results of the survey effort by method of contact resulting in a completed survey (i.e., mail, telephone). As Table 3 indicates, a majority of completed surveys were obtained through the first mailing of the survey (62.0%). The second wave mailing elicited a fair response rate bringing in another 779 completed surveys (18.6%) of the total completes. Telephone followup provided an additional 816 completed surveys--a very successful effort particularly considering that their sample was taken from among the respondents who had already been sent four survey mailings without achieving a response.

Table 3

Sample Completion Rate by Survey Method

| Mail Sample n=8,378* | | |
|---|---|---|
| | N | % of Completes |
| 1st Wave Mail Completes | 2,601 | 62.0% |
| 2nd Wave Mail Completes | 779 | 18.6% |
| Telephone Followup Completes | 816 | 19.4% |
| | 4,196* | 100.0% |

*Does not include completed surveys from Telephone Only Sample, n=162.

## Research Implications

The Army Experience Survey is the first known successful effort among the military services to survey veterans about their post-service experiences, and the impact of their service experience upon their civilian lives. Given schedule and budgetary constraints, the outcome of the tracing efforts and survey response rates are encouraging. Telephone followup of nonrespondents appears to be a highly effective method of increasing sample response rates and may prove to increase sample representativeness by reaching sample members less likely to complete written surveys.

# USAF Spouse Survey: The Final Chapter

Major Mickey R. Dansby
Captain Karl A. Ibsen
Leadership and Management Development Center
Maxwell AFB, Al 361'2-5712

The USAF Spouse Survey (AFSS) was completed by over 11,000 spouses of Air Force military and civilian personnel during the period 1982-1985. The survey was administered as a part of management consulting conducted by the Leadership and Management Development Center (LMDC). Short (1985) gives a brief history of the consulting process and describes the Organizational Assessment Package (OAP), the main consulting instrument with which the AFSS was linked. Because of its link with the OAP, the AFSS has proven to be a unique source of information on the relationship between work and family issues (Dansby, 1984; Dansby & Hightower, 1984; Flannery & Dansby, 1985). Ibsen and Austin (1983) give an overview of the purposes and history of the AFSS.

In late 1984, a decision was made to discontinue the management consulting program (by 1 Oct 86) because of manning constraints. Collection of data via the AFSS has already ceased. The purpose of the present paper is to summarize the final results of the AFSS for 1982-1985, with a breakdown of results by calendar year.

## Method

The AFSS consists of 73 attitudinal and demographic items. Responses to attitudinal items range across a seven point Likert scale with a "1" indicating strong disagreement or dissatisfaction and a "7" indicating strong agreement or satisfaction. Responses to the survey were collected at numerous CONUS and overseas bases during management consulting visits solicited by commanders of Air Force organizations (usually of wing size or equivalent). All military and civilian personnel present for duty were administered the OAP in group settings. Married personnel were given the AFSS to take home to their spouses. Approximately 35% of the spouses returned the completed surveys (in sealed envelopes) to a central collection point. Subsequently, the survey responses were added to the cumulative LMDC data base of AFSS responses.

Although the total data base includes data from 30 bases (19 CONUS and 11 overseas) in nine major commands [seven CONUS (71.8%) plus USAFE (25.4%) and PACAF (2.8%)], the sample was not selected to be representative of the Air Force. Furthermore, the distribution of commands sampled varies from year to year. Therefore, the reader is advised that generalizations to the Air Force as a whole, or comparisons across year groups for trends, must be approached with caution.

## Results

Table 1 provides a year-by-year and total summary for the demographic characteristics of the survey respondents. The table shows the total number of responses by year and gives demographic data as percentages of the valid responses to each item. The total number of responses may vary slightly for each item due to missing data.

Figure 1 shows the attitudinal items for the AFSS. Table 2 presents the item means, standard deviations and percent of respondents marking the item five or higher (%5+) for these items year by year and overall. For item 36 (satisfaction with the open mess) the results are given separately for spouses of officers and enlisted personnel. For items 44 and 45 (career intention), the scale has only six points; accordingly, the percentage marking four or higher is reported. Items 62 and 63 (TDY frequency and length) are primarily demographic in nature. No means or standard deviations are reported for these items (however, the table shows the %5+).

## Discussion

As was mentioned previously, the AFSS was not given to a true probability sample each year, nor were the commands stratified across years. Despite these limitations, the results show a remarkable consistency from year to year This fact, plus the broad representation of different commands and bases in the total sample, leads us to speculate that the results may be generalizable to the total population of Air Force spouses, at least insofar as broad attitudes toward the Air Force life are concerned. At the very least the data are representative of the bases at which they were collected, and the results reflect the attitudes of a significant portion of Air Force spouses.

Previous factor analyses (Dansby, 1984) resulted in the extraction of 14 factors with eigenvalues greater than one. These factors proved relatively stable over time (Dansby, 1984). We will discuss the total results of the AFSS in relation to some of the more interesting factors.

Factor 1 is a broad factor representing the spouse's identification with the Air Force. It includes items 16, 17, 19, 22, 27, 44, 45, and 71 (negative loading). Looking at the overall results for these items, we see a strong identification with the Air Force. About three quarters of the spouses are glad the member chose the Air Force as a career (item 27); a similar proportion want the military member to make the Air Force a career, if they have not done so already (item 44). Better than 60% would recommend the Air Force career (item 17) and feel involved with the Air Force lifestyle (item 16). Only about a third want the

Table 1

Demographic Summary of AFBS Data Base

| Characteristic | Level | 1982 n=3039 | 1983 n=1896 | 1984 n=3824 | 1985 n=2301 | Total n=11060 |
|---|---|---|---|---|---|---|
| Sex | Male | 8.7% | 11.2 | 9.8 | 9.3 | 9.6 |
| | Female | 91.3 | 88.8 | 90.2 | 90.7 | 90.4 |
| Age | 17-20 | 5.7 | 5.4 | 5.2 | 3.3 | 4.9 |
| | 21-25 | 24.0 | 19.0 | 23.4 | 18.4 | 21.4 |
| | 26-29 | 21.4 | 18.8 | 20.2 | 18.9 | 20.0 |
| | 30-34 | 24.0 | 20.4 | 19.8 | 24.4 | 22.0 |
| | 35-39 | 17.1 | 16.1 | 7.0 | 20.0 | 17.6 |
| | 40-49 | 9.4 | 15.8 | 10.0 | 12.8 | 11.4 |
| | 50 or more | 2.4 | 4.5 | 2.4 | 2.2 | 2.7 |
| Years Married | Less than 1 | 7.8 | 7.2 | 8.9 | 5.6 | 7.6 |
| | 1 but less than 4 | 22.7 | 22.3 | 26.4 | 21.6 | 23.7 |
| | 4 but less than 8 | 18.9 | 19.3 | 20.9 | 2?.7 | 20.5 |
| | 8 but less than 12 | 20.5 | 15.4 | 15.1 | 16.3 | 16.9 |
| | 12 but less than 16 | 14.9 | 13.4 | 14.0 | 17.0 | 14.8 |
| | 16 but less than 20 | 8.5 | 9.6 | 7.8 | 10.0 | 8.7 |
| | 20 or more | 6.6 | 12.8 | 6.8 | 6.8 | 7.8 |
| Ethnic group | Indian-Alaskan | 1.2 | .6 | 1.1 | 1.0 | 1.0 |
| | Asian-Pacific | 4.1 | 2.1 | 6.8 | 7.6 | 5.4 |
| | Black | 6.3 | 5.5 | 6.6 | 7.9 | 6.6 |
| | Hispanic | 2.5 | 1.8 | 4.0 | 7.7 | 4.0 |
| | White | 83.5 | 88.1 | 79.6 | 73.3 | 80.8 |
| | Other | 2.5 | 1.9 | 1.9 | 2.4 | 2.2 |
| Personnel Category (respondent's spouse) | Officer | 17.9 | 23.6 | 23.6 | 24.5 | 23.0 |
| | Enlisted | 61.4 | 60.3 | 70.2 | 69.0 | 68.7 |
| | Civilian | 7.4 | 16.1 | 6.2 | 6.5 | 8.3 |
| Years with the AF (respondent & spouse) | 2 or less | 5.9 | 6.7 | 8.4 | 5.3 | 6.8 |
| | 2-4 | 12.1 | 12.8 | 12.6 | 9.0 | 11.7 |
| | 4-8 | 17.6 | 18.8 | 23.6 | 18.9 | 20.2 |
| | 8-12 | 20.4 | 17.1 | 15.4 | 18.4 | 17.7 |
| | More than 12 | 44.0 | 44.7 | 40.1 | 48.4 | 43.7 |
| Time on Present Base | 6 months or less | 12.2 | 12.2 | 14.5 | 10.8 | 12.7 |
| | 6-18 months | 28.6 | 27.5 | 31.0 | 30.0 | 29.9 |
| | 18-36 months | 36.0 | 29.4 | 30.9 | 38.2 | 33.6 |
| | More than 36 months | 23.2 | 30.8 | 22.6 | 21.0 | 23.8 |
| Housing Information | On base with member | 36.7 | 37.8 | 43.5 | 44.4 | 40.9 |
| | On another base | 6.2 | 2.9 | 3.2 | 1.8 | 3.6 |
| | Off base renting | 33.9 | 24.7 | 25.9 | 32.9 | 29.4 |
| | Off base buying | 23.2 | 34.6 | 27.4 | 20.9 | 26.1 |
| Why Living On Base? | Better schools | 2.9 | 2.4 | 2.7 | 3.9 | 3.0 |
| | Off base expensive | 24.2 | 25.8 | 27.2 | 20.9 | 25.6 |
| | Off base unavailable | 2.6 | 1.1 | 1.0 | 1.2 | 1.5 |

| Characteristic | Level | 1982 n=30,9 | 1983 n=1896 | 1984 n=3824 | 1985 n=2301 | Total n=11060 |
|---|---|---|---|---|---|---|
| Why Living Off Base? | Better schools | 1.1 | .8 | .8 | 1 | .8 |
| | No base housing | 15.6 | 7.3 | 6.6 | 8.0 | 9.5 |
| | Investment | 13.1 | 16.0 | 17.3 | 13.5 | 15.2 |
| | Not eligible/on base | 6.5 | 11.1 | 8.7 | 7.5 | 8.3 |
| | Base housing bad | 12.4 | 13.5 | 13.3 | 16.4 | 13.7 |
| | Other | 12.9 | 13.8 | 10.6 | 13.5 | 12.4 |
| Education Level | Non-high school grad | 7.3 | 5.6 | 7.2 | 6.7 | 6.8 |
| | High school grad | 41.2 | 35.1 | 35.5 | 34.7 | 36.8 |
| | 1 yrs col or less | 21.9 | 21.9 | 24.6 | 24.7 | 23.4 |
| | More than 2 yrs col | 14.2 | 17.1 | 16.9 | 17.2 | 16.3 |
| | College graduate | 12.2 | 15.6 | 12.7 | 12.5 | 13.0 |
| | Masters or doctorate | 3.2 | 4.7 | 3.3 | 4.2 | 3.7 |
| Number of children at home | None | 26.8 | 27.3 | 28.4 | 26.0 | 26.7 |
| | 1 | 24.2 | 25.1 | 26.0 | 24.0 | 24.9 |
| | 2 | 35.4 | 32.1 | 31.5 | 34.4 | 33.3 |
| | 3 | 12.5 | 11.3 | 10.9 | 12.0 | 11.7 |
| | 4-5 | 3.0 | 3.9 | 2.9 | 3.2 | 3.1 |
| | 6 or more | 1 | 3 | 3 | 4 | 3 |
| Employment | Not Working | | | | | |
| | (don't desire) | 31.4 | 30.9 | 27.7 | 26.2 | 29.0 |
| | (not available) | 19.4 | 23.1 | 19.5 | 24.2 | 21.1 |
| | Part time work | 14.0 | 14.8 | 16.8 | 17.3 | 15.8 |
| | Active duty mil. | 10.5 | 10.0 | 9.4 | 10.9 | 10.1 |
| | Civil service | 7.0 | 6.5 | 5.3 | 6.1 | 6.2 |
| | Other | 17.4 | 14.6 | 21.2 | 15.3 | 17.8 |
| Usual Work Schedule | Not employed | 50.8 | 54.1 | 47.2 | 50.2 | 50.0 |
| | Day shift | 35.3 | 33.7 | 37.4 | 36.1 | 35.9 |
| | Other shifts | 7.1 | 5.8 | 8.4 | 6.5 | 7.2 |
| | Onstable hours/ on call | 6.7 | 6.4 | 7.0 | 7.2 | 6.9 |
| Reason for Working | Financial need | 20.3 | 17.4 | 21.4 | 17.4 | 19.6 |
| | "Extra" money | 10.4 | 9.6 | 11.5 | 11.0 | 10.8 |
| | Personal/profess growth | 16.1 | 17.0 | 17.8 | 18.0 | 17.2 |
| | Other | 2.6 | 2.3 | 2.4 | 3.3 | 2.7 |
| Student Status | Not a student | 87.0 | 85.0 | 84.9 | 86.3 | 85.8 |
| | Full time undergrad | 2.1 | 2.9 | 2.7 | 2.2 | 2.5 |
| | Part time undergrad | 6.7 | 7.8 | 8.3 | 7.4 | 7.6 |
| | Full time grad | .5 | .6 | .5 | .5 | .5 |
| | Part time grad | 1.9 | 2.0 | 1.5 | 1.6 | 1.7 |
| | Other | 1.8 | 1.7 | 2.0 | 2.1 | 1.9 |
| Volunteer Work | Non-volunteer | 75.4 | 70.6 | 71.1 | 68.0 | 71.6 |
| | On base vol | 12.1 | 11.5 | 12.6 | 16.5 | 13.1 |
| | Off base vol. | 7.7 | 12.0 | 10.1 | 8.6 | 9.5 |
| | Both on and off base | 4.8 | 5.8 | 6.2 | 6.8 | 5.9 |

Please indicate your agreement by choosing the phrase which best represents your attitude concerning the following statements

1 = Strongly disagree
2 = Moderately disagree
3 = Slightly disagree
4 = Neither agree nor disagree
5 = Slightly agree
6 = Moderately agree
7 = Strongly agree

16 I feel involved with the Air Force lifestyle

17. I would recommend an Air Force career for any young man or woman, including a son or daughter of mine

18 My participation in base or organizational activities is essential for my spouse to achieve his/her full promotion potential in the Air Force.

19 An Air Force career has as much prestige and status as a civilian career

20 I am interested in being informed and kept up-to-date on subjects related to the Air Force role and mission

21 It is important for me to know about the kind of work my spouse is doing

22. The Air Force has made considerable efforts to make service life more attractive for members and their families.

23. My spouse has to devote more time to "staying competitive" for promotion by means of service schools, college degrees, etc., than does his/her civilian counterpart

24. My spouse has been under a lot of pressure as a result of his/her Air Force job.

25 My spouse's abilities are fully used in his/her current job

26 My spouse has an important job

27 I am glad my spouse chose the Air Force as a career

28 My spouse feels positive about his/her contribution to the Air Force

29 My spouse has to devote more time to his/her job than his/her civilian counterpart

30. I would encourage my spouse to extend his/her military career if there were fewer moves

31 The effect of PCS moves on family life is an important factor in my spouse's career decision

32 Air Force leaders are sensitive to the needs of Air Force families

For the various services listed below, please indicate your level of satisfaction

1 = Extremely dissatisfied
2 = Moderately dissatisfied
3 = Slightly dissatisfied
4 = Neither satisfied nor dissatisfied
5 = Slightly satisfied
6 = Moderately satisfied
7 = Extremely satisfied

33. Base Exchange

34. Commissary

35. Military Medical Care

36. Open mess

37. Recreation center

38. Base library

39. Auto hobby shop

40 Bowling Center

41 Golf

42. Arts and Crafts

43. Child care

44. Which of the following best describes your desires for your spouse's career or employment intentions?

1. I would like my spouse to separate/terminate from the Air Force as soon as possible.
2. For the most part, I would like my spouse to not make the Air Force a career.
3. I am undecided as to my desires concerning my spouse making the Air Force a career
4 For the most part, I would like my spouse to make the Air Force a career
5 I would like my spouse to make the Air Force a career.
6 I would like my spouse to retire in the next 12 months.

45 Your spouse may have different career intentions than you would hope. Which of the following best describes your spouse's career or employment intentions?

1. Will separate/terminate from the Air Force as soon as possible
2. Will most likely not make the Air Force a career
3 May continue in/with the Air Force
4 Will most likely continue in/with the Air Force as a career
5 Will continue in/with the Air Force as a career
6 Planning to retire in the next 12 months

Figure 1 Attitudinal Items

Using the responses provided below, please indicate the extent to which you believe each of the factors listed is important in determining your spouse's career intention

1 = Not at all
2 = To a very little extent
3 = To a little extent
4 = To a moderate extent
5 = To a fairly large extent
6 = To a great extent
7 = To a very great extent

46 Job satisfaction
47. Status and prestige
48. Rate of pay
49. Medical/dental benefits
50. Job Security
51. Retirement
52. Patriotism
53. Other

Using the same set of responses, please indicate the extent to which each of the factors listed is important in how you feel about your spouse's career intention

54. Job satisfaction
55. Status and prestige
56. Rate of pay
57. Medical/dental benefits
58 Job security
59 Retirement
60. Patriotism
61 Other

TDY is defined as temporary military duty, and the maximum length of a TDY assignment is 179 days.

62. My spouse's job requires him/her to be TDY-

1. Less than once a year
2. Once or twice a year
3. 3 to 5 times a year
4. 6 to 9 times a year
5. 9 to 11 times a year
6. Once or twice a month
7. More than twice a month

63. How long does each TDY normally last?

1. Less than 3 days
2. More than 3 but less than 7 days
3. More than 7 but less than 14 days
4. More than 14 but less than 21 days
5. More than 21 but less than 30 days
6 More than 30 days
7. Duration varies widely

For the following items, select the most appropriate phrase and enter the corresponding number on the response sheet

1 = Not at all
2 = To a very little extent
3 = To a little extent
4 = To a moderate extent
5 = To a fairly large extent
6 = To a great extent
7 = To a very great extent

64. To what extent does the frequency of your spouse's TDY affect your family's life?

65. To what extent do the length of your spouse's TDYs affect your family's life?

66. To what extent do you believe TDY requirements influence your spouse's career intentions?

67. To what extent do the TDY requirements of your spouse's job influence your opinion of the desirability of the Air Force lifestyle?

Below are items which relate to your spouse's job. Read each statement carefully and then decide to what extent the statement is true of your spouse's job. Indicate the extent to which the statement is true for his/her job by choosing the most appropriate phrase.

1 = Not at all
2 = To a very little extent
3 = To a little extent
4 = To a moderate extent
5 = To a fairly large extent
6 = To a great extent
7 = To a very great extent

68. To what extent do your spouse's duty hours disrupt your family life?

69. To what extent is your attitude about your spouse's job an important consideration to him/her?

70. To what extent are you proud of your spouse's job?

71. To what extent would you be happier if your spouse was doing a similar job only as a civilian?

72 To what extent would you like your spouse to change the job he/she is now doing, but remain in the Air Force?

73. To what extent do you believe that the pay and allowances earned by your spouse are in proportion to the job he/she performs?

Figure 1 (cont.)

Table 7

Attitudinal Summary of AFSS Data Base

| State-ment No. | 1982 (n=3039) Mean | SD | %+ | 1983 (n=1896) Mean | SD | %+ | 1984 (n=3824) Mean | SD | %+ | 1985 (n=2060) Mean | SD | %+ | Total (n=10819) Mean | SD | %+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 4.65 | 1.91 | 61% | 4.71 | 1.90 | 62 | 4.81 | 1.87 | 66 | 4.88 | 1.89 | 65 | 4.76 | 1.88 | 63 |
| 17 | 4.54 | 1.96 | 56 | 4.86 | 1.88 | 63 | 4.92 | 1.87 | 65 | 4.77 | 1.93 | 61 | 4.77 | 1.91 | 61 |
| 18 | 2.97 | 1.99 | 24 | 3.20 | 2.03 | 29 | 3.51 | 2.07 | 34 | 3.44 | 2.09 | 32 | 3.29 | 2.05 | 30 |
| 19 | 4.33 | 2.14 | 52 | 4.63 | 2.11 | 56 | 4.73 | 2.09 | 60 | 4.47 | 2.15 | 54 | 4.55 | 2.12 | 56 |
| 20 | 5.56 | 1.58 | 78 | 5.70 | 1.44 | 81 | 5.66 | 1.53 | 80 | 5.58 | 1.59 | 78 | 5.62 | 1.54 | 79 |
| 21 | 6.16 | 1.31 | 89 | 6.25 | 1.20 | 92 | 6.25 | 1.22 | 91 | 6.10 | 1.34 | 89 | 6.18 | 1.27 | 90 |
| 22 | 4.25 | 1.94 | 55 | 4.37 | 1.92 | 58 | 4.44 | 1.91 | 59 | 4.41 | 1.95 | 58 | 4.36 | 1.92 | 58 |
| 23 | 4.79 | 1.79 | 56 | 4.81 | 1.79 | 56 | 5.03 | 1.76 | 61 | 5.05 | 1.79 | 61 | 4.92 | 1.78 | 59 |
| 24 | 5.28 | 1.78 | 73 | 5.21 | 1.78 | 73 | 5.35 | 1.74 | 75 | 5.42 | 1.73 | 75 | 5.32 | 1.75 | 74 |
| 25 | 4.25 | 2.28 | 52 | 4.32 | 2.27 | 53 | 4.40 | 2.24 | 55 | 4.38 | 2.26 | 54 | 4.33 | 2.25 | 53 |
| 26 | 6.00 | 1.44 | 86 | 6.04 | 1.41 | 87 | 6.09 | 1.40 | 88 | 6.14 | 1.40 | 88 | 6.06 | 1.41 | 87 |
| 27 | 5.27 | 1.79 | 68 | 5.52 | 1.67 | 74 | 5.58 | 1.67 | 76 | 5.46 | 1.75 | 73 | 5.45 | 1.72 | 73 |
| 28 | 5.49 | 1.73 | 76 | 5.76 | 1.58 | 82 | 5.74 | 1.60 | 82 | 5.69 | 1.66 | 81 | 5.66 | 1.65 | 80 |
| 29 | 5.49 | 1.74 | 72 | 5.35 | 1.77 | 67 | 5.53 | 1.69 | 73 | 5.66 | 1.67 | 73 | 5.51 | 1.71 | 71 |
| 30 | 4.65 | 1.97 | 50 | 4.64 | 1.93 | 48 | 4.73 | 1.96 | 51 | 4.61 | 2.02 | 47 | 4.66 | 1.96 | 50 |
| 31 | 5.02 | 1.89 | 62 | 5.00 | 1.89 | 61 | 5.07 | 1.86 | 62 | 5.06 | 1.95 | 63 | 5.04 | 1.88 | 62 |
| 32 | 3.39 | 1.90 | 35 | 3.61 | 1.92 | 38 | 3.62 | 1.90 | 39 | 3.54 | 2.00 | 37 | 3.52 | 1.91 | 37 |
| 33 | 3.83 | 1.90 | 45 | 4.11 | 1.85 | 51 | 4.24 | 1.84 | 54 | 4.01 | 1.92 | 48 | 4.05 | 1.87 | 50 |
| 34 | 4.51 | 1.87 | 60 | 5.05 | 1.70 | 70 | 4.77 | 1.86 | 66 | 4.62 | 1.86 | 61 | 4.70 | 1.84 | 64 |
| 35 | 4.26 | 2.03 | 53 | 4.33 | 2.03 | 54 | 4.11 | 2.06 | 50 | 4.25 | 2.09 | 52 | 4.20 | 2.04 | 52 |
| 36 (Off) | 4.05 | 1.66 | 42 | 4.19 | 1.71 | 42 | 3.64 | 1.73 | 31 | 3.87 | 1.73 | 37 | 3.89 | 1.72 | 37 |
| 36 (Enl) | 3.99 | 1.52 | 29 | 3.98 | 1.48 | 28 | 4.21 | 1.50 | 34 | 4.00 | 1.58 | 32 | 4.07 | 1.52 | 32 |
| 37 | 4.39 | 1.30 | 38 | 4.25 | 1.20 | 29 | 4.52 | 1.26 | 39 | 4.45 | 1.38 | 39 | 4.42 | 1.28 | 37 |
| 38 | 5.06 | 1.33 | 60 | 5.12 | 1.34 | 62 | 5.10 | 1.32 | 61 | 5.18 | 1.38 | 65 | 5.10 | 1.33 | 62 |
| 39 | 4.39 | 1.14 | 29 | 4.40 | 1.07 | 28 | 4.54 | 1.17 | 33 | 4.36 | 1.22 | 27 | 4.43 | 1.15 | 30 |
| 40 | 4.90 | 1.40 | 53 | 4.76 | 1.30 | 49 | 4.75 | 1.33 | 48 | 4.85 | 1.39 | 51 | 4.81 | 1.35 | 50 |
| 41 | 4.25 | 1.10 | 19 | 4.31 | 1.14 | 21 | 4.49 | 1.14 | 27 | 4.32 | 1.22 | 22 | 4.35 | 1.14 | 23 |
| 42 | 4.45 | 1.20 | 34 | 4.52 | 1.18 | 34 | 4.43 | 1.24 | 32 | 4.40 | 1.30 | 32 | 4.43 | 1.22 | 33 |
| 43 | 3.88 | 1.59 | 26 | 3.94 | 1.60 | 24 | 4.12 | 1.53 | 30 | 3.93 | 1.63 | 24 | 3.97 | 1.57 | 26 |

| State-ment No. | 1982 (n=3039) Mean | SD | %+ | 1983 (n=1896) Mean | SD | %+ | 1984 (n=3824) Mean | SD | %+ | 1985 (n=2301) Mean | SD | %+ | Total (n=11060) Mean | SD | %+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 44 | 3.97 | 1.26 | 70 (4+) | 4.08 | 1.20 | 73 (4+) | 4.08 | 1.21 | 73 (4+) | 4.13 | 1.37 | 73 (4+) | 4.04 | 1.23 | 72 (4+) |
| 45 | 4.08 | 1.23 | 72 (4+) | 4.13 | 1.17 | 75 (4+) | 4.12 | 1.21 | 74 (4+) | 4.21 | 1.30 | 76 (4+) | 4.12 | 1.21 | 74 (4+) |
| 46 | 5.75 | 1.49 | 81 | 5.85 | 1.37 | 84 | 5.78 | 1.41 | 82 | 5.76 | 1.52 | 81 | 5.77 | 1.44 | 82 |
| 47 | 4.42 | 1.69 | 47 | 4.46 | 1.64 | 48 | 4.54 | 1.65 | 50 | 4.52 | 1.72 | 50 | 4.48 | 1.66 | 49 |
| 48 | 5.26 | 1.62 | 69 | 5.33 | 1.53 | 72 | 5.24 | 1.59 | 70 | 5.20 | 1.67 | 67 | 5.24 | 1.60 | 70 |
| 49 | 5.01 | 1.64 | 64 | 4.91 | 1.69 | 62 | 5.12 | 1.62 | 66 | 5.10 | 1.71 | 65 | 5.04 | 1.65 | 65 |
| 50 | 5.58 | 1.45 | 79 | 5.72 | 1.40 | 81 | 5.66 | 1.43 | 81 | 5.62 | 1.51 | 78 | 5.63 | 1.44 | 80 |
| 51 | 5.66 | 1.49 | 80 | 5.69 | 1.47 | 80 | 5.71 | 1.48 | 80 | 5.76 | 1.50 | 80 | 5.70 | 1.48 | 80 |
| 52 | 4.97 | 1.68 | 61 | 5.08 | 1.62 | 63 | 5.29 | 1.60 | 70 | 5.30 | 1.63 | 68 | 5.16 | 1.63 | 66 |
| 53 | 4.22 | 1.96 | 42 | 4.29 | 1.91 | 43 | 4.14 | 1.97 | 46 | 4.60 | 2.09 | 46 | 4.31 | 1.95 | 44 |
| 54 | 5.81 | 1.52 | 82 | 5.93 | 1.42 | 84 | 5.87 | 1.46 | 83 | 5.85 | 1.53 | 82 | 5.85 | 1.48 | 83 |
| 55 | 4.47 | 1.78 | 50 | 4.55 | 1.76 | 51 | 4.62 | 1.72 | 53 | 4.60 | 1.78 | 52 | 4.55 | 1.74 | 52 |
| 56 | 5.39 | 1.64 | 73 | 5.47 | 1.53 | 75 | 5.44 | 1.55 | 74 | 5.43 | 1.64 | 73 | 5.42 | 1.58 | 74 |
| 57 | 5.25 | 1.71 | 70 | 5.20 | 1.74 | 69 | 5.39 | 1.64 | 73 | 5.39 | 1.71 | 73 | 5.31 | 1.69 | 71 |
| 58 | 5.73 | 1.44 | 82 | 5.80 | 1.35 | 84 | 5.81 | 1.41 | 83 | 5.75 | 1.49 | 81 | 5.78 | 1.42 | 82 |
| 59 | 5.66 | 1.52 | 79 | 5.68 | 1.50 | 80 | 5.74 | 1.49 | 81 | 5.79 | 1.53 | 80 | 5.71 | 1.50 | 80 |
| 60 | 4.74 | 1.80 | 55 | 4.84 | 1.75 | 57 | 5.05 | 1.73 | 63 | 5.08 | 1.78 | 62 | 4.93 | 1.76 | 60 |
| 61 | 4.10 | 2.02 | 40 | 4.19 | 1.97 | 41 | 4.25 | 2.03 | 44 | 4.50 | 2.21 | 43 | 4.20 | 2.01 | 42 |
| 62 | - | - | 6 | - | - | 12 | - | - | 7 | - | - | 11 | - | - | 8 |
| 63 | - | - | 40 | - | - | 24 | - | - | 34 | - | - | 36 | - | - | 34 |
| 64 | 3.27 | 2.05 | 27 | 3.15 | 1.95 | 23 | 3.27 | 1.97 | 44 | 3.49 | 2.08 | 27 | 3.28 | 1.99 | 26 |
| 65 | 3.64 | 2.13 | 34 | 3.49 | 2.04 | 31 | 3.64 | 2.06 | 33 | 3.86 | 2.11 | 36 | 3.65 | 2.08 | 34 |
| 66 | 3.30 | 2.06 | 30 | 3.21 | 1.98 | 27 | 3.41 | 2.04 | 31 | 3.45 | 2.09 | 29 | 3.34 | 2.03 | 29 |
| 67 | 3.51 | 2.18 | 32 | 3.41 | 2.08 | 30 | 3.53 | 2.09 | 32 | 3.72 | 2.16 | 34 | 3.53 | 2.11 | 32 |
| 68 | 3.33 | 1.97 | 27 | 3.05 | 1.93 | 22 | 3.23 | 1.93 | 25 | 3.47 | 2.12 | 29 | 3.26 | 1.97 | 26 |
| 69 | 5.13 | 1.58 | 66 | 5.15 | 1.59 | 68 | 5.24 | 1.57 | 71 | 5.26 | 1.61 | 69 | 5.19 | 1.58 | 69 |
| 70 | 5.69 | 1.43 | 78 | 5.76 | 1.43 | 81 | 5.82 | 1.40 | 82 | 5.81 | 1.47 | 82 | 5.74 | 1.44 | 81 |
| 71 | 3.85 | 2.12 | 37 | 3.56 | 2.12 | 32 | 3.74 | 2.06 | 34 | 4.06 | 2.18 | 38 | 3.79 | 2.09 | 36 |
| 72 | 3.17 | 2.22 | 28 | 3.09 | 2.18 | 27 | 3.13 | 2.18 | 27 | 3.37 | 2.31 | 29 | 3.16 | 2.20 | 28 |
| 73 | 3.19 | 1.60 | 18 | 3.23 | 1.59 | 20 | 3.26 | 1.60 | 19 | 3.31 | 1.87 | 18 | 3.21 | 1.60 | 19 |

military member to leave the Air Force and pursue a similar career as a civilian (item 71). Almost 60% perceive the Air Force as going to considerable effort to make service life more attractive (item 22). These results show fairly strong spousal support for the Air Force.

The second factor represents job benefits as a retention influence. Items 49, 50, 51, 57, 58, and 59 load with this factor. Since items 49, 50, and 51 are highly correlated with items 57, 58, and 59 (respectively), we will discuss only the first set of items. Almost two thirds of the spouses see medical benefits as a determinant of career intention to at least a "fairly large extent." About 80% of the spouses indicate job security and the retirement program are strong determinants of career intention.

TDY attitudes (items 64-67; Factor 3) show a relatively small impact on family life and career intentions. Less than a third of the spouses say TDY length or frequency influences their family life (items 64 and 65), or the member's career intentions (item 66) and opinion of the desirability of Air Force life (item 67), to a "fairly large extent" or more. On the other hand, results for Factor 4 (satisfaction/prestige as a career influence) show the spouses believe job satisfaction (item 46), status and prestige (item 47), and pay (item 48) are important career determinants. Over 80% rate satisfaction (item 46) as important to a "fairly large extent" or more; the corresponding proportions for status (item 47) and pay (item 48) are 50% and 70%.

Items (37-42), loading on the fifth factor, measure satisfaction with recreation facilities. This factor will not be discussed, since satisfaction varies from base to base and depending on frequency of use. Factor 6 (identification with job) is interesting. Most of the spouses (over 80%) are proud of the member's job (item 70) to at least a "fairly large extent"; few (less than 30%) want the member to change jobs within the Air Force (item 72). A resounding majority (nearly 90%) agree that the member's job is important (item 26); however, only a little better than half agree that the member's abilities are being fully used on the job (item 25). Most spouses (about 80%) agree that the member feels positive about his or her contribution to the Air Force (item 28).

Factor 7 reflects attitudes on basic services (items 33-36 and 43). For this factor, we note that the commissary (item 34) receives the highest ratings of satisfaction, followed by the base exchange and medical care, and finally the open mess and child care services. Pressure from the job (Factor 8) appears to be a significant concern for the spouses. About three fourths agree that the member has been under a lot of pressure as a result of his or her Air Force job (item 24). While only a fourth of the spouses respond that the member's job disrupts family life to at least a "fairly large extent" (item 68), over 70% agree that the member devotes more time to the job than his or her civilian counterpart (item 28). About 60% agree that the member must devote more time to "staying competitive" than a civilian counterpart (item 23).

Factor 9 (other influences on career decisions) yields little meaningful information. Factor 10 (patriotism as a career influence; items 52 and 60) is rated high by the spouses. About two thirds believe patriotism influences the member's career intention to a "fairly large extent" or more. The eleventh factor (desire for information about the member's job; items 20 and 21) has importance for Air Force commanders and supervisors. Over 80% of the spouses agree that they are interested in being kept up to date on the Air Force role and mission (item 20), and over 90% want to know about the work the member is doing (item 21).

Factors 12-14 will not be discussed (they are addressed in Dansby, 1984). Two items that have considerable importance for Air Force leaders should be noted, however. Over 60% agree that PCS moves have an important effect on family life and retention (item 31). Finally, less than 40% of the spouses agree that Air Force leaders are sensitive to family needs (item 32).

The results of the AFSS present some challenges for Air Force leaders. Clearly, the spouses like and support the Air Force; but just as clearly, they expect the Air Force to be responsive to the needs of the family. Open communication between families and Air Force officials needs to be maintained so leaders can be responsive. Further research with the AFSS, or a similar instrument, could serve to keep Air Force leaders in touch with the "pulse" of Air Force family life. Response to family needs may be critical in retaining a dedicated, experienced, quality force.

## References


Dansby, M. R. (1984). A proposal for the revision of the U.S. Air Force Spouse Survey (Report No. 84-0645) Maxwell AFB, Al Air Command and Staff College.

Dansby, M. R. & Hightower, J M. (1984). Family and work in the Air Force. Proceedings, Psychology in the Department of Defense, Ninth Symposium, 455-459.

Flannery, P. A. & Dansby, M. R. (1985). USAF Family Survey: A revision of the USAF Spouse Survey (Report No. LMDC-TR-85-3). Maxwell AFB, Al: Leadership and Management Development Center.

Ibsen, K A. & Austin, J S. (1983). Initial development of the USAF Spouse Survey. Proceedings of the 25th Annual Conference of the Military Testing Association, 633-638.

Snort, L. O. (1985). The United States Air Force Organizational Assessment Package (Report No. LMDC-TR-85-2). Maxwell AFB, Al. Leadership and Management Development Center.

# PREDICTORS OF PROPENSITY FOR CONTINUING EDUCATION AMONG ARMY CHAPLAINS

J. ERIC PIERCE, Ph.D., Chapel of the Four Chaplains
LAWRENCE A. GOLDMAN, Ph.D., USA Soldier Support Center-NCR

Background. Army chaplains, upon induction, have taken a minimum of 90 semester hours of academic work beyond the college level. As it requires roughly three years of full-time post graduate study to accumulate these 90 semester hours, one might expect chaplains to be individuals who are positively disposed toward continuing education. Chaplains, however, display varying amounts of positive attitudes toward continuing education. Personal experience, as well as responses to the instrument used in the present research, indicate that some chaplains are far more likely than others to seek or accept opportunities for continuing professional education. While there may be many contributing factors which would help to explain such a variety in attitudes, it would be most beneficial for the Chaplain Branch of the U.S. Army to be able to isolate some of the most sensitive predictors of propensity to pursue continuing education. For the purpose of this study, "continuing education" was defined in an Army-wide survey of chaplains conducted in 1984 "as a carefully guided reading program, a directed course of study at home or away from home, formal classwork, or seminary workshops of two or more weeks duration".

The Chaplain Branch administers a professional educational system of its own, as do other branches of the Army. This system includes the offering of courses at the U.S. Army Chaplain Center and School at Fort Monmouth, NJ, most of which are mandatory, as well as selection of certain officers to participate in The Army Educational Review Board program. Under this program, and because of special professional needs of the Branch and the services it is called upon to offer the Army community, the Army selects chaplains for special training at civilian institutions. A reliable set of predictors of propensity among chaplains to seek continuing education would facilitate this process, especially if future studies were to establish a positive correlation between prediction scores and successful academic performance. For the present study, we focused upon isolating the most reliable predictors from which such scores could be derived.

Major Variables and Models. The methodology used for this study consisted of deriving concepts from extant literature about continuing education among professionals and from pilot interviews with chaplains in the field and at the U.S. Army Chaplain Center and School. From these concepts, measurable variables were derived, and a questionnaire was constructed. The original study attempted to explore two research

questions: "What tasks are characteristic of the various duty positions of chaplains?", and "What is the nature of the relationship between the chaplaincy and the continuing education experience?". It was from exploration of the latter question that we discovered certain predictors of propensity toward involvement in continuing education among chaplains by isolating common factors among respondents who expressed a high degree of likelihood that they would seek further education.

Chaplains are clergy persons. The exercise of their profession involves a particular set of skills that can be taught and practiced under supervision. It entails an institutional setting through which service is rendered, and a formalized ethic of service. A minimum level of training for admission to the profession (seminary) is followed for many in the profession by continuing education and training to hone and preserve old skills, and to acquire new ones. The beliefs, attitudes and intentions of chaplains and other clergy are obviously among the behavioral determinants of seeking and participating in continuing education. Fishbein and Ajzen (1975) presented a theoretical framework for examining the relationships among these variables. They demonstrated that belief about an object influenced attitude and that, in turn, influenced behavior. Groteleuschen and Caulley (1977) applied the framework to continuing education as the type of behavior influenced. Consequently, they evolved a predictor model which, after some revision of its statistical formula, was useful in the present study. Bonn (1974) had devised previously a specific predictor model for participation by clergy in continuing education.

While income and previous educational experience are generally accepted as powerful contributing factors toward participation in continuing education, they were not expected to prove strong variables among chaplains for the following reasons:

1. Uniform minimum education requirement of 90 semester hours of graduate study for admission to the Army Chaplain Branch.

2. Competitive screening by most large denominations of chaplain candidates tending to promote admission of highly qualified persons.

3. Uniform pay for chaplains, by rank.

4. Government financing and mandating of some continuing education.

5. A professional value system favorably disposed toward continuing education.

Because, then, of a given degree of uniformity in so many areas, it was decided to concentrate upon beliefs, attitudes and intentions of chaplains regarding continuing education. To this end, the Groteleuschen and Caulley model was useful. The model is stated mathematically as:

$$B \sim I = w_1 (A) + w_2 (SSN) + w_3 (SPN), \text{ where}$$

B = The behavior in question,
I = Intention to perform behavior B,
A = Attitude toward performing behavior B,
SSN = Subjective Social Norm, and
SPN = Subjective Personal Norm, with $w_1$, $w_2$, and $w_3$ being empirically derived weights. Questionnaire items were provided in the model from which scores could be derived for the variables.

The Subjective Social Norm and Subjective Personal Norm were derived in the model from scores on pairs of questionnaire item indicators relating to belief and motivation. The impact of social interaction can be seen by Groteleuschen and Caulley's use of these two concepts as factors in the prediction of intention to participate in continuing education. The Subjective Social Norm relates to the actor's perception of the reaction of significant others to his participation. The Subjective Personal Norm is concerned with the actor's own beliefs about such participation. The Symbolic Interactionists would insist upon combining these two "norms" in any consideration of predictors of intention, and Groteleuschen and Caulley have done so by their formula.

Groteleuschen and Caulley had suggested that the weights in their formula be determined by running a multiple regression analysis for an entire group of respondents, using I as the criterion and A, SSN, and SPN as the three predictors. The standardized regression coefficient of each component would serve as its weight in utilizing the formula for determining I. To eliminate the problem of using I before its value is computed, we used partial correlation coefficients of each of the three components. The obtained coefficients were then used as the weights for the components.

Aside from the Bonn Predictor model and the Groteleuschen and Caulley model, we utilized scores from the following (each representing a composite of several indicators):

1) Stress and General Dissatisfaction Test
2) Perceived Barriers to Participation in Continuing Education
3) Felt Need for Continuing Education
4) Planned Participation in Continuing Education
5) Perceived Adequacy of the Respondents' Training

6) Willingness to Participate in Continuing Education
7) Status Inconsistency Test

Findings. There was a very strong relationship between present or recent participation and future participation in a deliberately planned program of continuing education experience. Ninety-five percent of chaplain respondents who were currently or recently (within 12 months) involved in a continuing educational experience expressed that they were "somewhat likely" or "very likely" to seek further education. Of those chaplains not currently or recently involved in such an experience, only 73 percent were either "somewhat likely" or "very likely" that they would be so involved in the future. Thus, past or present participation appeared to be an indicator of future participation.

A stepwise multiple regression analysis was conducted to isolate the best predictors of participation in continuing education. The seven attitudinal variables and the two predictor models comprised the independent measures. Responses ("Yes" or "No") to the question, "Are you presently participating or have you in the past 12 months participated in a deliberately planned program of continuing education?", constituted one dependent measure. Responses on a 5 point scale, ranging from "Not likely at all" to "Very likely" for the question, "In the next 5 years, how likely is it that you will participate in a deliberately planned program of continuing education?", comprised the other dependent measure. As shown in Table 1, the significant indicators of participation, in order of significance, were: The Planned Participation Test, the Groteleuschen and Caulley model, the Status Inconsistency Test, and the Perceived Barriers Test.

As shown in Table 2, the Planned Participation Test, the best independent measure of participation in current or recent continuing education, was also identified as the best predictor of perceived future participation in such a program. The simple $r$ of .80 between the Planned Participation Test scores and perceived future participation was highly significant. Felt Need for Continuing Education was also a statistically significant predictor of future educational propensity, but to a lower degree than the Planned Participation Test.

Table 3 shows the inter-correlations of the five independent measures identified as significant predictors of the two dependent measures.

TABLE 1 - SIGNIFICANT MEASURES OF CURRENT (OR RECENT) PARTICIPATION IN A PROGRAM OF CONTINUING EDUCATION

| Independent Measure | Multiple R | $R^2$ | $R^2$ Change | Simple r* |
|---|---|---|---|---|
| Planned Participation Test | .318 | .101 | .101 | -.32 |
| Groteleuschen and Caulley Predictor Model | .342 | .118 | .017 | -.22 |
| Status Inconsistency Test | .354 | .125 | .007 | -.05 |
| Barriers to Participation in Continuing Education | .363 | .132 | .007 | -.21 |

*Based on a dichotomous scale where "1" = yes and "2" = no

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

TABLE 2 - SIGNIFICANT MEASURES OF FUTURE PARTICIPATION IN A PROGRAM OF CONTINUING EDUCATION

| Independent Measure | Multiple R | $R^2$ | $R^2$ Change | Simple r |
|---|---|---|---|---|
| Planned Participation Test | .801 | .641 | .641 | .80 |
| Felt Need for Continuing Education | .817 | .668 | .027 | .13 |

* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

TABLE 3 - INTER-CORRELATION MATRIX OF SIGNIFICANT INDEPENDENT MEASURES OF CURRENT (OR RECENT) AND FUTURE PARTICIPATION IN A PROGRAM OF CONTINUING EDUCATION

| | PLANPART[1] | GROTCAUL[2] | STINCTEST[3] | BARRPART[4] | FELTNEED[5] |
|---|---|---|---|---|---|
| PLANPART[1] | 1.00 | .35 | -.07 | -.32 | .06 |
| GROTCAUL[2] | | 1.00 | -.10 | -.38 | .13 |
| STINCTEST[3] | | | 1.00 | .08 | .00* |
| BARRPART[4] | | | | 1.00 | -.03* |
| FELTNEED[5] | | | | | 1.00 |

1 Planned Participation Test

2 Groteleuschen & Caulley Predictor Model

3 Status Inconsistency Test

4 Barriers to Participation in Continuing Education

5 Felt Need for Continuing Education

* Not statistically significant from zero.

Conclusions. Those variables which proved to be most sensitive and useful as predictors were the five identified previously. The Groteleuschen and Caulley model, although living up to our expectations, was not quite as sensitive as the Planned Participation Test. Still, it should be retained for use in future studies. The Bonn Predictor model and three of the composite variables used in the present study (Willingness to Participate, the Stress and Dissatisfaction Test, and Perceived Adequacy of Training) proved to be less sensitve and therefore were not retained.

Useful future studies to isolate reliable predictors would include a longitudinal study to discover whether or not respondents who perceive a strong likelihood of participation in a continuing education program actually do participate within five years. Another useful study would be to track the academic performance of respondents in continuing education programs subsequent to administering the predictor tests. A high positive correlation between test scores and grades would reinforce the predictive value of the tests.

## REFERENCES


Bonn, Robert L. Continuing Education Participants: Who, How Many, Types of Programs. Richmond: SACEM, 1974

Fishbein, M. and Ajzen. Beliefs, Attitudes, Intention and Behavior. Reading, Mass: Addison - Wesley, 1975.

Groteleuschen, A. and Caulley, D. "A Model for Studying Determinants of Intention to Participate in Continuing Professional Education". Adult Education, 28 (1977), pp. 22-37.

# Composite Ratings as a Performance Criterion

Jonathan Warren
Consultant

Peter Newton
National Security Agency

John Bondaruk
National Security Agency

A continuing issue in studies of job performance is establishing acceptable, useful criteria. Whether the purpose of a study is to evaluate selection or promotion procedures, or training effectiveness, or to compare different classes of employees, the usefulness of the results depends heavily on the accuracy of the criteria. In a long history of attempts to improve performance criteria, behaviorally anchored rating scales and behavioral observation scales are probably the most popular current choices. Their use is accompanied by a concern for specifying and rating separately the various components of performance important in any given job. The separate ratings may or may not then be integrated into a composite global rating. The gains expected from being more precise in what we attempt to assess have been disappointing. We may do better by being more modest in our attempts to pin down a criterion with great precision while paying closer attention to some of the mundane aspects of human judgment. The rest of this paper describes such an effort.

In a study of selection procedures in a large government organization, a composite rating procedure for assessing job performance was devised with attention to the practicalities of recording accurate, thoughtful human judgments. First, since large numbers of ratings were required, the demands on the raters in time and effort should be reasonable. The process should be quick and easy to avoid rater fatigue and loss of interest. Second, the judgments asked for should be realistic, understandable (within the frame of reference of the employees rather than psychologists), clearly related to the job being assessed, and should refer to performance normally observable by the raters. Third, the ratings should permit the application of multiple perspectives. The same qualities are assessed, perhaps using the same form, but the judgments are made from different assumptions, or for different stated purposes.

The procedure used called for the raters to make three judgments about each of from six to twenty employees whose work they knew. All three judgments referred to general or global performance in a single job for which a detailed job description had recently been established and which was well known by the raters. An initial list was given each rater of up to 20

employees he or she had recently supervised. Before starting the rating process, each rater deleted the names of persons he or she felt unable to rate and added others to a maximum of 20. The average number of persons rated was about eight.

The instructions asked the raters to assume they were in a newly formed organization in charge of hiring a number of new employees for the job being rated. The persons hired were to be productive immediately without further training. The raters were instructed to identify the three persons on their list they would hire first. Then they indicated the best person among those three, and then the three next best from among the persons not yet selected.

As long as at least six names were on the list, this process selected the single most effective employee, two who were next best, and three who ranked just below them. The remaining names constituted a fourth group who trailed the others. Thus every person on the list was ranked in terms of overall productivity in the job, but without fine distinctions except at the top of the order. Avoiding those fine distinctions, which are often difficult to make and rarely informative, is one way the process was simplified and shortened without sacrificing good information.

For the second task, the raters were informed that they were to select ten persons for their new organization from a pool of about 100 applicants that included the persons on their list. They were to identify any of the listed employees the raters thought were so competent that they would hire them immediately without bothering to check the qualifications of the other applicants. This judgment, which is made fairly quickly, indicates whether any or all of the persons rated are unusually competent.

Finally, for their third task, the raters indicated on a five-point scale their absolute judgment of the overall capability of each listed employee. The five scale points were labeled from "Barely proficient" to "Among the best I've seen," with similar descriptions for the three intermediate points.

The use of three kinds of ratings kept some variability in them, balancing some of the advantages and disadvantages of relative and absolute judgments. The first step forced the raters to discriminate among the employees at four levels of relative capability. The instructions for the absolute rating scale informed the raters that they were not required to use the full range of the scale, that they could rate everyone at the top, the bottom, or anywhere else. Yet the discriminations they had just been forced to make in ranking the employees no doubt imposed some tendency to use more than a limited segment of the scale.

The variability of the absolute ratings was quite good, with

acceptable proportions of ratings in the lowest levels. On the five-point absolute scale, the percentages of ratings from lowest to highest were 3, 10, 28, 37, and 22. From lowest to highest among the four levels of relative ranks, the percentages were 54, 22, 16, and 8. The ranks thus discriminated best at the upper end of the distribution of performance, while the absolute ratings discriminated well throughout the range but best at the lower end. Slightly more than a third of the employees (37 percent) were judged capable enough to be hired immediately, providing a coarse but quick confirmatory indicator of the absolute level of performance.

About 700 supervisors judged a total of more than 5,000 employees who represented nine different jobs in five job families, with at least 240 persons in each job. Almost 2,000 employees were judged by at least two supervisors; almost 400 had been judged by four.

To estimate the reliability of the judgments, the employees were grouped according to whether they had been rated by two, three, or four supervisors. In a table with employees as rows and ratings as columns, the placement of ratings in the rows was arbitrary. Thus one of the ratings of someone who had been rated twice could appear in either the first or second column. A rating for someone rated three times could appear in any of the first three columns. No one was rated more than once by the same rater. The four columns therefore represented four independent sets of ratings.

The level of agreement between all six possible pairs of columns was calculated using gamma, a measure of agreement between two sets of ranks. Median gammas for the absolute ratings and the relative ranks were respectively .47 and .37. This procedure includes error attributable to the different judges rating each employee, which reduces the levels of interjudge agreement from those that would be reached if the same supervisors had judged all the employees, a condition not feasible with a large number of employees. For six different job families, the median coefficient between absolute ratings and relative ranks given by the same rater was .60.

With large numbers of employees judged by several raters and each rater providing three separate judgments, a number of composite ratings can be formed. A composite scale formed by combining the three ratings by a single supervisor showed an intraclass correlation coefficient of .75. The scale had a range of 2 to 12, a mean of 8.1, and a standard deviation of 1.8. Combining all three judgments by two sets of supervisors produced a scale with a greater range and about the same reliability, .77. Other composites produced reliabilities of .70 to .75, but all required judgments by more than one supervisor. The three different judgments by the same supervisor matched the reliability of the best scale formed by the judgments of more than one rater.

The criteria were used in a validity check of a battery of selection tests for employment. The predictors were tests of cognitive skills that formed five factors--verbal ability, reasoning, clerical ability, spatial perception, and knowledge of science. For many of the employees, the tests had been taken a number of years before the judgments of job performance were made. Within the different job categories, multiple correlation coefficients between optimal combinations of predictors and multiple criteria ranged from about .20 to .45 with a median of .28.

The validity coefficients for composite criteria formed from the judgments of two, three, and four supervisors were .24, .26, and .34 respectively. The numbers of cases in the groups of two, three, and four supervisors were respectively about 2,000, 800, and 400. As expected, combining the judgments of more than one supervisor apparently increases the validity of the criterion, although the gain may not be appreciable.

The best composite scale in terms of reliability and validity was the six-element composite formed from three judgments by each of two judges. Virtually on its heels, though, with respect to both reliability and validity, was the single-judge scale formed from the three ratings. The probable tendency of judges to be moe careful in their absolute ratings after having been forced to rank the persons being rated, the greater sensitivity of the ranks at the upper end of the scale and of the ratings at the lower end of the scale, and the probable corrective influence of the binary decision to hire a person immediately all may contribute to the usefulness of the single-rater composite scale. Its ease of administration is another important element. All three ratings of as many as 20 persons could be made in about 15 minutes. The different perspectives the three judgments require may provide a useful, global criterion of job perfo-mance.

# MAXIMIZING CRITERION VARIANCE IN VALIDATION RESEARCH: IS THIS ALWAYS BEST?

Captain Robert J. Angus
and
Major Reginald T. Ellis

Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario, Canada

The Canadian Forces Personnel Applied Research Unit (CFPARU) is responsible for conducting a continuing research program to maintain a reliable and valid set of selection tests for use in the Canadian Forces (CF). The selection test battery currently in use is the Canadian Forces Classification Battery (CFCB), which was implemented operationally in 1981. The technical training courses, for which this battery is used to select students, are continually changing; consequently the CFCB must be closely monitored to ensure that the best students continue to be selected to fill recruiting vacancies.

## Criterion Data Collection System

Initial CFCB validation studies were conducted using training performance criterion data collected specifically for that project. Recognized limitations in the quality, consistency and comprehensiveness of these one-shot data led to a decision to establish an on-going data collection system, which would start to generate data at the same time as the operational implementation of the new battery. This Criterion Data (CD) collection system was set up to support periodic studies to revalidate CFCB standards against basic trades training performance for over 60 entry level occupations.

The CD collection system provided for the acquisition of standard format training performance measures on every graduate from CF basic trades training courses (Ellis & Saudino, 1980; McMenemy, Amyot, & Enkurs, 1984). Implicit in the design of this system were two assumptions. The first was that maximizing variance in the criterion measures would yield better estimates of validity. The second was that the capacity to identify, and exclude from the analysis, those trainees who were unsuccessful for reasons unrelated to ability (i.e., motivational failures) would similarly yield better validity estimates. Thus, steps were taken to supplement readily available student computer records consisting of pass/fail information, with more precise performance

---

The views and opinions expressed in this paper are those of the authors and not necessarily those of the Department of National Defence.

measures for both the passing group (which almost invariably represents more than 90% of the sample) and for the non-successful group. Supplementary information for the passing group consisted of performance measures such as class standing or course grade. For the non-successful group, supplementary information was aimed at providing a means of discriminating between failures related to lack of ability and those related to lack of motivation.

The latter requirement was addressed by devising a system of disposition and reason-for-disposition codes for use in conjunction with automated individual training records. This required training authorities to denote, within the automated training record, the disposition action ("release" from the CF, "reassignment" to a different trade, or "recourse" to a later course serial), and the reason for non-success (lack of ability, lack of motivation, medical problems, etc.) for each non-passing trainee.

## Validation Studies

Based on this data collection system, a standardized approach to constructing data sets and conducting validity analyses was established to cycle through revalidation studies on various trades as the need to do so was identified and/or as sufficient criterion data accumulated (Miller & Angus, 1985). The validation model involves the identification of the best CFCB single or composite predictor on the basis of a comparative correlational analysis. Included in the model is a procedure for selecting an appropriate cut-off score on the basis of a standard contingency table reflecting success rates associated with an array of selection ratios.

During the first of a series of revalidation studies based on the new model (Miller & Angus, 1984), it was determined that the best CFCB predictor for the trade under study yielded a validity coefficient of .31 (corrected for restriction of range). In constructing a selection ratio versus pass-rate contingency table, it was expected that as the selection cut-off score increased, the "efficiency" of the selection process would also increase (i.e., the percentage of students above the cut-off who passed the course would rise in a similar fashion). However, this pattern was not observed in the contingency table for this predictive validity study. In fact, it was observed that the success rate (selection efficiency) remained constant irrespective of the selection ratio used. Further investigation of data from the 18 students who were unsuccessful during this training revealed that nine of these trainees had achieved predictor scores above the 40th percentile. Having already identified the most valid predictor of training performance for this particular trade, it was inconsistent that students with predictor test scores in the upper part of the distribution should perform so poorly during training, if their training problems were truly ability-related. In spite of the fact that those nine students had much higher selection test scores than was necessary to meet the minimum selection standard for this course (a percentile score of 10), they failed to successfully complete the training and were eventually classified as ability-related failures.

Student "Reason for Failure" Misclassification

It is apparent that some or all of these nine students may not have failed for ability-related reasons. Rather, it seems more likely that they were incorrectly classified as ability-related failures, and should have been assigned to a motivation-related failure category. If they had been classified as motivation-related failures, their scores would not have been included in these analyses since only those students who are successful, or are unsuccessful for reasons attributable to ability, are to be included in the analyses. To examine this possible misclassification issue, the total sample of trainees was rank-ordered by predictor test scores, and those unsuccessful trainees falling above the median for this whole group were identified as "potentially misclassified". Further analyses were then conducted to determine the validity coefficients for the following sub-groups of students:

a. all those who had passed the course;

b. all those who had passed the course less the nine potentially misclassified students; and,

c. the group of 18 students who failed the course.

It was hypothesized that if misclassification had occurred, then the following results could be expected from these analyses:

a. the validity coefficient for the group who passed the course should be greater than the validity coefficient for the full group of trainees;

b. the validity coefficient for the full group less the potentially misclassified failures should be even greater than that of the passing group alone;

c. the validity coefficient for the group of unsuccessful trainees should be very low since half of the group are students that are potentially misclassified.

The results of these analyses are presented in Table 1. Examination of this table shows that the validity coefficients are completely consistent with the hypothesized relationships. Inclusion of data from students who were unsuccessful on the course for ostensibly ability-related reasons, instead of contributing positively to the precision of the validity analysis, was detrimental to its overall accuracy. This fact is demonstrated by an increase in the validity coefficient from .31 to .37 as a direct result of the exclusion of records pertaining to failed trainees, and an additional increase to .39 when only the potentially misclassified (failed) trainees are excluded. Further, the validity coefficient for the unsuccessful group of students shows that not only was the correlation between their scores and performance on basic trades training not significiantly different from zero, but the correlation was negative, or in the OPPOSITE direction to what would be expected. In other words, the inclusion of scores for misclassified failed students erroneously and significantly decreasing the validity coefficient.

Table 1

An Examination of the Change in
Validity Coefficients as a Result of Separating
Successful and Unsuccessful Trainees

| Group | n | Corrected Validity Coefficient |
|---|---|---|
| All trainees | 195 | r = .31 |
| Successful trainees only | 177 | r = .37 |
| All trainees less the misclassified group | 186 | r = .39 |
| Unsuccessful trainees only | 18 | r = -.14* |

Note: All correlations are significant (at the .01 level) except the one designated by an asterisk.

Reasons for Misclassification

Misclassification in this situation is understandable if one considers some of the circumstances specific to the CF trades training system. The CF operates on a one person/one job basis, with very strict and finite limits imposed on the numbers of personnel permitted within each rank level in each trade. A recruit is therefore assigned to a single, specific trade at the point of enrolment, and only into a trade for which a specified training vacancy is open at that time. Recruits often don't get their "first choice" of trade, but may agree to enrolment in a less preferred trade in order to gain initial entry to the CF.

However, the opportunities for post-enrolment reassignment to another trade are relatively limited. If a trainee discovers, once trades training has commenced, that the originally assigned trade is not a vocation in which s/he can be happy or satisfied, there are few options open in terms of changing trades. An outright request for reassignment to another trade is in most cases not actioned, and until recently often resulted in the trainee being released if s/he was adamant about leaving the originally assigned trade. This was intended to prevent students from requesting reassignment on a "whim" if the training in their present trade was at a particularly demanding or stressful point. Early release from the CF has not, in recent times, been an attractive option with young trainees, who face the subsequent prospect of high civilian unemployment rates.

The CF does try to accommodate those trainees who do their best and work hard during trades training, but are unable (perhaps through a lack of aptitude) to successfully complete the course. These trainees are quite often recommended, by their course instructors, for reassignment to another CF trade where they might have a better chance of success. Consequently, it can be seen that trainees who appear to their instructors to have the ability to succeed, but are simply not positively motivated to

do so, are generally not considered good risks for retention in the CF in any trade; while hard working trainees who lack the ability to succeed in one trade are often recommended for reassignment to a different trade.

To the bright but vocationally dissatisfied young trainee, awareness of these informal policies can logically lead him or her to the conclusion that the route to reassignment involves convincing the instructor that, in spite of his or her best efforts, the course material was beyond comprehension. Hence, it is possible that the student may intentionally fail his or her initial basic trades training course, while purposefully avoiding giving the appearance of not being motivated, in order to be recommended for reassignment to another CF trade. Following from this, the course instructor would likely assign such a student to an ability-related failure category, despite the fact the student failed because of a lack of positive motivation to continue in his or her present trade, and not because of a lack of ability.

While plausible, this scenario may seem to represent a rather cynical view of the ethical standards of some of our more able enrolees. Alternative explanations which are equally plausible are available. For example, basic trades training instructional staff may be unable, or perhaps unwilling, to identify the reasons for failure as being motivationally related. This could arise from a sense of unease about passing judgement on the psychological or emotional make-up of their students, or perhaps from an enlightened view that it is often not the trainee's fault that the initial trade assignment decision was a poor one. This view would hold that willingness to accept a less preferred trade to gain entry is evidence of positive motivation towards military service in general, and that this argues for retaining the individual if possible. Cognizant of the impact that a "motivational failure" label may have on the likelihood of retention/reassignment, the instructor may deliberately, and for the best of reasons, misclassify the student as an ability-related failure.

The above discussion points to the need to incorporate a wider array of inputs into the trade assignment decision process. While aptitude is an important element, non-cognitive factors such as vocational interest, temperament, etc., play an important role in determining training success. Rather than treating these factors as "error", and removing them from analyses which are constrained by a narrow, aptitude-based focus, steps need to be taken to standardize and systematize non-cognitive inputs to the process so that matching the person to the job is based on a broader and more complete picture of the relevant factors.

It is suggested that the assumption that maximizing criterion variance will invariably yield better empirical validity estimates tends to be associated with this narrow, aptitude-based focus. The design of this validity study was, in fact, a direct result of acting upon this assumption, since it represents an attempt to maximize the number of records in the data set, while excluding criterion data which were "contaminated" by motivational influences. Improper identification of trainees who should be included in the data set can occur for a number of possible reasons. Whatever the reason underlying this phenomenon, it seems clear that such misclassifications can and do occur.

The measurement of non-cognitive variables which are relevant to the selection/decision-making process is not, in itself, a particularly difficult task. However, this class of measures is much more susceptible to influence by environmental factors and to faking by individual examinees than cognitive ability tests. This, in the same fashion as aptitude tests, renders the incorporation of these measures into the selection process highly problematic. The answer appears to lie in improving the quality of the counselling provided as part of the recruitment process. Organizational recruitment involves two related but separate decision processes, that of the organization, and that of the applicant. Cognitive tests clearly can and should play a major role in the organizational decision about accepting a given applicant. Counselling relates to the individual's decision, and it is here that motivational inputs are most relevant and most easily incorporated. The CF Career Information System (Ellis, 1983), the US Army (JOIN) System, and the US Navy Personnel Accessioning System (NPAS) are all attempts to systematize and incorporate both cognitive and non-cognitive elements into the total accessioning process. Until such systems are fully in place, and reliable data on motivational factors can be collected, the narrow, aptitude-based focus will likely continue to prevail. However, motivational issues will nevertheless continue to impact on training success.

Consequently, it is important to be aware of this potential source of data contamination within the widely accepted practice of attempting to maximize criterion variance when conducting predictive validity studies. Is maximizing criterion variance always best? Yes, but only if those records in the criterion data, that contain information on unsuccessful trainees, are free of misclassification errors.

## REFERENCES


Ellis, R.T. (1983, October). Improved recruit counselling with the Canadian Forces Career Information System. Paper presented at the 25th Annual Conference of the Military Testing Association, Gulf Shores, Alabama.

Ellis, R.T., & Saudino, D.A. (1980). Development of a selection model based on the experimental classification battery-men (Working Paper 80-13). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

McMenemy, J.P., Amyot, K.A., & Enkurs, I.A. (1984). Criterion data collection system: Policy and maintenance procedures (Working Paper 84-7). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Miller, D.L., & Angus, R.J. (1984). Predictive validity of selection tests for the Steward 862 trade (Technical Note 20/84). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Miller, D.L., & Angus, R.J. (1985). A revalidation model for the Canadian Forces Classification Lattery (Working Paper 85-2). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

# Literacy, Readability and Knowledge

Frederick R Chang

*Navy Personnel Research and Development Center*
*San Diego, California 92152*

The fields of reading and reading instruction have been highlighted by a number of "great debates " The most popular of the debates was the one made famous by Jeanne Chall (1967) concerning the relative merits of phonics vs whole-word decoding in reading instruction Another debate, which has occurred in the military, has concerned the issue of whether or not the armed services should provide basic skills instruction to those personnel who may need it (Sticht, 1982) The papers in this symposium reflect elements of two other distinctions that I feel are important to the issue of remedial education in the Navy The first issue concerns the distinction between general literacy vs specific literacy, and the second issue concerns the distinction between reading "skills" and the content knowledge that readers possess I will comment on these two issues briefly, and will describe some recent research that we have carried out that is germane to these issues

## General Literacy vs. Specific Literacy

The paper by Dr Moracco discussed the Navy's attempts to move from general literacy training to more specific literacy training, and indeed this is a distinction that is fundamental to this symposium The paper by Dr Idar discussed some of the policy implications of this change, and I would like to focus my discussion on some of the cognitive aspects In teaching "general literacy," one typically refers to the process of giving readers general reading skills that they can then transfer to the reading of materials in many different content domains The teacher of general literacy is seen as empowering the reader with a set of reading skills that will generalize to the efficient reading of textual materials in any subject In general the public school system teaches general literacy to its students In teaching "specific literacy," on the other hand, there is the recognition that people read texts for specific purposes and in specific content areas The specific literacy teacher attempts to teach the reader what he or she needs to know in order to perform a given task Since it is generally held that it takes knowledge to get knowledge, it therefore stands to reason that if you give instruction to a reader in a content area, it will be easier for that reader to learn new information themselves in the content area Most civilian adult basic education programs subscribe to some form of the specific (functional) literacy approach

The clash between the approaches comes in dealing with the issue of remediation Basic skills training in the military has had strong supporters on both sides of the issue (see Sticht, 1982 for a review) The detractors from the specific literacy approach claim that in teaching students about specific domains, the students will not acquire the necessary skills to read efficiently in all domains in which they will need proficiency The fear is that the students will be artificially "propped up" in a particular domain, only to fail, miserably, when asked to read in another domain The supporters of the specific literacy approach to remediation point out that general literacy training for children in the public schools and general literacy training for adults in the armed services are quite different Children in the public schools have many years to gain literacy skills, young adults in the military do not Given that one knows quite well the sorts of literacy tasks that young adults in the military are to perform (i e , reading for learning or reading for doing something), then it makes sense to provide basic skills instruction in the context of the specific literacy task After all, in general literacy students are sked to read in many different

content areas, why not have them reading in the domain in which they will be working?

The debate will not soon be settled, but it is interesting to ask the question What is learned in general literacy? The answer to this question is certainly most complex, and can be addressed at many different levels At the lexical level, though, it is certainly the case that one goal of general literacy is to provide readers with a large enough vocabulary so that they can read in many domains Nagy and Anderson (1984) have argued convincingly that it is simply not possible for people to obtain the vocabulary that they have through direct instruction in vocabulary There are simply too many words, it would take too long It seems reasonable to suppose then, that much vocabulary is acquired by reading in many different content areas (Nagy, Herman & Anderson, 1985)

How much vocabulary is needed to read in a variety of different content areas? An analysis of the vocabulary used in many different publications reveals that relatively few words are used most commonly. A variety of different word lists have been assembled that contain the majority of the vocabulary in common usage The different lists are naturally quite similar, and contain relatively few root words (a few thousand) It is generally found, that these words account for the large majority of general use in many popular documents It was of some interest to me to compare a general literacy text, with a specific literacy text to see what proportion of the words contained in each, appeared on a common word list That is, it is of interest to know if general literacy texts contain a vastly higher proportion of common vocabulary than specific literacy texts If so, then one might argue that students in reading specific literacy texts might not be exposed to and learn the common vocabulary needed for efficient reading in other domains For the general literacy text I chose the comprehension paragraphs from the Gates-MacGinitie reading test level D The Gates-MacGinitie test is one of the most widely used tests of general literacy in the United States and is the general literacy test administered to all Navy recruits For the specific literacy text, I chose some passages from the Experimental Functional Skills Program reading assessment battery The general and specific literacy texts contained roughly the same number of words For the common word list, I selected the 4908 most frequent words on the Kucera and Francis (1967) list (These words corresponded to a frequency of occurrence of 20 or greater )

The results were that 86% of the words on the general literacy text were on the common word list, and 90% of the words on the specific literacy text were on the common word list This is an interesting finding and suggests that in general most of the words from both texts are ones that are used most commonly The 14% that are not, from the general text, presumably derive from some particular content area (after all, all the paragraphs on the Gates-MacGinitie test are about some domain) Similarly, the 10% of the words from the specific text that were not on the common word list are presumably specialized Navy words An implication from this very informal finding is that, at least at the lexical level, perhaps much of what is thought to be learned only in general literacy is also learned in specific literacy

**Processing Skills vs. Knowledge**

Another distinction, closely related to the general versus specific literacy distinction, concerns processing skills used in reading versus the knowledge that the reader has of the domain to be read Research in experimental psychology has been dominated by a concern for the processing skills used in reading The concern has been both methodological (e g , Chang, 1983, Kieras and Just, 1984) and substantive (e g , Crowder, 1984, Gibson and Levin, 1975). One outcome of the emphasis on processing skills has been that some of the findings from the basic research laboratories have been interpreted fairly literally into training approaches with sometimes negative results One example of this outcome concerns the finding that good readers, in general, exhibit very different eye movement patterns from poor readers Good readers 1) make fewer eye movements per line, 2) have shorter fixation durations, and 3) make fewer regressive eye movements, than poor readers On the basis of these findings, several eye movement training programs were developed The thinking was that all that one needed to do was to train the poor readers to move their eyes in reading just like good readers, and the poor readers would then be reading just

like the good readers Naturally, such training failed miserably (see Gibson and Levin, 1975)

This example is not meant as an indictment against the study of basic reading processes – clearly such study is important However, the emphasis on the study of reading processes has lead in part to the development of some "process oriented" remedial reading programs For many readers, these programs have not been successful The more general lesson to be learned from this is that process differences between groups of readers do not necessarily translate well into instructional programs Such process differences may be only symptomatic of more profound differences

In recent years, there has been an increasing emphasis in cognitive science research in the study of the importance of knowledge in cognition More and more studies in cognitive psychology treat prior knowledge not as a nuisance variable to be eliminated in the study, but rather as the important variable to study In part, the recent emphasis on knowledge based research has been spurred on by developments in artificial intelligence research In researchers attempts to make computers smart, they have found that knowledge of the domain plays a critical role (Barr and Feigenbaum, 1984) This agrees quite nicely with the finding that expert and novice human chess players do not differ in general processing skills (i e , short term memory processes), instead they differ in that expert chess players have a great deal of highly organized domain specific knowledge about chess positions Increasingly, researchers in reading and reading instruction are recognizing the importance of studying the prior knowledge of the reader, (e g , Chiesi, Spilich and Voss, 1979, Kieras and Johnson, 1984) and recognizing the importance of the tradeoff between processing skills on the one hand and knowledge on the other and what the implications are of this tradeoff to remedial training

**Knowledge and Readability**

The concepts behind the functional reading assessment battery, described in the paper by Dr Sticht, have been extended for use in estimating the reading demands of documents I will briefly describe some exploratory research that we have done (this work was done in collaboration with Dr Sticht), as it begins to shed light on the role of prior knowledge on reading and readability

Readability researchers have for many years been interested in determining ways to best match readers to documents That is, the goal has been to provide a way to select a document for a reader so that the document is not too difficult (or too easy) for the reader to comprehend The result of the research has been among other things, the development of many different readability formulas (see Klare, 1974 for a review) Very briefly, the formulas attempt to relate surface structure characteristics (e g , sentence length and word length) of the document to comprehensibility It has been found empirically that these characteristics are correlated with measures of comprehensibility (cloze tests, multiple choice tests etc ) While there have been many criticisms of these formulas (e g , Duffy, 1985) they continue to be used

One major problem with these traditional formulas is that they treat all readers identically That is, a given document is presumed to be equally comprehensible to all readers regardless of a reader's prior knowledge of the subject matter Clearly this is a profound problem Therefore, the goal of our research was to assess the influence of prior knowledge on the comprehensibility of a document and to develop new readability formulas that take account of this prior knowledge

We studied a sample of 296 Navy recruits who had taken the Navy knowledge test as part of the functional reading assessment battery (described in the paper by Dr Sticht) This test does not involve reading any paragraphs, subjects are simply asked questions about the Navy, thus the test measures their prior Navy knowledge Based on the results of this test subjects were broken into four groups (of roughly equal size) that reflected performance on the test Group 1's (low knowledge) performance was lowest on the knowledge test and Group 4 (high knowledge) was the best Groups 2 and 3 were intermediate

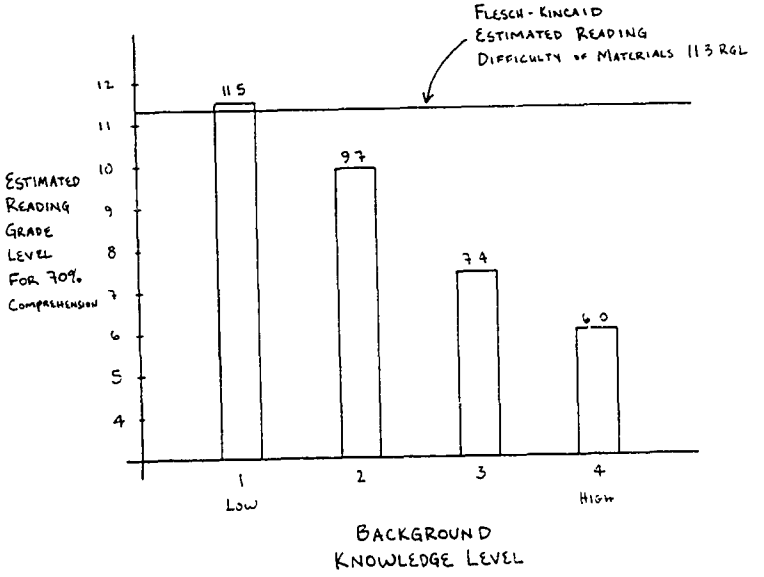
All groups read eleven different Navy related passages For three of the passages subjects performed a reading-to-learn task (the multiple-choice questions had to be answered from memory, the passages were not available for question answering), and for the other eight passages

the subjects performed a reading-to-do task (the passages were available for consultation in answering the comprehension questions) In the preliminary results reported here, all eleven passages were collapsed into a single analysis, in a later report we will break these apart

We performed a standard readability formula analysis (see e g , Caylor, Sticht, Fox & Ford, 1973) but we did it separately for each of the four knowledge groups For each knowledge group we determined the reading grade level (RGL) required to comprehend each passage at the 70% level (a linear interpolation method was used to determine scaled RGL) Then four different surface structure features of the passages were used to predict scaled RGL average sentence length, average word length, proportion of content words and average content word length The four different formulas differed primarily in their intercept and were combined into single formula that contained surface structure components and a knowledge component

The results are clear in showing three general trends 1) in general, the more readers knew about the domain, the easier the passages were to comprehend, 2) traditional readability formulas will overestimate the readability demands of a passage for those who have some prior knowledge in the content area, and 3) prior knowledge accounted for more of the variance in the prediction of comprehensibility than did the surface structure features

Using the new knowledge-based readability formula the estimated reading difficulty of each passage was computed for each group (the estimates are in RGL units) The estimated difficulty averaged over all passages is presented in Figure 1 for the four knowledge groups That is, for each knowledge group the figure shows the estimated RGL needed to comprehend the passages at the 70% level The solid line at the top of the figure shows the estimated reading difficulty of the same passages by the standard DoD Flesch-Kincaid readability formula (Kincaid, Fishburne, Rogers & Chissom, 1975)



Consider the low knowledge group One can see that the estimated reading difficulty for the low knowledge group is almost identical to the reading difficulty estimated by the Flesch-Kincaid formula This indicates that in considering readers with very little prior knowledge of the content area, the standard DoD readability formula predicts actual performance relatively well However, as prior knowledge of the content increases one can see that the traditional formula consistently overestimates the RGL needed to comprehend the passage Thus, the traditional formula treats all readers as if they have little or no knowledge of the domain If the reader does know something about the domain, then a given document will be easier to comprehend than would be indicated in a traditional formula

In computing the knowledge-based readability formula we found that the surface structure variables accounted for roughly 26% of the variance in predicting the scaled RGL values The knowledge variable, on the other hand accounted for about 43% of the variance Thus, the knowledge variable accounted for a substantially greater proportion of the variance, and this difference was statistically reliable

The claim has been made in this research that the more readers know about the domain to be read, the easier a passage is to comprehend It should be pointed out that readers in this study evidently did indeed read and extract information from the test passages and did not simply answer the comprehension questions without reading the passages An indication of this comes from the fact that overall performance on the Navy knowledge test (no reading passages involved) collapsed over all groups was 47% while test performance based on the reading passages was 74% Thus, the information extracted and learned from the passages accounted for a sizable increase in test performance

It should be emphasized that these are only preliminary results, and that much work needs to be done However these findings are potentially extremely important We have demonstrated that prior knowledge of a content area has a large effect on the comprehensibility of a particular passage, and we have been able to quantify that effect in a precise way We have also developed a technique that can be used quite generally to assess the role of prior knowledge in any domain and incorporate the knowledge influence into a readability formula The new technique will allow us to construct a variety of different formulas, based on different measures of prior knowledge, to assess the readability of both technical and instructional documents In this way we will be able to match better, readers to documents taking into account what the readers already know about what is to be read

## REFERENCES


Barr, A. & Feigenbaum, E.A (1981) *The Handbook of artificial intelligence, Vol 1* Stanford, CA HeurisTech Press

Caylor, J S , Sticht, T G , Fox, L C & Ford, J P (March 1973) *Methodologies for determining reading requirements of military occupational specialties* (HumRRO Tech Rep 73-5) Alexandria, VA Human Resources Research Organization

Chall, J (1967) *Learning to read The great debate* New York McGraw-Hill

Chang, F R (1983) Mental processes in reading A methodological review *Reading Research Quarterly, 18,* 216-230

Chiesi, H L, Spilich, G J & Voss, J F (1979) Acquisition of domain-related information in relation to high and low domain knowledge *Journal of Verbal Learning and Verbal Behavior, 18,* 257-273

Crowder, R G (1982) *The psychology of reading An introduction* New York Oxford University Press

Duffy, T M (1985). Readability formulas What is the use? In T Duffy & R Waller (Eds ) *Designing usable texts* New York. Academic Press

Gibson, E J & Levin H (1975) *The psychology of reading* Cambridge, MA· The MIT Press

Johnson, W & Kieras, D E (1983). Representation-saving effects of prior knowledge in memory of simple technical prose, *Memory & Cognition, 11,* 456-466

Kieras, D E & Just, M A (Eds ) (1984) *New methods in reading comprehension research* Hillsdale, NJ: Lawrence Erlbaum Associates

Kincaid, J P , Fishburne, R.P., Rogers, R.L & Chissom, B S (February 1975). *Derivation of new readability formulas (Automated Readability Index, Fog count, and Flesch Reading Ease Formula) for Navy enlisted personnel* (CNTT Research Branch Rep 8-75), Millington, TN Chief of Naval Technical Training

Klare, G R (1974-1975) Assessing readability, *Reading Research Quarterly, 10,* 62-102

Kucera, H & Francis, W (1967) *Computational analysis of present-day American English* Providence, R I Brown University Press

Nagy, W E & Anderson, R C (1984) How many word are there in printed school English? *Reading Research Quarterly, 19,* 304-330

Nagy, W E , Herman, P A & Anderson, R C (1985) Learning words from context Reading Research Quarterly, 20, 233-253

Sticht, T G (June 1982) *Basic skills in defense* (HumRRO-PP-3-82) Alexandria, VA Human Resources Research Organization


Footnotes


The views, opinions, and/or findings contained in this paper are those of the author and should not be construed as an official Department of the Navy or Department of Defense position, policy or decision, unless so designated by other official documentation

The work on readability was done in collaboration with Dr Thomas Sticht, a more formal, co-authored paper describing this research is forthcoming I am grateful to Dr William Montague for generally useful discussions concerning the role of knowledge in cognition and to Dr Edwin Aiken for helpful comments on an earlier draft of this paper Thanks are due Tami Lopez for assistance in the data analysis

# The XFSP: Reading and Mathematics Project

Thomas Sticht, Louis Armijo, Natalie Koffman, Kent Roberson
U.S. Naval Postgraduate School

In an ongoing program of instructional development, the authors and other colleagues from the Naval Postgraduate School and the Navy Personnel Research and Development Center are developing 45 hour developmental reading and mathematics programs for the Navy under sponsorship of the Chief of Naval Education and Training. The programs we are developing will replace programs being offered by some dozen contracting organizztions world-wide. The contractor programs are "general" in their orientation, whereas the programs we are developing are Navy-related, and they are oriented to assist enlisted personnel meet requirements for career promotions to higher paygrades and responsibilities.

In this paper, which will focus on the XFSP (Experimental Functional Skills Program): Reading, I will first discuss the conceptual framework for the development process, including a simple model of the *human cognitive system* and the concept of *functional context training*. These conceptual frameworks help to give the development process broad, general directions. In this regard, they provide what is typically missing in current instructional systems design (ISD) methods, and that is a view of human cognition and the conditions for its use that can provide heuristics for needs assessment, task analysis, and program design.

Following the discussion of the underlying concepts for program development, I will briefly describe the program we have developed as it stands at the present time, and then I will describe the assessment test battery we have developed to assess program effectiveness. Finally, a small scale study to compare the Navy-related reading program to the general reading program offered by a local contractor is described.

### *Human Cognitive System Model*

In conducting the development of the XFSP. Reading program we have worked from a stripped-down, simplified model of a human cognitive system and the processes the system uses for extracting and representing information in the environment. The model is schematized in Figure 1. The model contains three major components, two of which are "inside the head", and the third component which is "outside the head" and includes various information displays produced by or comprehended by the cognitive components. The latter include the *knowledge base* , which is a long term memory that contains all the information and knowledge possessed by the person, and the *processing skills* , including language, that operate on the information in both the knowledge base and the external environment to produce comprehension, communication, and thinking. The model as depicted in Figure 1 is of one who possesses both oracy and literacy information processing skills. The literacy processes include all those used to recode written language into internal forms comparable to those used in oral language, and, in addition, to perform all those literacy tasks that are not instances of writing as a second signaling system for speech. The tasks unique to literacy are those made possible by the properties of graphic displays: they are more or less permanent (thereby permitting study), they can be arrayed in space (permitting the construction of forms, signs, flow charts, graphs, etc.), and they can use the properties of light (contrast, color) to guide attention and facilitate information processing.

According to the model of Figure 1, then, the performance of literacy tasks requires knowledge about what one is reading or writing (including mathematics knowledge when reading in that domain), processing skills for thinking about what to communicate or for comprehending what others communicate; and, of course, graphic displays of information in

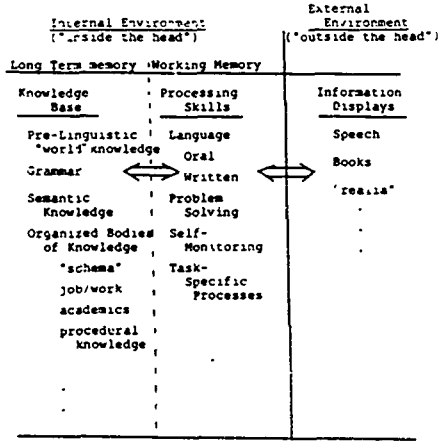



Figure 1

Cognitive system model for instructional program development

the environment to be processed for meaning. An important implication of this analysis is that it reveals that literacy, considered as the ability to comprehend and use the graphic symbol systems of writing, graphing, illustrating, mathematics, and so forth, can be enhanced by improving either one's knowledge base in a given task domain, or one's processing skills,or, as in the case of improving the readability of materials, through the redesign of the graphic information displays the cognitive system must deal with, or a combination of these factors. Use of this conceptual framework in determining the learner's needs in literacy programs is illustrated later on, following a discussion of the *functional context* approach to instructional design that has undergirded the XFSP:Read program development activities.

*The "Functional Context" Concept*

Along with the simple cognitive model of Figure 1, we have followed what is known by some as the "functional context" approach to education and training development (Shoemaker, 1967). The essence of this approach is contained in two major goals for instruction. First, always try to make the instruction as meaningful to the learner as possible in terms of the learner's prior knowledge. This facilitates the learning of new information by making it possible for the learner to relate it to knowledge already possessed, or to make it possible for the learner to transform old knowledge into new knowledge. Second, as much as possible, use the materials and equipments that the learner will use after training or education as part of the instructional program. This will motivate the learner by showing that what is being learned is relevant to a future goal, and it will promote transfer of learning from the classroom to the next training or "real world" activity. In short, the functional context method of instructional design attempts to motivate and promote learning and transfer by making the program meaningful in terms of the learners past, present, and future.

*The XFSP: Read Program*

The XFSP: Read program was developed following the guidance of the concepts described above applied in studies of what kinds of tasks Navy personnel perform using reading skills in training and job settings. In this research, students, instructors, and job performers in ten Navy jobs were interviewed and asked for information regarding two major types of reading tasks: reading-to-do something and reading-to-learn something. In a reading-to-do task, the person is performing some job task, needs some information from a document, looks-up the information, holds it in working memory long enough to apply it, and can then forget it. In a reading-to-learn task, the person reads information to be stored in long-term memory as part of the knowledge base, and then retrieves it (or a reconstruction of it) for use at some later time, such as taking an end of week test, or for performing a task on the job.

The interviews with personnel revealed that reading-to-do and reading-to-learn were performed to about the same extent in school situations, but on the job reading-to-do comprised about three-fourths of the reading tasks. It was also found that the processing skills performed in reading-to-learn were more complex than those used in reading-to-do. Whereas the latter emphasized information location and extraction skills, such as use of tables of content, indexes, "thumbing" or "flipping" through or searching tables and figures, reading-to-learn involved more elaborate activities to merge new information with old knowledge. The primary reading-to-learn processes were categorized into four groups: (1) reread or rehearse processes, in which learning was accomplished by rereading some portion of the material, or was repeated in some way over again to oneself; (2) question/problem solve processes, in which learners asked themselves questions about the material, or solved problems in textbooks; (3) relate/associate processes, in which the learners transformed what they read into some other form, either by paraphrasing, making internal images, watching a movie or demonstration and relating what they had read to the new information gained from these "iconic" or "realia" information displays (see Figure 1); and (4) focus attention processes, such as highlighting with colored pens, underlining, summarizing or some similar methods for focusing attention on a limited aspect of the material, usually in conjunction with a reread or rehearse activity later on.

The interviews also revealed the role of the knowledge base in performance of reading tasks. For instance, it was found that close to 60% of job tasks involving reading had been performed previously, and for about half of the 325 reading tasks cited by the sample of 178 personnel, additional reading related to the task had been performed, and for two-thirds of these cases, the related reading helped in reading the material cited in the interview (see Sticht, 1977 for a more complete description of this work).

*The Career Progression Reading Program*

On the basis of the foregoing research and additional study of the reading demands of the Navy environment, we have designed and developed a reading program that has a functional context for Navy personnel in that the program uses Navy content derived from materials they must study to pass promotion tests, and the information processing skills are of immediate use to them. Most of the students are "mid-level" literates, with reading skills in the 6th to 10th grade levels, and most have had one one or more years of duty, and so they are able to relate their knowledge base to the content of the course.

The instructional program is designed to be delivered either by a teacher using paper-and-pencil materials in a classroom, or by a teacher, paper-and- pencil materials and computers

in a classroom, or in a learning center using computer-based instruction alone. Figure 2 shows the teacher-, book-, and computer-based instructional classroom delivery system that is the central model for the delivery system. The basic course is three weeks in duration and students attend for three hours a day for a total of 45 hours of instruction (the time for instruction was specified by the Navy Campus office) The books include a special reader, which, based on the model of Figure 1, we call a Navy Knowledge Base. This reader contains extracts and revisions from the technical manuals that personnel must read and learn from to pass promotion tests or perform higher level duties. The contents were selected because they contain the information that is on the practice tests in the manuals and are deemed important by the Navy management.

The second book is called the Information Processing Skills book, and it is modeled after the processing skills component of the model cognitive system. This book presents lessons and practice in performing various processing skills for reading-to-do and reading-to-learn tasks using the Navy Knowledge Base book. The goal here is to present *externally* a knowledge base and processing skills to serve as didactic tools for talking about the *internal* knowledge base and processing skills and to hence make students aware of their cognitive systems and how to apply them to doing and learning literacy tasks

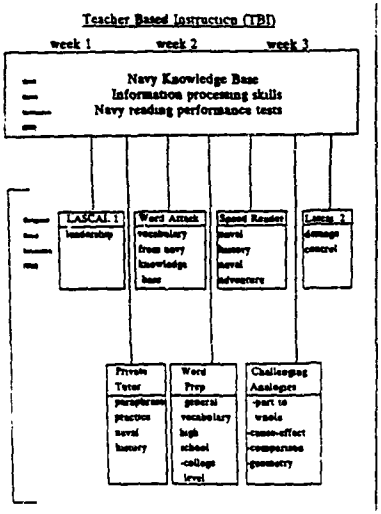



Figure 2

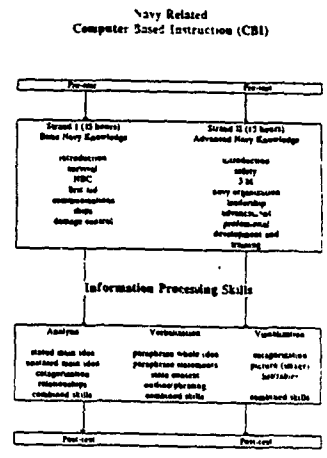



Figure 3

The computer based instruction consists of public and private domain programs that serve as sort of "electronic worksheets" that the teacher can assign students to do while he or she works with other students on a small group basis. The computer software was selected to permit the teacher or curriculum specialists to personalize the instruction, and so each program has an editor feature that permits one to enter content works or paragraphs for the specific literacy domain they wish to teach. The Lascai, Word Attack, Speed Reader, Word Prep, and Ghallenging Analogies programs all have a game format that operates automatically on the content that is edited into the program. The Private Tutor is sold by IBM and is a very inexpensive, easy to use authoring system for preparing indivdiual lessons or entire computer-based instructional courses.

Figure 3 shows the computer-based instruction that has been developed for use in a learning center as a stand-alone, self-paced instructional program. The program has two Strands, each providing about 15-20 hours of instruction depending upon the reading skill level of the user. The Strand I material contains content appropriate for sailors seeking promotion to lower level supervisory positions, while the Strand II material is primarily for those seeking higher level promotions (these are only approximations, since some contents are useful across

the board in both strands). The bottom part of Figure 3 shows the processing skills taught in both strands of the computer-based instruction.

*Functional Reading Assessment Battery*

In addition to the functional, Navy-related, standardized teacher-, book-, and computer-based instructional programs, we have also developed a new reading test battery to test Navy-related reading skills (reading-to-do and reading-to-learn) as well as Navy knowledge. The latter, that is, knowledge gained, is rarely measured in reading programs because reading is considered as a content-free, process skill. But the cognitive model of Figure 1 makes clear that knowledge of what one is reading is required to make reading comprehension possible. So we are assessing the improvement in knowledge as a function of participation in the functional reading program. Figure 4 summarizes features of "traditional" reading tests and our new "functional" reading battery.

*An Evaluation Study*

In a small evaluation study, the improvement of a sample of students who took a "general" reading program offered by education contractors was compared to the improvement of students in our "functional" reading program on three tests: a general reading test that gives grade levels of performance, our Navy reading-to-do test and our Navy Knowledge test (the reading-to-learn test was not available at this time). The results are summarized in Figure 5. The data show that, in general, people tend to learn what they are taught. The "general" reading program did better on the general reading test, but this did not transfer to the Navy reading and knowledge tests to any significant degree. The Navy-related reading program, on the other hand, resulted in only a little improvement on the general reading test, but made consistent improvements where it counts for Navy personnel, that is, in their Navy reading and knowledge.

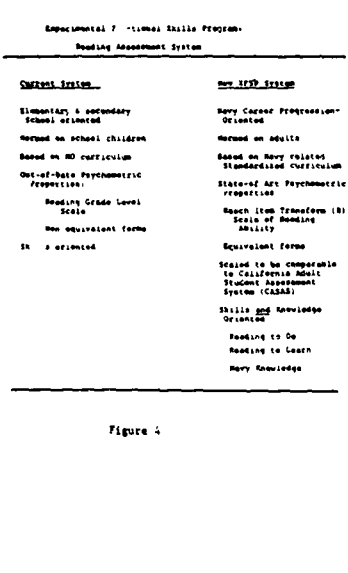



Figure 4

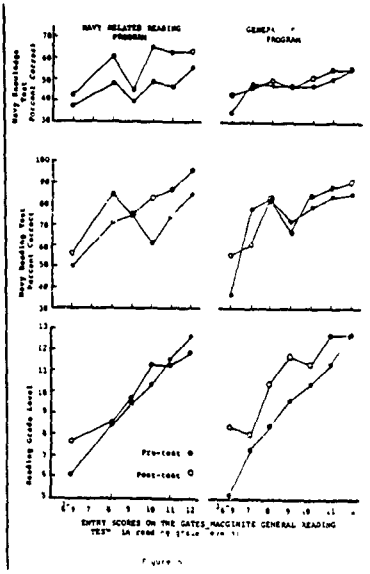



Figure 5

*Future Tasks*

At the present time we are polishing up the XFSP:Read materials and packaging them for implementation and further evaluation. We have also started the development of an XFSP:Mathematics program based on the same general conceptual framework discussed herein. Finally, we have initiated some exploratory study of the use of the new Navy reading assessment battery in developing an approach to estimating the reading demands of Navy training and jobs using the reading-to-do and reading-to-learn tasks as separate criterion variables, and including Navy knowledge and other knowledge domains as variables in the predictor along with the more traditional passage-based variables (e.g., word and sentence length) used in readability formulas.

*Reference*

Shoemaker, H. (1967). The Functional Context Method of Instruction. HumRRO Professional Paper No. 35-67, Alexandria, VA: Human Resources Research Organization.

*Footnote* Colleagues who have worked on one or more aspects of this project include Drs. R. Weitzman; Frederick Chang; William Montague; and Monique Allen; Joan Molleur; Susan Wood, Jim Joelson; Jim Powers; Tom Kasten; and Otto Kruse. Project Monitor is Dr Judy Moracco of CNETT/TTI.

# A CONCEPTUAL FRAMEWORK FOR OCCUPATIONAL EXPLORATION

John J. Pass, PhD

Navy Personnel Research and Development Center
San Diego, California 92152-6800

## Introduction

The need for improved dissemination of military career information for purposes of occupational exploration and person-job matching (PJM) has long been apparent. With the advent of voluntary forces, the military was removed from automatic consideration in the career plans of the nation's youth. The problem has been to provide young people with the means of career exploration in order to acquaint them with the benefits of military training and career opportunities.

## Basic Components

Essentially, an occupational exploration system must include: (1) personal assessment procedures to measure the individual along a number of cognitive and noncognitive dimensions in order to provide selection information to the organization and enhance the individual's self-awareness; (2) occupational information that is accurate, timely, sufficiently comprehensive, and personalized to the individual; and (3) a bridging mechanism between personal information and job information that might include a vocational guidance or counseling process.

The final design of any occupational exploration system will be determined by its purposes, the constraints of cost and technology, and the ingenuity of the designer. Figure 1 illustrates some examples of the types of information that could be included. First, there are the individual dimensions or attributes to be measured. Second, there is the occupational information that would be useful for an individual to use in gaining some understanding of requirements and conditions of occupations or careers. Third, typical search strategies that are useful in the attempt to match the individual characteristics with organizational and occupational requirements are shown.

Search strategies are a critical component of an occupational information system, one that deserves careful consideration. Upon the search strategy rest many of the system's time and cost requirements. Equally as important is the effect of the search strategy on the person-job match. Any effective procedure for matching persons with jobs must take into account both individual and institutional characteristics. Individual characteristics include abilities, preferences, interests, and goals. Institutional characteristics include priorities, objectives,

training program vacancies, and personnel requirements. An effective search strategy, then, enables a narrowing down of alternatives so that information is developed or presented for those occupations that are best suited to the individual. Ultimately, the search strategy conduces to a tentative choice of an occupation or an occupational field.

System evaluation is an additional component of an occupational exploration system. This can prove to be an important element. One part of this component can be a program that seeks to obtain user evaluation of the system. A series of multiple-choice questions may be administered on-line to assess general satisfaction with the guidance process and the interactively programmed computer system. Other, more formal evaluative research can also be employed. The information from both sources can be used as feedback to help improve the system.

| APPLICANT ASSESSMENT | OCCUPATIONAL INFORMATION |
|---|---|
| -Aptitude | -Job Requirements |
| -Interests | -Training Provided |
| -Medical Evaluation | -Physical Demands |
| -Work Experience | -Career Benefits |
| -Career Maturity | -Job Openings |

SEARCH STRATEGIES

-Accessing Occupational Information
-Matching Personal and Occupational Characteristics
-Narrowing of Options

SYSTEM EVALUATION

-User On-Line Evaluation
-Interview
-Impact Evaluation

Figure 1. Occupational Exploration System

## Current Situation and Prospects

Unfortunately, although the need has been recognized for many years, occupational information is frequently scanty or outdated, and difficult to access. Many times, there is a lack of individual assessment methods so that personal considerations can be dealt with. Furthermore, current occupational exploration

systems in both civilian and military settings are often notably deficient in provision of accurate and comprehensive military occupational information.

However, there are a number of bright spots in this setting. One of them is computerization. Computerization offers a major hope for amelioration of the deficiencies in availability and accessibility of accurate, timely, and personalized occupational information. Computers can provide immediately scored self-assessment instruments, didactic counseling materials, and comprehensive, up to date, rapidly retrieved, personalized job information. Computers can deliver standardized occupational information to the job applicant, thus minimizing or at least limiting the introduction of information bias. Individuals benefit from complete information and impartial guidance in selecting an occupation, and counselors are freed of the need for broad and current knowledge of military occupations and training.

Furthermore, in the civilian community, the past generation has witnessed the development of excellent computerized occupational information systems. And, more recently, reliable military career information has been made available under the sponsorship of The Office of The Assistant Secretary of Defense (FM&P), in support of a policy advocating improved guidance and counseling, and the provision of honest, accurate military career information to high school and college students as well as to prospective enlistees.

## The Military Setting

Within the military research community, there is progress in the practical evolution of systems and techniques to accomplish those ends, with the design and development of systems and techniques that have wide applicability in a variety of civilian and military settings. Let me cite a few examples that incorporate various components shown in Figure 2 above.

In FY75, the Navy Personnel Research and Development Center (NAVPERSRANDCEN) began work on an advanced developmental effort called Project CONTRACT (Computerized Navy Techniques for Recruiting, Assignment, Counseling, and Testing). Two products of CONTRACT were the development and demonstration of a computerized system called the Navy Vocational Information System (NVIS) and the development of an optimal personnel assignment algorithm for use with the computerized entry level job assignment system employed by the Navy.

One research objective of NVIS was to develop a prototype interactive computerized occupational information system that would provide young men and women with personalized occupational guidance and a list of related civilian and Navy jobs that demonstrated a good match between their own attributes and typical job requirements.



Copy available to DTIC does not permit fully legible reproduction

NVIS databases contained information on 279 civilian jobs, 114 worker trait groups, 79 entry level Navy occupations, and more than 100 specialized Navy jobs. NVIS maintained an interactive dialogue with the user via a cathode ray tube.

NVIS proved to be the precursor to an expanded Navy occupational information system because an outgrowth was the conceptualization and advanced development of a microcomputer based person-job matching system, another prototype computerized vocational guidance system called AGENA (Automated Guidance for Enlisted Navy Applicants).

AGENA used interactive programming to lead the person through a logical, thought-provoking dialogue that introduced the system and equipment; proceeded through preliminary aptitude screening; progressed, via learning how to plan for a career and discover personal interests and aptitudes; and ultimately explored a number of Navy entry-level occupations that matched personal interest and aptitudes, with the opportunity to assess the availability of options in a rather wide timeframe.

Within the broad function of person-job matching, three subfunctions were supported in the AGENA system: (1) aptitude screening; (2) vocational guidance; and, (3) assignment prediction. AGENA took personal and organizational factors into consideration and, coupled with information on when the individual wanted to enter the Navy, determined which entry level jobs were most appropriate for meeting individual and organizational considerations. The availability of these entry-level assignments was indicated for a three-month timeframe.

AGENA information on the entry-level Navy jobs was available in two formats. The abbreviated version, designed for video terminal display, included five sections: (1) general description, (2) related civilian jobs, (3) qualifications, (4) working conditions, and (5) Navy opportunities. An extended description, available in hardcopy as an option, included all sections of the abbreviated description plus three additional sections: (1) what the people in the rating do, (2) sea/shore rotation, and (3) the training provided by the Navy.

The substantial contributions Navy training can make to total career development discussed in AGENA, along with a brief discussion of the general value of Navy training and experience. On-line access to Civilian Occupations Database gave descriptions of civilian occupations (or clusters of occupations) related to the Navy assignment opportunities on the video display terminal and a hardcopy output from the printer. Descriptions of civilian occupation included: (1) general description, (2) qualifications and training (3) pay wan working conditions, (4) employment outlook, and (5) related Navy jobs. Finally, a brief discussion of additional benefits of Navy enlistment (e.g.,medical benefits) was included.

AGENA was shown in a demonstration version in the fall of 1981. Subsequently, NAVPERSRANDCEN undertook supportive R&D for the Army Joint Optical Information Network (JOIN). The purpose of this research was to design, develop, test, and incorporated in the JOIN system, a computerized vocational guidance system, and to develop other capabilities including computerized adaptive screening and assignment prediction. Products of that effort have been incorporated into JOIN or are under further development.

To the list of military laboratory efforts in the area of occupational information systems would have to be added the important work being carried on by the Canadian Forces Personnel Applied Research Unit, in Toronto. Their Canadian Forces Information System offers several excellent examples of the use of the components outlined above, including orientation video and realistic job preview.

However, the examples I have given were to illustrate the incorporation of the various elements or components into a system that delivers accurate and consistent military occupational information in a user friendly way.

## Conclusion

There is growing interest and progress in military occupational exploration. While many issues remain in the development of such a system (e.g., level of occupational information specificity, degree of automation., and the location of the system within the organization), these issues tend to be more easily resolved once the objectives for the system are established. We can certainly envision two clear objectives: to increase the awareness of military career options; and to conduce to more optimal person-job matches that will enhance job satisfaction and productivity while decreasing attrition and personnel turbulence.

The potential payoff of pursuing this R&D area in terms of increasing the numbers of military applicants and decreasing attrition definitely warrants additional effort. Work is moving ahead on the conceptual design of a comprehensive computerized vocational guidance system that assists the job seeker in self exploration, provides accurate and consistent occupational information, and helps in relating personal characteristics and occupational choices. To enhance awareness of military career options, and make military job training a viable element of career planning to be seriously considered by those entering the work force, is the goal of these projects.

# Development of the Career Maturity Assessment

Esther E. Diamond, Ph.D.
Educational and Psychological Consultant
Evanston, Illinois

## Introduction

Many recruits entering military service have not engaged in thoughtful career planning, or adequately assessed themselves, or sought, located, and assembled career information, or developed an appropriate set of vocational coping behaviors. Consequently they are not yet ready to make informed occupational choices.

To assist recruits in the career choice process, the design, development, and pilot testing of a career maturity assessment (CMA) instrument was undertaken, for eventual incorporation into a Computerized Vocational Guidance (CVG) system for use in recruiting. CMA scores would determine the point in the guidance sequence to which applicants would branch.

## Background

As Crites (1973) has pointed out, prevailing views of vocational behavior before the 1950s were almost entirely nondevelopmental. Vocational decision-making was considered to be a time-bound, largely static, one-time event, which usually occurred upon high school graduation, "when an adolescent took stock of himself and the world of work and then decided what he was going to do" (p. 5). Super's seminal work in vocational development theory (1953,1957) was among the first to introduce the concept of career maturity, defined as the readiness of an individual to make career decisions expected at a particular age. He defined the term further as "...the repertoire of coping behavior leading to outcomes...a developmental rather than an outcome construct" (1974, p.11). That decision-making plays a major role in vocational development has been underscored in the work of most major career development theorists (e.g., Super, 1984; Ginzberg, 1984; Crites, 1973, Osipow, 1983). In general, self-concept or self-knowledge and decision-making skills appear to be common elements in most of the research reviewed. In developing the CMA, then, these concepts were the major focus.

## Instrument Development

The Navy Personnel Research and Development Center (NPRDC) specified that CMA items should be written at the sixth-grade reading level and should not exceed 30, with maximum administration time of five minutes; face validity should be high; responses should be true/false or Likert-type; scoring should permit dichotomous reporting of results; scale reliability should be at least .65; and the instrument should be validated against a suitable commercially available instrument.

Following review of the related literature, a pilot version was constructed, with items drawn from a number of sources, including the literature review, the author's professional experience in developing a number of career guidance programs, and the objectives and results of the

National Assessment of Career and Occupational Development (National Assessment of Educational Progress, 1976). The CMA at first consisted of 38 items, to permit later elimination of any that proved unsatisfactory on the basis of item statistics.

Three responses were possible for each statement: Y (Yes, the statement is generally true), N (No, the statement is generally not true), and ? (Uncertain). The Y and N responses had a score value of 0 or 2, depending on the particular item, while question marks had a score value of 1. Following are three examples of the items:

I envy people who don't have to work.

I am afraid to try new things.

I have a good idea of my interests and abilities and how they relate to different occupations.

Variables

The total score on My Vocational Situation (MVS) (Holland, Daiger, & Power, 1980), a 26-item vocational decision-making instrument with three scales, was selected as the criterion variable.

To study the validity of the CMA for subgroups, classification variables were collected for age (under 25, 25 and over), gender (male, female), and level of education (less than high school graduation, high school graduation, some college, four-year college graduation).

Sample

The sample consisted of U.S. Army recruits from nine forts across the country. The total number tested was 405 (323 males and 62 females), but because of errors in the administration, not all recruits completed both the CMA and the MVS. CMA answer sheets for 401 recruits and MVS answer sheets for 304 recruits were analyzed.

Data Analysis

The 401 CMA cases had either no missing responses or less than 25% missing. Missing responses were coded as ? and scored as a 1. For purposes of calculating individual item contributions to reliability, only cases with no missing responses (348) were used. For the MVS, 304 cases were used. Intercorrelations were obtained between all variables. Alpha was computed for total CMA and MVS and for CMA with each item, in turn, deleted.

Results

Correlation between total CMA score and total MVS score was .75, indicating a strong relationship between the CMA and the criterion measure, with more than 56% of the variance (.75 squared) accounted for. Alpha was .83. To reduce the total number of items in the CMA, the individual item data described previously and the correlations with total score were examined. The four items with the lowest correlations with total score and whose

elimination would increase alpha the most were dropped. NPRDC agreed to let the instrument stand at 34 instead of 30 items. A second data analysis was then run.

Alpha for the revised CMA increased to .85, and correlation with total MVS increased to .78, indicating an even stronger relationship than before, with 60 percent of the variance (the squared correlation of .78, which is equal to .6084), accounted for. Total score mean was 51.04, the median 53.02, and the mode 54. The cutoff was set at the median, 53.

The mean for females was higher than that for males, and there was very little difference between means for the two age groups, although the older group scored higher. Means increased with educational level.

Summary statistics for the total CMA sample, based on the 34 items, are presented in Table 1. Table 2 gives means and standard deviations by gender, age, and educational level. The z-ratios computed for the differences between uncorrelated means showed no significant differences between males and females, age groups, or educational levels. Table 3 gives intercorrelations with MVS, gender, age, and educational level.

Table 1

Summary Data for Total Career Maturity Assessment Sample
(N=401; highest possible score = 68)

| | | | |
|---|---|---|---|
| Mean | 51.04 | SD | 10.73 |
| Median | 53.02 | SEmeas | .54 |
| Mode | 54.00 | Skewness | - .75 |

Table 2

Career Maturity Assessment Means and Standard Deviations by Gender, Age, and Level of Education

| Variable | N | Mean | SD |
|---|---|---|---|
| | (By Gender) | | |
| Male | 319 | 50.51 | 10.97 |
| Female | 82 | 53.08 | 9.52 |
| | (By Age) | | |
| Under 25 | 364 | 50.56 | 10.62 |
| 25 and over | 37 | 55.75 | 10.76 |
| | (By Level of Education) | | |
| Less than HS grad | 57 | 48.68 | 11.68 |
| HS grad | 205 | 50.33 | 10.82 |
| Some college | 125 | 52.41 | 10.00 |
| College grad | 14 | 58.63 | 7.04 |

Table 3

Intercorrelations of Career Maturity Assessment (CMA) with My Vocational Situation (MVS), Gender, Age, and Educational Level

| | MVS | Gender | Age | Ed Level |
|---|---|---|---|---|
| CMA | .78** | .10* | .14** | .17** |
| MVS | | .08 | .13** | .11* |
| Gender | | | .05 | .18** |
| Age | | | | .29** |

*p < .05
**p < .01

Discussion and Recommendations

The CMA demonstrated several characteristics that support its potential usefulness as an indicator of career maturity. Reliability is .85. Correlation with the MVS indicates that it probably measures the same construct and that both concurrent and construct validity could be inferred--to the extent, at least, that they can be inferred for the MVS. Holland et al (1980) describe the MVS as having "a moderate degree of reliability and promising validity" (p. 1). The latter is described as the construct validity of the three MVS scales--Vocational Identity, Occupational Information, and Barriers. For total MVS score for the sample tested with both the MVS and the CMA, alpha was .88. The two instruments are intended to measure the same kinds of vocational needs or problems on the part of the respondent. Since the CMA score distribution was negatively skewed, with the mode higher than the median and the median higher than the mean, one might conclude that the items were not too difficult for most recruits to read and understand.

There are preferable ways of determining the face validity of the items and of setting the cutoff score. Budget and work-time restrictions, however, precluded such activities as the use of judges for determining face validity; administration of a longer, better-researched criterion measure than the MVS; factor analysis to study the construct validity of the CMA; and use of a standard-setting procedure such as the Angoff, Nedelsky, or Ebel methods (Livingston & Zieky, 1982) for establishing the cutoff score. These activities are suggested as recommendations for future, continued development.

In addition, the reliability of the dichotomous decision will need to be obtained empirically after the cutoff score has been in operation for an adequate period of time. The decisions made at various points above and below the cutoff will need to be examined to determine whether those who really need help with vocational decisions are being identified and whether false positives and false negatives are at an absolute minimum.

Furthermore, since the CMA has no subscales that might indicate specific problem areas (as in the MVS), there is a need to identify the kinds of problems that individual recruits have with regard to vocational

decision-making--for example, obtaining a better picture of their abilities and interests, locating occupational information, clarifying their goals, resolving value conflicts, or improving their decision-making skills. It may be possible to identify such problems through the CVG program. Recruits should also be given the option of asking for help if they feel they need it, regardless of whether they are above or below the CMA cutoff score. In fact, keeping track of such requests might provide still another way of checking on the accuracy of decisions based on the cutoff.

One last word should be said about setting the cutoff score at the median: Fortunately, the decision involved is not capable of having adverse impact on those below the median. The score will not be used for employment or promotion decisions. Its use should be entirely benign, both from the service's and the individual recruit's point of view; and, of course, it should be experimental until such time as policy and further research sanction its operational use.

---

Note: Financial support of this work by Battelle Scientific Services Program, in partial fulfillment of contract DAAG-81-D-0100, is hereby acknowledged. The views, opinions, and/or findings contained in this report are those of the author and should not be construed as an official Department of the Navy position, policy, or decision, unless so designated by other documentation.

---

References

Crites, J.O. (1973). Theory and research handbook: Career Maturity Inventory. Monterey, Calif.: McGraw-Hill.

Ginzberg, E. (1984). Career development. In D. Brown, L. Brooks & Associates, Career choice and development (pp. 169-191). San Francisco: Jossey-Bass.

Holland, J.L., Daiger, D.C., & Power, P.G. (1980). My vocational situation. Palo Alto, CA: Consulting Psychologists Press.

Livingston, S.A., & Zieky, M.J. (1982). Passing scores. Princeton, N.J.: Educational Testing Service.

National Assessment of Educational Progress (1976). The first national assessment of career and occupational development. Career and occupational development report no. 05-COD-00. Denver: Author.

Osipow, S.H. (1983). Theories of career development (3rd ed.). Englewood Cliffs, N.J.: Prentice-Hall.

Super, D.E. (1953). A theory of vocational development. American Psychologist, 8, 185-190.

Super, D.E. (1957). Psychology of careers. New York: Harper.

Super, D.E. (1974). Measuring vocational maturity for counseling and evaluation. Washington, D.C.: National Vocational Guidance Association.

# NAVY R&D IN SUPPORT OF THE ARMY JOIN SYSTEM: LEVERAGING THE GOVERNMENT RESEARCH DOLLAR

Herbert George Baker,
Bernard A. Rafacz,
and
William A. Sands

Navy Personnel Research and Development Center
San Diego, CA 92152-6800

Inter-service competition and the consequences of rivalry among the various branches of the armed forces are legendary. Nevertheless, there are instances of highly productive, cooperative efforts that conserve precious research dollars and shorten the time line between conception and implementation. An excellent example is the research and development conducted by the Navy Personnel Research and Development Center (NAVPERSRANDCEN) in support of a new computerized system being developed by the Army for its recruiting forces.

## Background

For nearly a decade, NAVPERSRANDCEN has held a leading position in the development of prototype automated systems that have direct applicability to military recruiting. Previous efforts of this research laboratory have resulted in a number of innovative products such as mobile van-based computerized military occupational information systems, complex algorithms that predict entry level job openings for periods as long as three months, and the Navy's automated classification and assignment system.

In the late 1970's NAVPERSRANDCEN began work on a sophisticated computerized system for the Navy Recruiting Command (NAVCRUITCOM). This was the Navy Personnel Accessioning System (NPAS) (Baker, Rafacz, & Sands, 1983a). It addressed a broad range of concerns at the recruiting station level; i.e., at the most forward terminus of recruiting operations. Subsumed within NPAS were components that harnessed the power of the microcomputer to the tasks of applicant scre ning, vocational guidance, assignment prediction, and recruiting management support (Baker, 1983; 1985b).

A demonstration version of this system was completed in 1981. Notwithstanding the successful development of this prototype, funds were unavailable for further prosecution of this effort. In addition, NAVCRUITCOM had no immediate plans for automating its recruiting station functions. Consequently, the NPAS project was terminated. Nevertheless, the body of expertise that had been built up in the course of developing the NPAS system held excellent potential for application in other arenas of personnel systems research.

Gearing up for the aggressive recruiting that will characterize the next several years, the Army launched a bold, innovative program to streamline and improve its recruiting and accessioning methods. The United States Army Recruiting Command (USAREC) is the proponent and developing agency for the Joint Optical Information Network (JOIN) System which is in the forefront of this effort (Bryan, 1982). Research and technical advisory services are being provided by the

U.S. Army Research Institute for the Behavioral and Social Sciences (ARI).

JOIN, a stand-alone microcomputer-based system, is in many respects similar to NPAS, while being technologically superior by virtue of advanced microcomputer capabilities. It has been implemented in Army recruiting offices nationwide. In the course of the recruiting process, Army enlisted applicants directly interact with this state-of-the-art accessioning system.

When the Army became the first to apply the benefits of automation to the front line recruiter, the opportunity for rapid inter-service technology transfer became apparent. Research managers of both laboratories became aware of a mutual R&D opportunity; one that could more swiftly advance the Army's project, while keeping intact a Navy research group with a proven track record. Accordingly, ARI proposed that the NPAS research team work on the JOIN System under ARI funding. The vehicle for this inter-service cooperative effort was an inter-laboratory agreement which provided for a 3-year effort beginning in FY82, and terminating on 30 September 1984 (Sands, Gade, & Bryan, 1982).

## Task Areas and Products

Five major task areas provided the focus for the efforts of the Navy research group. Both in-house and contractor resources were enlisted in the ensuing work. Some of the products resulting from this project were directly adaptable from previous Navy innovations, while others represent new endeavors on the part of both laboratories.

Aptitude and adaptability screening was a task area where significant accomplishments were soon evident. The paramount result was the development of the capability to administer, score, and interpret a computer-based adaptive test for screening enlistment applicants. The specific product was the Computerized Adaptive Screening Test (CAST), which was designed to replace a conventionally administered paper-and-pencil instrument. Its purpose is to indicate probable success or failure on a comprehensive test battery used for applicant selection by all the armed services.

Test items had previously been developed under a NAVPERSRANDCEN contract with the University of Minnesota. Algorithms for item calibration, adaptive item selection and administration, scoring, and interpretation of test results were either developed or adapted from those extant in the psychological literature. An experimental version of the test had been completed for NPAS (Baker, 1983a; 1984a); and it proved readily portable to the Army's JOIN system.

Computer software to support the administration, scoring, and interpretation of the CAST on a developmental microcomputer system was written and documented (Baker, Rafacz, & Sands, 1983b). Subsequent to field testing and data collection (Sands & Rafacz, 1983), data analyses and instrument refinement were completed. CAST was refined and programmed to operate on the operational microcomputer system. When it was implemented nationwide on the JOIN

System, CAST became the first large scale operational use of computerized adaptive testing (Sands & Gade, 1983).

Thereafter, NAVPERSRANDCEN researchers continued to serve in a consulting role, assisting in program modifications required to interface the CAST with existing computer programs in the JOIN system, and modifications that permit on-line data gathering for further test validation.

Also developed under this task area was an instrument designed to assess military adaptability. This instrument is based on a values congruence approach in contrast to the more common method of empirically linking biodata with tenure. It was pilot-tested on a sample of recruits (N=540). The result was the identification of four items that discriminate between adaptables and nonadaptables. While further validation is indicated, an instrument of this type could be inserted into the automated screening procedures with little impact on applicant processing time or on the recruiter-applicant interaction.

A second major research focus was vocational guidance. The objective here was to conduct preliminary research in the design of a computerized vocational guidance (CVG) system that would be compatible with the Army recruiting environment. NAVPERSRANDCEN experience in developing interactive, user-friendly guidance systems proved to be a valuable and transferable resource.

Literature review, interviews with system developers, and hands-on appraisal were combined in a survey of available CVG systems, assessing their adaptability to Army recruiting purposes. The study produced preliminary indications of the components that would be needed in a recruiting-oriented CVG system. It concluded that it would be infeasible to adapt any extant system, and recommended development of an automated system specifically tailored to the needs of military recruiting and accessioning (Baker, 1984b).

Subsequently, work on a conceptual model of a recruiting compatible CVG system was initiated. A major step in developing such a system is detailing the organizational and operational constraints that would impact its design. Consequently, a study was undertaken and its results reflected in a report addressing these issues (Baker, 1985a).

Additional studies in vocational interests and values and their assessment delineated the many issues and controversies surrounding these areas of psychological concern, as well as the availability of appropriate instrumentation. Efforts to identify a suitable vocational interest inventory for administration to Army applicants at the recruiting station led to recommendations for the Vocational Interest Career Examination (VOICE) (Alley, 1978) as the most suitable inventory available.

An instrument to assess the career maturity of enlistment applicants was also developed (Diamond, 1985). In use, this instrument will determine whether an applicant requires more or less guidance in occupational exploration. It will also indicate to the recruiter the strength of applicant job preferences, thereby

assisting in focusing the sales interview with applicants whose strong enlistment motivation is job training. The instrument is being considered for incorporation into a JOIN-based system that is being designed to provide Army job information to students in community colleges.

Attention was directed within this task area to the preliminary classification of Army entry level occupations according to the Holland coding schema (Holland, 1973), and to a study of the feasibility of using expressed preferences in a recruiting-oriented CVG system. The occupational classification will facilitate job exploration and the elicitation of preliminary job choices. Expressed preferences may prove to be a method for rapidly focussing the classification interview, as well as demonstrating that applicant preferences are in fact being considered.

Work in a third task area, recruiting management support, sought to enlist the computer in lightening the recruiter's clerical task load. NAVPERSRANDCEN had developed several automated forms and reports generation capabilities as part of the NPAS effort (Baker et al., 1983a). Here again, technology transfer opportunities were obvious.

Interviews with Army recruiters and recruiting managers, and previous experience in the field were employed in assessing needs for forms generation, reports generation, and general word processing. Preparation of the Application for Enlistment (DD Form 1966), a multipage document, is a major administrative burden within the recruiting process. Information needed for creating an applicant file of data to produce the DD Form 1966 was determined and functional requirements for a computer printer which would generate the form were specified. Interactive computer dialogues, and complementary printout capabilities for the DD Form 1966 were developed, tested, and demonstrated.

To enable Army automation specialists to begin the task of porting the software to the operational system, documentation was expedited and delivered to ARI and USAREC concurrently. A program for an experimental "free-form" automated Application for Enlistment (capturing the same personal data as the DD Form 1966 but printing out on plain paper) was also developed and documented.

Sales technology is an integral, indeed vital, part of the recruiting process. The objective in this fourth task area was to develop an automated methodology for assessing enlistment motivation. Interviews with recruiting managers led to selection of the paradigm already employed in the Army recruiter's sales package as the basis for instrument development.

The product of this task area is a formalized technology for employing an established Army sales technique. When automated, and incorporated into a recruiting sales presentation on the JOIN system, this instrument would determine an applicant's dominant buying motive (DBM), i.e., the strongest motives toward enlistment. This information would allow the recruiter to focus on the appropriate features and career benefits, quickly tailoring the sales presentation to the applicant.

A fifth and final task area was that of personnel assignment. Using expertise gained in the development of NPAS and other automated assignment systems, coupled with information gained through interviews with Army recruiting and automation managers, the research team determined the required developmental efforts to interface the JOIN system with extant and planned Army-wide computer networks. This study highlighted the problems inherent in developing an assignment-prediction system, one that would provide information on entry level job openings and produce more optimal person-job matches.

## Conclusions

This inter-laboratory cooperative effort proved very fruitful. Several applicant assessment instruments were developed, tested, and refined. One is already in operational use in the recruiting milieu, and in itself represents a milestone in computerized adaptive testing. Other instruments are being evaluated for further validation and possible implementation. Studies that are critical to the design of recruiting compatible automated systems were completed and their results made available to the research and operational communities through a number of working papers, technical reports, professional papers, and journal articles. A number of software packages were completed.

Necessary groundwork has been accomplished in a wide variety of research areas: on this base, the Army can develop enhanced capabilities for the JOIN system more rapidly. Because the recruiting operations of all the armed services are similar, products of this effort have high potential for further inter-service technology transfer. Endeavors such as this are exemplars of a developing trend within military personnel psychology R&D: maintenance of service-specific expertise bases, allied with decreasing research parochialism (Wiskoff, 1985). In sum, this project dramatically demonstrated the value of inter-service cooperation, an excellent example of the leveraging of the government research dollar.

## References

Alley, W. E. (October 1978) Vocational Interest-Career Examination: Use and application in counseling and job placement. Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Baker, H. G. (May 1983) Navy Personnel Accessioning System (NPAS): II. Summary of research and development efforts and products. (NPRDC SR 83-35). San Diego: Navy Personnel Research and Development Center.

Baker, H. G. (January 1984) Computerized Adaptive Screening Test: Development for use in military recruiting stations (NPRDC TR 84-17). San Diego: Navy Personnel Research and Development Center.a

Baker, H. G. (February 1984) Computerized vocational guidance (CVG) systems: Evaluation for use in military recruiting (NPRDC TR 84-21). San Diego: Navy Personnel Research and Development Center.b

Baker, H. G. (July 1985) Designing a vocational guidance system for military recruiting: Problems and prospects - I. Organizational and operational considerations (MPL TN 85-7). San Diego: Manpower and Personnel Laboratory, Navy Personnel Research and Development Center.a

Baker, H. G. (1985) A prototype computerized vocational guidance system for Navy recruiting. Journal of Computer-Based Instruction, 12(3), pp. 76-79.b

Baker, H. G., Rafacz, B. A., & Sands, W. A. (May 1983) Navy Personnel Accessioning System (NPAS): III. Development of a microcomputer demonstration system (NPRDC SR 83-36). San Diego: Navy Personnel Research and Development Center.a

Baker, H. G., Rafacz, B. A., & Sands, W. A. (May 1983) Initial development of a computerized adaptive screening test (CAST) for use in military recruiting. Paper presented at the annual conference of the International Personnel Management Association (IPMAAC), Washington, DC, 22-26 May 1983.b

Bryan, J. D. (October 1982) Recruiting evolves toward tommorrow. All Volunteer. Fort Sheridan, IL: U.S. Army Recruiting Command.

Diamond, E. E. (July 1985) Development of the career maturity assessment (MPL TN 85-7). San Diego: Manpower and Personnel Laboratory, Navy Personnel Research and Development Center.

Holland, J. L. (1973) Making vocational choices: A theory of careers. Englewood Cliffs, NJ: Prentice-Hall.

Sands, W. A. & Gade, P. A. (1983) An application of computerized adaptive testing in U.S. Army recruiting. Journal of Computer-based Instruction. 10, (3&4), 87-89.

Sands, W. A., Gade, P. A., & Bryan, J. D. (1982) Research and development for the JOIN system. Paper presented at the 24th annual conference of the Military Testing Association, San Antonio, TX, 1-5 November.

Sands, W. A. & Rafacz, B. A. (1983) Computerized adaptive testing (CAT) for the U.S. Army JOIN system. Paper presented at the 25th annual conference of the Military Testing Association, Gulf Shores, AL, 24-28 October.

Wiskoff, M. F. (August 1985) Military psychology and national defense: Making an difference. Division 19 Presidential Address delivered at the 93rd annual convention of the American Psychological Association, Los Angeles, CA, 23-27 August.

# Vocational Interests as Predictors of Army Performance[1]

Hilda Wing
U.S. Army Research Institute

Bruce N. Barge and Leaetta M. Hough
Personnel Decisions Research Institute

In: Elements of a Military Occupational Exploration System
Military Testing Association, October 1985
San Diego, California

Measures of vocational and occupational interest have been used in selection for Army enlisted occupations for many years. In this paper we will describe how such measures have been used in the recent past, review current Army research which will link such measures to performance in Army jobs, and identify critical issues that must be resolved in order for interest measures to be effective in a selection and classification program.

## The Army's Use of Interest Measures in Selection/Classification

Use of vocational interest measures for classification into Army training was part of the Army's selection and classification for enlisted personnel from 1958 until 1980. The Army Classification Batteries, followed by the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7, included forms of the Army Classification Inventory (ACI), which contained sentences describing activities with which an applicant could agree or disagree. Four scale-scores were obtained from each applicant: Combat, Administrative, Mechanical, Electronics. These scale-scores were incorporated with ASVAB cognitive ability subtest scores to produce Aptitude Area (AA) composites. For example, the Combat AA included both ability and interest measures. Empirical data supporting this use had been provided by the developers of ACB-73 (Maier & Fuchs, 1972). Interest measures were dropped from the Army enlisted classification system with the introduction of new ASVAB forms in October of 1980.

The Army's current Project A is, among other things, the largest selection and classification research effort to date. The initial function of Project A is to validate the ASVAB against Army performance. An additional aspect of Project A's mission is to develop new predictors which will cover attributes that the ASVAB does not. ASVAB is more than adequate for selection into Army training (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984). What we are more concerned about is classification and, in addition, performance on the job, successful completion of the first tour, and reenlistment eligibility. To that end there has been

---

[1]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 Deparment of the Army.

developed an evolutionary model of predictor space. This model conceives of predictor space as having three components. First, the cognitive-perceptual component includes measures of verbal, quantitative, and spatial abilities. Next, the perceptual-psychomotor component includes perceptual speed and accuracy, short-term memory, multi-limb coordination, and movement judgment. We have developed a mini-battery for this component which is administered on an IBM-compatible microcomputer with a custom-designed response pedestal. Finally, the non-cognitive component covers both biographical/temperament (personality) and vocational interest measures.

The evolution of this model of predictor space has been firmly anchored to data, as follows. Project A has so far completed research on a first cohort of Army enlisted personnel, those entering in FY 1981 and 1982. The second cohort includes those soldiers who entered the enlisted service during FY 1983-1984. It includes both longitudinal and concurrent components; the longitudinal is included in the concurrent. For the longitudinal effort, we developed our first test battery, the Preliminary Battery, from readily available, off-the-shelf paper and pencil measures of cognitive and non-cognitive attributes. We administered the Preliminary Battery prior to training, to soldiers in four selected MOS, from October 1983 through June 1984. This year we obtained measures of training success and early attrition for this sample.

This summer, we are testing the concurrent component of this 1983-84 cohort with a second, new battery, in conjunction with a full complement of performance measures. We have added another 15 MOS, and the perceptual-psychomotor component of predictor space is being evaluated with micro-computers. Data collection should be complete by late November, 1985. While we have no analyses completed for what we are calling the Trial Battery, we do have some information about its immediate fore-runner, the Pilot Trial Battery.

A complete longitudinal effort is planned for the FY 1986-1987 cohort, to begin sometime next year. There will be the Experimental Battery, which will be much like the Trial Battery, to be administered to soldiers entering training in each of our selected MOS. Subsequently, we will be administering the appropriate performance measures to these soldiers. At the same time, we also plan to evaluate the performance of second-tour members of our 1983-1984 cohort.

Results for the Preliminary Battery

The Preliminary Battery included the Air Force Vocational Interest Career Examination (VOICE), which assesses 18 basic interests (Alley & Matthews, 1982). Because of the research on the Holland hexagonal model of vocational interests, we investigated its appropriateness. We factor-analyzed both the items and scales of the VOICE. We were able to recover the 18 basic interest scales quite nicely from the item factor analyses (Hough, Dunnette, Wing, Houston, & Peterson, 1984). We were able to find the Realistic group of occupational interests, but all the others clumped mostly into one group. In hindsight this made perfect sense. The majority of occupations in the enlisted military service are Realistic in

nature, as they are jobs in the skilled trades. There are a handful of Investigative occupations, some Conventional, and some Social occupations. For virtually no occupation in the enlisted ranks does the Artistic or Enterprising label fit.

What evidence was there of criterion-related validity for these interest measures? Available criteria were existing training grades and early attrition (status as of December 1984, or an average of one year of service). For training, the cognitive tests of the Preliminary Battery appeared to have some predictive power, although the coefficients were not large and not much larger than those obtained for the ASVAB. The attrition analyses are currently incomplete. This criterion will be especially hard to predict because the early attrition was fairly low, about eight percent. While some of the VOICE scales were significantly related to attrition in each of the four MOS, the correlations were quite low. The coefficients for some of the biodata/temperament scales, which evaluated aspects of socialization, were higher than those for interests. The domain of causes for discharge in the Army extends from "disciplinary" through 'for good of service" to "unsuitable unknown." It is likely that early attrition in the Army, particularly that through the Trainee Discharge program, may be more disciplinary than anything else. Thus, the predictiveness of the socialization scales is understandable.

The VOICE scales were not related to any great extent with the other measures evaluated, including the ASVAB. It is likely that as various criteria mature (later attrition, re-enlistment) or are administered as part of the Project A data collection (commitment, effectiveness), these early measures of vocational and occupational interests will have a better chance to demonstrate what they can do.

Results from the Pilot Trial Battery

The Pilot Trial Battery was field tested during the fall of 1984. Soldiers supplied data to evaluate the properties of the battery, including test-retest stability. We called our interest measure here the "Army VOICE," or AVOICE. We obtained this by starting with the VOICE, cutting back items on most of the 18 scales while adding scales for Army interests which are not duplicated in the Air Force, such as Infantry, Armor/Cannon, Science/Chemical Operations.

Psychometrically, the new instrument worked well, except that the factor analyses yielded the same pair of factors as before. For the Pilot Trial Battery, these factors appeared to be described better as "Combat" and "Combat Support," rather than "Realistic" and "Non-Realistic." This is a matter of taste rather than substance, as there is confounding of terms. The Combat occupations are Realistic while the Combat Support occupations cover the other five corners of Holland's hexagon. But, this is, we judge, the occupational reality of the Army enlisted world. The reliability and stability of the interest scales were excellent, in the .80's and .90's. There were no performance criteria available for this sample, but we did inspect the overlap of the interest measures with the remaining components of the Pilot Trial Battery and the ASVAB. The intercorrelations between AVOICE scales and other scales were generally low.

Preparing the Trial Battery from the Pilot Trial Battery consisted mainly of cutting back, so that a 6-7 hour battery was reduced to one requiring less than four hours. The AVOICE in the Trial Battery being administered now includes 176 items and takes about 15-20 minutes to administer. It will provide scores for interests in 22 Army occupations.

Issues in the Operational Use of Interest Measures in Selection and Classification

We see at least five major issues to be confronted in determining when and how to use measures of vocational interests in selecting and classifying for military enlistment. The first four are clearly technical while the last is more of a policy issue which can be informed by our technology.

First, the complete hexagonal model of Holland's vocational interest theory appears to be inappropriate for predicting performance in Army occupations. We tend to forget the context-sensitivity of models. The domain of Army jobs maps onto only a portion on the theorized hexagonal interest space, mainly that corner called Realistic. All of the other Army jobs, which could be characterized as involving interests from the Investigative, Social, and Conventional corners, appear to clump together. At this time we do not know whether this simple differentiation will provide all the predictability possible, given the available criteria, or whether further distinction into occupational scales will be warranted. But, it is clear that approaches using a complete Holland model will have limited applicability for the spectrum of Army enlisted occupations.

Second, the selection of appropriate criteria for vocational interests to predict is a major concern. Should criteria be those we consider as maximal effort, such as job knowledge tests and hands-on measures? Or should they be typical effort types of measures, such as motivation? We really need to know more about these criteria. One of the goals of Project A is to improve our conceptual understanding of the criterion space. This is clearly a worthy and necessary goal.

Third, how should predictors and criteria be used? The primary function of any interest measure is to direct the individual towards some occupations and away from others. That is, the object is classification. Regardless of the specific criteria, there are questions about the form of the predictors to use. Should we use scores from occupational scales, or should we use factor scores? Should we use single scores, or do we need to investigate configurations, or profiles? How should we combine interest measures with measures from other domains, such as the cognitive? It could be that positive interest in a specific area can compensate, to some extent, for lower ability for that area (Matthews, 1982). What are the characteristics of the sample sizes, the psychometric properties of the measures, that must be present for us to be able to make any kind of definitive statement concerning such claims?

Fourth, what exactly are we trying to predict: Success or avoidance of failure? This is the more complex issue concerning the fact that the

Army, and perhaps most employers in general, cannot always use people in what those people are best at. For example, one of the MOS in Project A is the Combat Medic. We have administered a complete battery of performance measures to several hundred Combat Medics so far in addition to the Trial Battery. However, at this point in time the United States is currently not in any general armed conflict, and there is little opportunity for these soldiers to practice their training in any realistic environment. Some of them may be working in maternity wards while others spend most of their time in the motor pool. We find it difficult to understand exactly how an interest in medical activities, absent other information, will be predictive of important criteria for these soldiers. Other examples are possible. How should interest measures be used in such cases?

The fifth and final issue concerns where in the enlistment process is it most appropriate to use interest measures? In the All-Volunteer Army, they may be more appropriately used by the recruiter. Should they be used in a mandatory or advisory way? Perhaps this is a technical question as much as are the other four: Are interest measures more predictive, of whatever criteria we can come up with, in whatever psychometric fashion determined effective, when these measures are used in an advisory fashion rather than a mandatory one?

This report has provided a brief description of how the Army is investigating the use of vocational interests in predicting performance in Army jobs. Project A will be providing vast amounts of data which will better inform our use of these measures. However, this use may be complex. The empirical data will, we trust, point us towards better use.

## References


Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination: A description of the instrument and possible implications. Journal of Psychology, 112, 169-193.

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J., & Peterson, N. G. (1984). Covariance analyses of cognitive and noncognitive measures in Army recruits. Paper presented at the convention of the American Psychological Association, Toronto, Ontario, Canada.

Maier, M. H., & Fuchs, E. F. (1972). Development and evaluation of a new ACB and aptitude area system (Technical Research Note 239). Alexandria, VA: U.S. Army Research Institute.

Matthews, M. D. (1982). Vocational interests, job satisfaction, and turnover among Air Force enlistees. Paper presented at the Fourth Annual Learning Technology Congress and Exposition, Society for Applied Learning Technology, Orlando, FL.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative Armed Services Vocational Aptitude Battery (ASVAB) area composites (Technical Report 651). Alexandria, VA: U. S. Army Research Institute.

# Army Career Vocational Guidance As A Recruiting Tool

Allyn Hertzbach, Deirdre J. Knapp, and Richard M. Johnson
U.S. Army Research Institute for the
Behavioral and Social Sciences
Alexandria, Virginia 22333-5600[1]

What is Army recruiting doing in the vocational guidance business? Surely, the Army can find more efficient ways to meet its manpower needs. This point of view is not unreasonable and might find much support, especially among the other services. But the Army is considering the feasibility of providing career vocational guidance (CVG) via the JOIN system. This paper addresses the overall approach currently being pursued by the Army Research Institute (ARI) and the Army Recruiting Command (USAREC). First, the rationale for the project is discussed, and a description of the components of the vocational guidance package follows. The feasibility research now being conducted with the American Association of Community and Junior Colleges (AACJC) is briefly described and is followed by a look at what future research might be pursued.

## Program Rationale

Anyone familiar with career vocational guidance knows that automated career information delivery systems (CIDS) are available in many schools and employment centers. There are states that have such systems available at little or no cost to residents. In fact, the military services are considering providing software that would allow high schools to use the Armed Services Vocational Aptitude Battery (ASVAB) as the ability measure for the most often used automated delivery systems. There is even some consideration being given to developing a DoD CIDS that could be deployed across the country to encourage young people to take the ASVAB and to consider joining one of the military services.

The Army began planning to develop its own service specific vocational information system before DoD began to examine the possibility of providing its own program in career vocational guidance. And the Army is still considering the feasibility of an automated system of its own. The main reason for continuing this work is simply that the Army has the most difficulty of all of the services in meeting its manpower needs. Another compelling reality is that the recruiting environment is becoming increasingly competitive in that the Army has to vie with the other services, as well as an increasing number of private sector employers. A third reality is that the pool of the nation's young people is shrinking and will not begin to increase until the mid 1990s. Finally, the Army has the automated means (JOIN) to prepare a tailored sales package that would greatly benefit its recruiters, not to mention

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army

other possible deployments of the system, as in high schools or two year colleges.

Aside from the sales advantages to be derived from improving the JOIN sales package with career information and vocational measures, the Army recruiter would be able to direct or funnel applicants more systematically by using occupational groupings rather than being asked about specific jobs and and having to soft pedal or evade the issue of specific jobs or particular kinds of job training. The Vocational Guidance package would have available to it groupings of skill clusters developed by the Army Recruiting Command, rather than particular military occupational specialties, so that the prospect could view more generalized job information. Therefore, the applicant could see a realistic preview of relevant jobs via JOIN videodisc presentation without also being predisposed to a particular job, upon which he might predicate enlistment. If the other services and DoD approved, the Army might also be interested in including the Military Occupational And Training Data (MOTD) as another resource for its vocational guidance package, which collapses military jobs across the services and does a crosswalk with civilian jobs.

## Anatomy of the Army Career Vocational Guidance (CVG) Package

The CVG package being developed for the JOIN system will be compatible with the current JOIN software, replacing some components and adding other components. A brief description of the JOIN system is included for those readers unfamiliar with the system. The Joint Optical Information Network (JOIN) is a Z80 based microprocessor with 64k RAM, I/OS operating system with two double-sided five and one quarter inch floppy disc drives. The JOIN system includes a keyboard, color monitor, dot matrix printer, modem, and videodisc player. The videodisc player is under random access control through the JOIN system. The system is employed by a recruiter with a prospect to determine the individual's needs, interests, wants, as well as to determine the prospect's likelihood of qualifying for enlistment. The system also presents information about Army jobs, benefits, and enlistment options. The JOIN system is currently deployed at almost all Army recruiting stations.

The Career Maturity Assessment (Diamond, 1984) is being considered for inclusion in the Army's CVG package. The instrument was developed for the Naval Research and Development Center and provides a score that indicates whether an individual is ready to make reasonable career decisions or not. Though there is disagreement about what the actual career maturity dimensions are, Super (1974) suggests that career maturity is the readiness to make reasonable career decisions at given decision points in career development. He describes career maturity as the congruence between a person's vocational behavior and the behavior expected at that person's age. Diamond (1984) reports that self-knowledge, self-concept, and decision making skills appear to be the common elements in research aimed at describing and explaining career maturity. Career maturity information can be very useful for both the prospect and the recruiter. This information should be valuable to the recruiter and help him to recognize less stable decision making potential. The

recruiter might want to take more time to clarify needs and interests with individuals who do not meet the cutoff score on the instrument and expect more fluctuation or poorer selections from "immature" individuals. "Mature" individuals, of course, are not immune from mistakes or fluctuations, especially at the ages when Army recruiters are likely to see them.

Two interest inventories are being considered for inclusion in the CVG package. Both of these instruments are currently being developed by ARI. One is an in-house effort, the Job Interest Survey(JIS), and the other, the Vocational Interest Profile (VIP), is an ARI contract effort (Faust, Unger, and Schmitz, 1985). Test development for these instruments continues and is not yet far enough advanced to speculate about which will be selected. One possibility being considered is combining the best items of the two tests and automating the result. Thus far, only preliminary testing of the instruments has occurred (Faust, Hertzbach, and Knapp, 1985) with a small sample of 160 soldiers. This pilot effort was implemented to work out the administration, to test the reading level of the instruments, to determine how long it took soldiers to take the test, and to determine the clarity of the instruments. The two instruments were administered to new recruits at Army reception stations as part of the 1985 New Recruit Survey. Thus, more extensive analyses will be performed to determine the potential usefulness of the instruments.

The next element of the CVG is job information, and there are a number of possibilities. The JOIN system is particularly well equipped for providing large reference resources. With a few 5 1/4" diskettes very large amounts of information can be made available to the prospect and recruiter. There are several sources of information which could be useful, but the most important consideration, short of accuracy and comprehensiveness of the job information, is the policy that the recruiter cannot, indeed, must not, sell specific jobs. With this restriction in mind, the skill cluster groupings developed by USAREC and Career Management Fields (CMF, groups of MOSs) are more appropriate than is a directory of all of the specific Army jobs (MOS). Another excellent source of job information is the Military Occupational and Training Data software developed for joint service use. The difficulty encountered in using this software is the number of diskettes that would be required to add it to the package. Increasing the number of diskettes makes the package a bit more awkward to use, not to mention increasing the chances for losing or impairing the materials. Despite these considerations, the MOTD software has many advantages and with the approval of the joint service oversight committee could aid a great deal of useful information to the Army's CVG package. The available software for including realistic and comprehensive job information is an important part of the CVG and should greatly enhance its sales utility.

Aside from job information, there are a number of incentives and Army programs about which prospects should be informed. JOIN is an excellent medium for presenting this kind of information. There are many varieties of information that can be tailored for specific audiences. If the prospect is a college student or has some college credit, information about advanced grade for college credit or the Army College Fund or opportunities to become an officer (OCS and ROTC) should be presented. This part of the package furthers the sales dialogue and begins to focus on the reasons for joining and the issue of joining. With the acquisition of information and prequali-

fication (via the Computerized Adaptive Screening Test (CAST), the JOIN ability measure that predicts Armed Services Vocational Aptitude Battery (ASVAB) performance), a well-informed Army recruiter and the prospect can begin to discuss the tradeoffs for enlistment. If the prospect is likely to qualify mentally and physically, the general information can be personalized and presented in the most attractive way possible. The recruiter knows what job interests and needs the individual has and can link these to available incentives and programs. This kind of interaction could also largely be automated and allow the prospect to go through the package alone at his school counseling office.

## CVG Feasibility

Now that the rationale and the components of the Army's CVG program have been presented, the practical consideration of whether or not this package can be deployed beyond the recruiting station becomes important. The Army needs to get to the high quality market and not just wait for that market to walk into the recruiting station. One market that the Army has never systematically approached is the two year college market, and these schools seem to be an excellent place to develop and test the feasibility of the Army's CVG. ARI currently has a contract with the American Association of Community and Junior Colleges(AACJC) to determine if the CVG can function to the mutual advantage of the Army and the two year colleges. If the CVG is seen as appropriate, the package could also be employed, with some modification, in high schools, as well. However, the primary purpose for the CVG is for the Army recruiter in uncovering dominant buying motives and to structure the prospect's job expectations (without selling a particular job) and to match the prospect's needs and interests with Army opportunities.

Briefly, the goal of the contract effort with AACJC is to develop six sites at junior colleges to test and develop the CVG. Attempting to get an estimate of the potential market, targeting advertising to the market, discovering dominant buying motives, and identifying a tailored sales approach are particular aims of this work. JOIN systems are to be provided to each site, including the appropriate software and other materials. The best methods for implementing the CVG are also to be determined--stand alone or with the college counselor. Of interest also in this research effort is identifying barriers to enlistment and most effective ways of stating tradeoffs for the prospects and the recruiters.

This research effort is more an exploratory and qualitative project than a tightly designed experiment. Aside from the aforementioned purposes of the research, there is a need to establish good rapport with the two year schools, as well as with the AACJC. The schools and the AACJC need to understand that we are not trying to compete with them for their students, rather that we are providing ways for students to finish school and enabling them to further their education (Army College Fund, GI Bill). Having the schools collect and provide critical commentary on the CVG components, testing the CVG's reception at the schools, and learning how best to interface the Army recruiter with the schools are qualitative issues that are also being addressed. This project will be completed at the end of fiscal year 1986.

## Discussion

The Army Career Vocational Guidance research project will assess vocational interest and, perhaps, career maturity; assess ability (CAST); provide realistic military occupational information; provide incentive and program information; and articluate clear and attractive tradeoffs to encourage prospects' decisions to enlist. The strategy for implementing this program is to make it available to Army recruiters on JOIN and to provide it to institutions (e.g., two year colleges and high schools) for the benefit of their students. The students will receive vocational information about themselves, as well as information about the Army opportunity. They will also learn about military occupations and the training required for these occupations. This service will be provided at no cost to the schools or the students. The Army recruiter will receive feedback in some form from the schools using these packages. Perhaps, schools will send prospect lists of students who express some interest in learning more about the Army.

The actual viability of an Army CVG is, at this point, undetermined. There is no question, however, that the Army needs to explore all avenues that might produce high quality prospects in this difficult recruiting environment. However, this effort can only be made to good effect with the cooperation of the other services. If there is an all service effort, the Army CVG program will, of course, need to be made compatible with whatever is implemented, including the possibility of scrapping any school deployment. However, the information that is learned from our current effort with the AACJC should be useful in both the Army's effort, as well as the all service, DoD effort.

As previously suggested, the future of the Army Vocational Guidance package depends in large part on it viability in the schools and to the extent that it can be coordinated with all service efforts. But, assuming that the program is viable, the next phase of the development would be to design a comprehensive empirical research effort to test the hypotheses developed from the current effort with the AACJC and build the package, implement it, assess its effectiveness, fix it, and deploy it as policy makers deem appropriate.

## References


Diamond, Esther E. (1984). Development of the Career Maturity Assessment (Draft Report). San Diego, CA: Naval Personnel Research and Development Center.

Faust, D., Unger, K. & Schmitz, E. (1985). Development of a Prototype Army Vocational Interest Inventory for the Enlisted Personnel Allocation System (Tech Report). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Faust, D., Hertzbach, A. & Knapp, D. (1985). Field Trials of Two Vocational Interest Assessment Techniques: The Vocational Interest Profile (VIP) and Job Interests Survey(JIS) (Unpublished Report). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Super, D. E. (1974). Measuring Vocational Maturity for Counseling and Evaluation. Washington, D.C.: National Vocational Guidance Association.

Weltin M. (1984). Job Interests Survey (unpublished technical papers). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

# Developing New Attribute Requirements Scales for Military Jobs

Elizabeth P. Smith
U.S. Army Research Institute[1]
5001 Eisenhower Avenue
Alexandria, Virginia 22333-5600

Conducting empirical validity investigations to predict job performance is not always feasible. Even when empirical approaches are undertaken, such as the ongoing ARI Project A to improve the selection, classification and utilization of enlisted personnel, it is rarely possible to include all jobs within an organization. Given the complexities of empirical validation, it is necessary to develop other methods for matching people to jobs and optimizing their performance.

One approach is to obtain rational estimates of the human attributes (i.e., abilities, characteristics, and interests) which are required for successful job performance. When gathered systematically from qualified judges, these estimates can be summarized as profiles of required attributes. Then, measures of individuals' attributes can be matched to such profiles for selection and classification purposes. In addition, knowledge of required attributes is potentially useful for (a) designing new systems and training programs that are within the capacities of available personnel and (b) generalizing empirical validity data to new and different jobs, by grouping them on the basis of similarity of attribute profiles (Fleishman, 1982; Pearlman, 1980). The latter application is especially pertinent to the Army's Project A, which is collecting validity data for only 19 Military Occupational Specialties (MOS).

A well-researched method of determining ability requirements is the rating scale approach developed by Fleishman and his associates (see Fleishman & Quaintance, 1984 for a comprehensive summary), based on a taxonomy of 40 cognitive, perceptual, physical and psychomotor abilities. With these scales, a rater decides if an ability is necessary for errorless job performance, and, if so, estimates the level required on a 7-point, behaviorally-anchored scale.

Early outcomes from Project A provided an opportunity to develop a new set of rating scales based on a new taxonomy of human attributes. An expert judgment task (Wing, Peterson, & Hoffman, 1984) obtained estimates of validity for 53 predictors against 72 criterion constructs from 35 personnel psychologists. Factor analysis of the data yielded 21 clusters of the 53 cognitive, perceptual, psychomotor, temperament and interest predictor variables. A predictor test battery based on these 21 clusters has been developed and is being validated. The purpose of this paper is to discuss the initial construction and testing of a new set of scales for estimating job requirements which is based on these 23 clusters (hereafter called "attributes"). As more data become available, it is expected that the taxonomy of predictors (and test battery) may change. The rating scales will be revised to reflect these changes.

A set of scales based on the Project A taxonomy has several potential advantages over the Fleishman ones. The most salient feature is that obtained profiles of attribute requirements will directly correspond to

[1]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.

Project A validity data. It will include temperament and interest measures that are not among the Fleishman scales and will not include those attributes/abilities for which no predictor tests are given. Additional benefits (e.g., lower cost, more efficiency) may be possible with this set of scales. It was designed to be used by work supervisors rather than personnel psychologists and contains primarily Army-specific behavioral anchors with only about half as many attributes to rate as Fleishman's.

For any rating scales to be useful in practice, they must give reliable and valid scores. The effort reported here examined issues related to the reliability of the ratings. Validity investigations will occur later. The following issues were examined here. First, how closely do raters agree, i.e., how high is interrater reliability? Second, how well do the scales differentiate across attributes (i.e., yield non-flat profiles) within a job and across the attribute profiles of different jobs? Finally, can the scales be used to identify attributes for which differences in level of the attribute most influence performance? For some attributes, higher levels may be required for better performance whereas for others, once a minimal requirement is met, having a greater amount of the attribute has no additional effect on performance.

## METHOD

Subjects. Thirty-six Non-commissioned Officers (NCOs) from the Cannon Crewman MOS and 39 NCOs from the Motor Transport Operator MOS, all males located overseas, participated as Subject Matter Experts (SMEs).

Instrument. The Attribute Assessment Scale, which was empirically developed for this research, consists of a set of behaviorally-anchored scales for 20 of the 21 attributes in the Project A taxonomy plus two additional attributes, Stamina and Physical Strength, which were thought to enhance face validity. A scale for Enterprising Interests was eliminated because it was impossible to generate items for this attribute which were sufficiently different from those falling under Self-Esteem/Leadership. The names of the attributes were modified from the original Wing, et. al. (1984) labeling for better comprehension by SMEs. The final instrument had one page per attribute. Below the definition at the top, there were three 7-point vertical scales, placed side-by-side, to enable three responses. A zero-point was added to indicate the attribute was not required at all. SMEs circled the number corresponding to the appropriate level for their job.

To construct the scales, comprehensive definitions for the attributes were developed so as to be readily understandable by people who were not trained in personnel research. A pool of items for potential anchors (i.e., behavioral statements) was generated. Ten items per attribute were ultimately selected, after screening by two to four other researchers. These were presented with the appropriate definition in an anchor-rating instrument. Initially, 26 NCOs from either the Administrative Specialist or Military Police MOS rated each item on the amount of the attribute represented by or needed for the behavior described. Items with mean ratings that were the highest, lowest, and closest to 4.0 (midpoint) that also had a standard deviation less than 1.5 were selected as scale anchors. Using these criteria, scales could be created for only 11 attributes.

After identifying difficulties related to (a) task comprehension, (b) response format, (c) failure of raters to differentiate effectively among items, and (d) a few of the definitions and items themselves, I revised the anchor-rating instrument and administration procedures, adding a

15-minute training period. This instrument was given to another sample of NCOs ($N$=28) from the same two MOS. From the second administration, using the criteria indicated above, three anchors were obtained for all but two of the attributes (Social Interaction and Stress Reaction for which only two anchors were selected) to form the Attribute Assessment Scale.

Procedure. SMEs rated the level of each of the 22 attributes that is required to perform Skill Level 1 (entry level) work under combat-readiness conditions in his own MOS for three performance levels: at the 15th, 50th, and 85th percentiles. In addition to the written instructions, SMEs received extensive training in how to complete the task, including a step-by-step demonstration of the actual rating process using the anchors as guides. Training and responses to questions took about an hour. Early ratings were checked to ensure comprehension of the directions before raters proceeded with the rest of the task. Ratings took about 30-45 minutes.

Analyses. Intraclass correlation coefficients (ICCs) were calculated from Raters X Attributes ANOVAs over all attributes and separately for the three major domains (i.e., cognitive/perceptual, physical/psychomotor, and noncognitive) for each of the three performance levels. The ICCs estimate the reliability of the mean ratings [$r(k)$; k=number of raters], an index of interrater reliability. Also, an MOS X Attributes X Performance Levels univariate repeated-measures ANOVA was performed.

## RESULTS

Eight Motor Transport Operators were eliminated from the analyses due to the logical inconsistency of their data. $R(k)$ coefficients over all attributes were, in increasing order by performance level, .75, .77, and .69 for Cannon Crewmen (k=36) and .74, .74, and .69 for Motor Transport Operators (k=31). For the domains, r(k) coefficients ranged from .61 to .79 across performance levels and MOS. There were two exceptions to this: Physical/psychomotor reliabilities were very low for both MOS at the 85th percentile [$r(36)$=.13; $r(31)$=.58] performance level.

None of the effects involving MOS for the MOS X Attributes X Performance Levels ANOVA were significant. There were significant main effects for Attributes [$F$ (2?,1365) = 6.98; $p$ = .0000] and Performance Levels [$F$ (2,130) = 398.35; $p$ = .0000] and a significant effect for the Attributes X Performance Levels interaction [$F$ (42,2730) = 2.51; $p$ = .0000]. Scheffe' comparisons between means within performance levels by MOS indicated significant differences between only the highest and lowest means, which ranged from 1.09 to 1.75. Means and standard deviations for all ratings are provided in Table 1.

## DISCUSSION

In comparison to the very high Intraclass Correlation Coefficients (ICCs) obtained by Fleishman and associates or those discussed by Rossmeissl (1985) within this symposium, the ICCs from this research are weak, especially since around 30 raters are needed to obtain coefficients of at least .60. ICCs are based on variance components. As such, low (or uninterpretable) reliabilities result if there is too great a between-subjects variance and/or too little within-subjects variance. The low reliabilities obtained here appear to be a function of both. Previous research on ability assessment has found mean ratings that varied from very low (even "Not required") to very high (7) across attributes. This was not the case here. The inclusion of three performance levels may have

Table 1

Mean and Standard Deviations of Ratings of Attribute Requirements for Cannon Crewman and Motor Vehicle Operator MOS at Three Performance Levels.

| Attributes | MOS[a] | Performance Level | | | | | |
|---|---|---|---|---|---|---|---|
| | | 15th Percentile | | 50th Percentile | | 85th Percentile | |
| | | M | SD | M | SD | M | SD |
| Cognitive | | | | | | | |
| Verbal Ability | C | 2.86 | ( .93) | 4.17 | ( .91) | 5.33 | (1.10) |
| | D | 2.87 | ( .85) | 4.32 | ( .85) | 5.32 | (1.08) |
| Memory | C | 2.44 | (1.18) | 3.89 | ( .89) | 5.53 | (1.11) |
| | D | 3.26 | (1.26) | 4.26 | (1.03) | 5.16 | (1.27) |
| Reasoning Ability | C | 2.78 | (1.31) | 3.94 | (1.17) | 4.89 | (1.33) |
| | D | 2.45 | ( .96) | 3.77 | (1.02) | 5.00 | (1.32) |
| Number Facility | C | 2.06 | (1.12) | 3.47 | (1.16) | 5.08 | (1.52) |
| | D | 2.48 | (1.18) | 3.90 | (1.30) | 4.94 | (1.48) |
| Mechanical Comprehension | C | 2.86 | (1.36) | 4.39 | (1.18) | 5.50 | (1.08) |
| | D | 2.97 | (1.33) | 4.23 | (1.14) | 5.20 | (1.00) |
| Information Processing | C | 2.50 | (1.08) | 3.58 | (1.30) | 4.81 | (1.37) |
| | D | 2.68 | (1.30) | 3.90 | (1.08) | 5.03 | (1.17) |
| Closure | C | 2.78 | (1.33) | 4.08 | (1.25) | 4.89 | (1.41) |
| | D | 2.68 | (1.42) | 3.87 | (1.45) | 4.61 | (1.67) |
| Visualization | C | 2.33 | (1.15) | 3.58 | (1.30) | 4.69 | (1.51) |
| | D | 2.29 | (1.30) | 3.42 | (1.43) | 4.26 | (1.84) |
| Perceptual Speed & Accuracy | C | 2.75 | (1.52) | 3.97 | (1.38) | 4.94 | (1.31) |
| | D | 2.97 | (1.33) | 4.16 | (1.29) | 4.71 | (1.40) |
| Physical/Psychomotor | | | | | | | |
| Physical Strength | C | 3.75 | (1.32) | 4.89 | (1.14) | 5.67 | (1.17) |
| | D | 3.58 | (1.39) | 4.61 | (1.20) | 5.32 | (1.25) |
| Stamina | C | 3.06 | (1.45) | 4.53 | (1.11) | 5.39 | (1.19) |
| | D | 2.68 | (1.30) | 3.94 | (1.41) | 4.84 | (1.63) |
| Multilimb Coordination | C | 2.80 | (1.47) | 4.20 | (1.21) | 5.26 | (1.40) |
| | D | 3.34 | (1.54) | 4.45 | (1.24) | 5.48 | (1.12) |
| Dextery | C | 3.00 | (1.35) | 4.47 | (1.08) | 5.50 | (1.08) |
| Non Cognitive | | | | | | | |
| Steadiness/Precision | C | 2.83 | (1.40) | 4.08 | (1.25) | 5.47 | (1.36) |
| | D | 3.06 | (1.29) | 4.52 | (1.09) | 5.29 | (1.07) |
| Social Interaction | C | 3.14 | (1.62) | 4.44 | (1.59) | 5.34 | (1.70) |
| | D | 2.58 | (1.71) | 3.74 | (1.44) | 4.65 | (1.70) |
| Stress Tolerance | C | 3.03 | (1.50) | 4.22 | (1.27) | 5.12 | (1.43) |
| | D | 3.10 | (1.47) | 4.27 | (1.34) | 5.27 | (1.20) |
| Conscientiousness | C | 2.66 | (1.24) | 4.09 | ( .89) | 5.31 | ( .99) |
| | D | 3.19 | (1.47) | 4.35 | (1.11) | 4.97 | (1.25) |
| Work Orientation | C | 2.91 | (1.46) | 4.29 | (1.18) | 5.54 | (1.17) |
| | D | 2.90 | (1.45) | 4.16 | (1.10) | 5.39 | (1.17) |
| Self Esteem/Leadership | C | 3.00 | (1.26) | 4.25 | (1.20) | 5.47 | (1.21) |
| | D | 2.48 | (1.55) | 3.84 | (1.37) | 5.00 | (1.41) |
| Athletic Ability/Energy | C | 2.89 | (1.35) | 3.92 | (1.16) | 4.94 | (1.19) |
| | D | 2.87 | (1.55) | 3.73 | (1.48) | 4.33 | (1.71) |
| Realistic Interests | C | 2.54 | (1.40) | 3.71 | (1.25) | 4.94 | (1.66) |
| | D | 2.48 | (1.12) | 3.61 | ( .88) | 4.61 | ( .99) |
| Investigative Interests | C | 2.00 | (1.43) | 3.44 | (1.61) | 4.61 | (1.78) |
| | D | 1.97 | (1.40) | 3.26 | (1.50) | 4.16 | (1.93) |

[a] C = Cannon Crewman

D = Motor Vehicle Operator (Driver)

had a strong, negative impact on these particular results. The demands of the task appeared to impose a unique kind of restriction in the range of possible ratings. That is, the effective range of ratings within levels covered only two or three points rather than the entire seven points. This outcome served to reduce within-subjects variability, as all ratings fell close together. Although SMEs were clearly advised not to respond according to belief that "better must mean more," the mean ratings suggest that a demand characteristic was created by the instructions to rate at three levels. The result was ratings of attribute levels which correspond to level of performance, with ceiling effects occurring at the highest level. These effects would explain the extremely low reliabilities for Physical/Psychomotor attributes at the 85th percentile.

The fact that attribute requirements were elicited for three performance levels also may have clouded the findings in another way and reduced interrater agreement, i.e., increased between-subjects variance. Although definitions were provided for the three performance levels, how the SMEs actually interpreted these definitions was unknown. SMEs may have had different interpretations of the attributes from our definitions as well as from one another. For example, their verbal reports seemed to indicate some tendency to interpret performance levels in terms of particular soldiers in their charge, rather than from a more general (and shared) view of job performance at a particular level. They also tended to rate attributes in terms of the characteristics of someone who performed at that level, rather than in terms of the actual requirements of the job. The performance criterion, then, was more ambiguous than expected, pointing out a clear need for a very specific definition of the criterion. It was apparent that understanding the task requirements -- what was meant by the performance levels and how to do three ratings at a time -- took more time and energy than actually doing the ratings. In short, the use of three performance levels may have made the task harder than was intended, and interfered with the SMEs' ability to rate true requirements.

Two other factors may have contributed to low interrater agreement. SMEs were not given written descriptions of what they were to rate. Instead they were asked to decide individually the nature and content of entry level work and, specifically, what it required in terms of attributes. Moreover, they were to rate the whole job -- all work within all duty positions -- and not just some specific task or set of tasks. This very broad scope allowed considerable opportunity for variance. As a result of personal experiences and/or selective memory, the SMEs could differ a great deal in what they were evaluating. Obviously, higher interrater agreement would be expected for narrower areas of consideration. In addition, some SMEs found the scale anchors frustrating rather than helpful. Raters appeared to have difficulty using anchors as reference points for comparing tasks within their MOS. Some tended to evaluate the job in terms of whether the exact tasks depicted were or were not an actual part of the job. With some anchors that depicted common soldier tasks, some SMEs had problems separating the overall soldier requirements from the specific job requirements. Thus, although very familiar behaviors were thought to be the best for illustrating a level of an attribute, this was not necessarily the case.

The results of the ANOVA indicate that attribute profiles for the two MOS are not significantly different. The effects that were significant, Attributes, Performance Levels, and their interaction, are most likely a function of the high statistical power related to the large number of

degrees of freedom, and so are not really meaningful. Despite this, the data provide some useful information. The minimal differences which do occur suggest that some differences (as well as similarities) between MOS may exist, but may be masked in the present research. In addition, rank orders of the magnitude of ratings were different for both MOS at all performance levels, again suggesting there may be some differences in patterns of attributes which need further examination. For instance, at the 85th percentile, Verbal Ability ranked tenth for Cannon Crewman but third for Motor Vehicle Operator, while Stamina ranked first and fifteenth respectively. If one were to select only the five variables with the highest ratings, the selection would be different for each MOS. However, the top five are not necessarily the most important attributes: They are ranked on level of required attribute and not on relative importance of the attribute.

In summary, NCOs appeared to understand, in general, how to use the set of scales constructed to rate job requirements. The requirement for three sets of ratings simultaneously, however, created some problems. First, the actual physical arrangement of the scales on the page confused people. Second, it seemed to impose limits on the magnitude of ratings assigned. Given the expanse of the criterion to be rated -- the entire MOS at Skill Level 1 -- and the limitations created by the design itself -- different performance levels -- the obtained indices of interrater agreement are reasonable.

These findings suggest that better reliability estimates could be obtained with fewer raters if SMEs were asked to rate requirements for a single performance level; i.e., to estimate the minimum level of an attribute required to perform the job successfully. Further, more reliable ratings may be obtained by changing to a generic set of scale anchors (e.g., very low,low, moderate, etc.) or otherwise replacing the present behavioral anchors and/or focusing raters' attention on evaluating a specific task, a well-defined set of tasks, or a written job description would yield better reliability coefficients. Elimination of the restriction in range of ratings which was created by including three performance levels, should yield better discrimination among the attributes within MOS, and differences in attribute profiles across MOS.

References

Fleishman, E. A. (1982). Systems for describing human tasks. American Psychologist. 30, 1127-1149.

Fleishman, E. A., & Quaintance, M. K (1984). Taxonomies of human performance. Orlando, FL: Academic Press Inc.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87, 1-28.

Rossmeissl, P. G. (1985, October). Computerized approaches for estimating ability requirements. Paper presented at the 26th Annual Conference of the Military Testing Association, San Diego, CA.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984, August). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada.

# METHODOLOGICAL PROBLEMS IN IDENTIFYING ABILITY REQUIREMENTS RELATED TO SOLDIER PERFORMANCE

Dr. Claude R. Miller

US ARMY TRADOC
SYSTEMS ANALYSIS ACTIVITY
(TRASANA)

## INTRODUCTION

The Training Effectiveness Analysis (TEA) Division of TRASANA recently completed a basic study of the procedures necessary to represent relationships between soldier variables and system performance in computerized combat models (1). The study was prompted by a concern for future system manning requirements. Census data indicate that the number of individuals available for recruiting is declining and will continue to decline well into the next decade (2). At the same time, the Army is undergoing a force modernization program that includes technologically advanced weapon systems that put a premium on highly skilled operators and maintainers. These concurrent processes indicate a need to identify soldier variables related to system performance so that the available manpower may be distributed among the various weapon systems in a manner that maintains or improves combat effectiveness.

Combat models may prove useful for estimating the extent to which relationships between soldier variables and performance affect combat effectiveness. However, prior to modeling, existing relationships between soldier variables and performance must be identified and mathematically described. The term "soldier variables" may refer to any number of physical or psychological attributes of an individual. Soldier variable data normally available, or easily obtained, include physical profiles, demographic descriptors, and results from the Armed Services Vocational Aptitude Battery (ASVAB). At present, the Army assigns soldier to Military Occupational Specialties (MOSs) on the basis of ASVAB scores and manpower needs. The usefulness of ASVAB scores to predict soldier performance in the field remains a matter of study. Supplemental and/or alternative selection and assignment procedures may be required in the future.

One possible supplemental/alternative approach is based on the work of Fleishman (3) who has studied the relationships between human abilities and task performance. Fleishman's work has suggested a taxonomy of cognitive, perceptual, and psychomotor abilities differentially related to performance on various types of tasks. It was in the context of Fleishman's work that an attempt was made in the TRASANA study to identify soldier variables related to system performance. This report summarizes some of the difficulties encountered in that effort.

## METHOD AND RESULTS

The referenced TRASANA study focused on Air Defense Artillery soldiers in MOS 16S, STINGER gunners. Subsequent to the referenced study, additional data were collected on five other Air Defense Artillery MOSs and two Field Artillery MOSs. Subject Matter Experts (SMEs) were asked to rate the extent to which each of 33 separate basic human abilities, taken from the work of Fleishman, was related to MOS critical task performance.

It was immediately obvious that having SMEs rate abilities required for each MOS critical task would impose an unreasonable workload on the SMEs that could produce results of questionable validity and reliability. For that reason, the survey instructions asked the SMEs to rate each ability in terms of how necessary the ability was to "successful performance in the MOS" without reference to specific critical tasks.

The format of the survey itself included:

1. The name of the ability (eg., Reaction Time)
2. A definition of the ability in the form of a question (eg., Does a gunner have to be able to react quickly to sights or sounds?)
3. A space to check "Yes", "No", or "Do Not Understand the Question."
4. A graphic seven-point rating scale. The scale included the integers from 1 to 7 and graphically indicated midpoints between integers. The only rating guidance provided on the scale itself was entered next to the numbers 1, 4, and 7. Next to the number 1 was the description "Ability Needed But Only To A Small Degree." Next to the number 4 was the description "Ability Needed To A Moderate Degree." Next to the number 7 was the description "Ability Needed To The Highest Possible Degree."

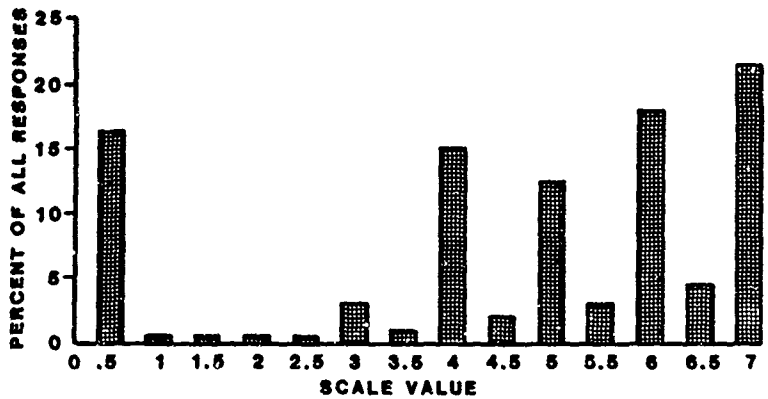
Prior to administration of the survey to SMEs, the survey was "pilot tested" with military personnel at TRASANA. Results from the pilot test indicated that many soldiers had difficulty understanding the question, i.e. the definition of the ability. Consequently, the definitions were rewritten using simpler language. Psychologists in the TEA Division were asked to compare the original and simplified definitions for comparability and generally agreed that the essence of each definition had been preserved. Nevertheless, opinion was not unanimous and some concern remained over whether all simplified definitions were accurate translations of the originals.

The final form of the survey was administered to a total of 85 SMEs from eight different MOSs. The number of SMEs for each MOS varied from n = 6 to n = 16. As stated previously, a seven-point rating scale was used for SMEs to rate how much of each ability was necessary to perform successfully in the MOS. Alternatively, the SMEs could check the space provided indicating that the ability was not needed at all. If an SME checked that an ability was not necessary, that response was given a scale value of zero.

Actual results in terms of how the abilities were rated for each MOS will not be addressed. Instead, the response pattern will be discussed. Given 85 SMEs rating 33 different abilities, a total of 2805 responses were possible. Of these, 219 (7.8%) were omitted, i.e. no scale response and no check that the ability was not needed. The remaining 2586 responses showed a distinct pattern that was very consistent across all MOSs. First, there was little use of the midway points graphically indicated on the scale. Only 12% of all responses were placed on the midway points. The preponderance of responses were given to the scale values 4, 5, 6, and 7. The scale values of 1, 2, and 3 were virtually ignored. The most frequently used scale value was 7 (22.7% of all responses). Figure 1 summarizes the response distribution pattern across the seven-point scale.

One final note regarding the pattern of responses seems of interest. The standard deviations among SME ratings for the highest rated abilities were consistently and significantly lower that the deviations among the lower rated abilities. The result was consistent for each group of SMEs surveyed.

Figure 1. Distribution of SME Responses to the Ability Rating Scales




DISCUSSION

The primary emphasis of the TRASANA study was to represent the relationships between soldier variables and performance in combat models using existing soldier variable data, not to identify a new set of soldier variables to explore. As a result, the relationships suggested by the SME ratings were not examined in detail, and validation of the SME ratings remains to be accomplished using actual performance data and scores from tests that measure the basic abilities.

Nevertheless, the experience gained from administration of the ability rating scales has implications for future scale development and interpretation. A major consideration is the workload put on the raters when ratings are required by specific tasks. To minimize the adverse impact of lengthy ratings, SMEs were asked to rate the abilities in terms of success in the MOS, not specific tasks. Although that procedure saves time, detailed information is lost. To recover at least part of that information, a group discussion was held with the SMEs after all ratings had been completed. From these discussions, it was possible to determine why certain abilities were rated low or high, and how certain abilities related to specific critical tasks.

A second consideration is how the abilities are defined. The SMEs had difficulty understanding the psychological terms used in many of the definitions, and psychologists had some difficulty with simplified definitions. To ensure common understanding of terms is a long standing principle of survey construction, but one too often assumed. Future scale development should take the time to develop sound and understandable definitions.

A final problem is the tendency of raters to judge an ability as needed to a moderate-to-high degree, or not needed at all. That tendency was noted on the first administration of the survey. In subsequent administrations raters were asked to save their highest ratings for those few abilities strongly required in the job. There is no evidence that such verbal instructions had any effect on the rating pattern. On two surveys, raters actually took their pencil and extended the scale out to 10 points, then checked "10." Use of anchor points may have reduced the tendency toward high ratings, but time constraints precluded development of verbal anchors. Another possible approach to the response pattern problem may be through the use of variance as well as mean ratings to establish the criterion used to identify an ability as needed for a job. Raters tend to agree on those abilities strongly related to successful job performance. The exact method for establishing criteria remains to be determined.

REFERENCES


1. Miller, C. R. Modeling Soldier Dimension Variables in the Air Defense Artillery Mission Area. TRASANA TEA-2-85, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, NM, February 1985.

2. "Soldier 90: Demographic Data And Trends," US Army Soldier Support Center, National Capital Region, Ft. Benjamin Harrison, IN, September 1982.

3. Fleishman, E. A. Toward a Taxonomy of Human Performance, American Psychologist. Vol. 30, 1975.

# Comparison of Weapon Systems Using Ability Requirements Scales

Jane M. Arabian
U.S. Army Research Institute for
the Behavioral and Social Sciences

The capability and complexity of weapon systems have been increasing. Meanwhile, census data indicate that there will be fewer 18-24 year olds between 1980 and the 1990's. The greater sophistication of military weapons together with the decrease in the size of the potential military applicant pool make it increasingly important not only to design systems that are within physical and cognitive capabilities of the available personnel but also to match correctly the skills and abilities of individuals to the skill and ability requirements of jobs (Shields & Baker, 1981). Inappropriate personnel classification could result in poor human resource management as well as a failure to benefit from the maximum functional capabilities of the weapon system. The best personnel classification system, however, will not be able to overcome system designs that require unavailable personnel attributes.

In order to optimize man-machine system performance, the engineering design process needs to be sensitive and responsive to the impact of the system on manpower, personnel, and training (MPT) issues. The MPT community, in turn, must provide information to the designers as early as possible in the system development process. The earlier in the process, the more likely it is that information will be based on subject matter experts' (SME) best estimates derived from their professional experience. Later on in the process, information can be based on empirical data obtained from the actual system. However, many of the design decisions which affect MPT occur before it would be possible to collect data with the actual system.

The Army Research Institute conducted a project to examine the Weapon System Acquisition Process (WSAP) and identify critical points in the process for MPT information (Promisel, et al, 1985). The project has been referred to as the Reverse Engineering Project because it traced the development and acquisition process in reverse, from the end-product back to the beginning (Concept Exploration Phase in the WSAP). Scientists working on the project studied as much available documentation, both historic and current, as possible for selected weapon systems.

The goal of the project was to examine how the performance of systems was affected by the specific system requirements and engineering and management decisions made earlier in the development process. For example, with regard to personnel requirements, project staff identified personnel requirements for the system as stated in the acquisition documents (e.g., where did the requirement come from and what kinds of data supported the requirement). Project staff also identified points in the acquisition process where it seemed feasible that more or different kinds of information could or should have been considered. Positive impact on system development from timely personnel information use was also noted. Recommendations were then made about how the acquisition process could be augmented to ensure that personnel requirements information would be developed and applied as early as possible in the WSAP.

The WSAP for Stinger, a man-portable air defense system, and its predecessor, Redeye, were examined first. One of the more salient findings was that very little personnel information was available throughout the acquisition cycle. Requirements for height and ASVAB scores were available, but there was little, if any, information about perceptual and psychomotor skill

and ability requirements. As the project staff became more familiar with the two systems through the requirements documentation, it seemed that the newer system would place more demands on the soldier than the older system. This seemed especially true if the new system's capabilities were to be fully realized. The new system did not appear to take the potentially increased psychomotor requirements into account; soldiers who operated the old system were expected to operate the new system effectively.

The research study described in this report was designed to provide information about the skill and ability requirements of the two weapon systems and to determine whether empirical support could be obtained for the impression that Stinger required more and/or higher levels of specifiable skills and abilities than Redeye. Since this type of information would be particularly valuable early in the WSAP, before the system would be physically available, the feasibility of using rating scales based on expert judgements to collect this information was examined.

## METHOD

Subjects. Two separate samples of subjects served as SME raters. The first sample (Scientist Sample) consisted of eight Army Research Institute research scientists who were participating in the Reverse Engineering Project. They were familiar with the WSAP requirements documentation and testing results for the Redeye and Stinger systems. The second sample of SME raters (Instructor Sample) consisted of ten military instructors for the Redeye and Stinger training courses. They had been experienced system operators before becoming training instructors.

Rating Instruments. Both samples of raters employed the Fleishman abilities taxonomy (Fleishman, 1975). The list of abilities appears in Table 1.

Table 1

Fleishman Ability Taxonomy

| | |
|---|---|
| 1. Oral Comprehension | 21. Visualization |
| 2. Written Comprehension | 22. Static Strength |
| 3. Oral Expression | 23. Explosive Strength |
| 4. Written Expression | 24. Dynamic Strength |
| 5. Memorization | 25. Trunk Strength |
| 6. Problem Sensitivity | 26. Stamina |
| 7. Originality | 27. Extent Flexibility |
| 8. Inductive Reasoning | 28. Dynamic Flexibility |
| 9. Category Flexibility | 29. Gross Body Equilibrium |
| 10. Deductive Reasoning | 30. Speed of Limb Movement |
| 11. Information Ordering | 31. Gross Body Coordination |
| 12. Math Reasoning | 32. Multi-Limb Coordination |
| 13. Number Facility | 33. Wrist-Finger Speed |
| 14. Fluency of Ideas | 34. Finger Dexterity |
| 15. Time sharing | 35. Manual Dexterity |
| 16. Flexibility of Closure | 36. Arm/Hand Steadiness |
| 17. Speed of Closure | 37. Control Precision |
| 18. Selective Attention | 38. Rate Control |
| 19. Perceptual Speed | 39. Reaction Time |
| 20. Spatial Orientation | 40. Choice Reaction Time |

An ability rating scale for each ability includes a definition of the ability and a rating scale anchored with behavioral examples. A rating of "1" indicated that the rater estimated that very little of a particular ability would be needed for a job or task; a rating of "7" indicated that a great deal of the ability was estimated as necessary. In addition to the scales themselves, binary decision flow diagrams have been developed (Malamad, et al, 1980). The flow chart leads the rater through a series of yes/no decisions which, when completed, results in specification of the abilities required to perform the job. Figures 1 and 2 provide examples of

the flow diagram and ability scales, respectively. The scale on the left uses military examples for anchors, as in the computerized assessment system described below; the scale on the right in Figure 2 uses non-military examples, as in the paper and pencil version of the ability assessment system.

Figure 1

Flow Diagram

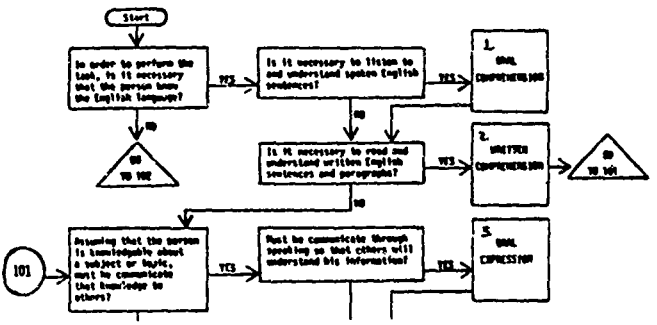
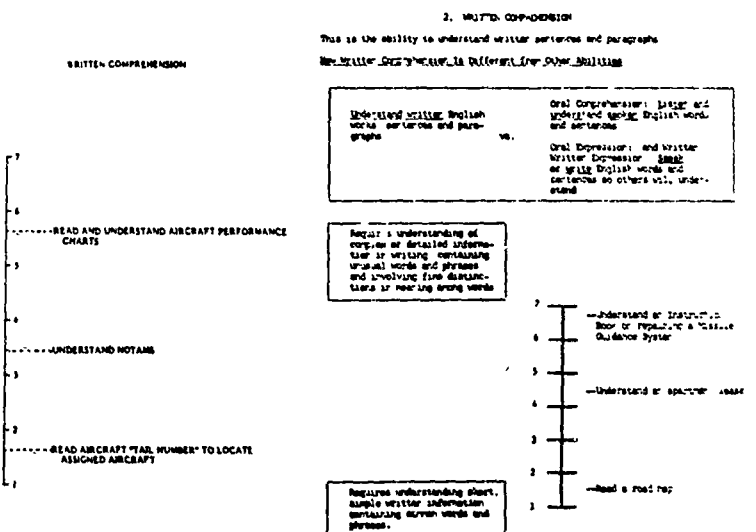



Figure 2

Ability Scales



Raters in the Scientist Sample employed a computerized version of the flow diagrams with ability scales (Rossmeissl, et al, 1983). This computerized ability assessment program was written in Applesoft© BASIC language and presented to the raters on an APPLE II microcomputer.

Raters in the Instructor Sample used a paper and pencil format to perform the ability assessments. The decision flow diagram was presented in one booklet with the ability scales in a separate packet. Raters worked through the flow diagram and when they came to a point for rating the level required of an ability they switched to the packet of scales, located the correct rating scale, made their rating, and returned to the flow diagram to continue the procedure. The scale anchors for the paper and pencil assessment were generic (non-military) behavioral examples (Theologus, et al, 1970).

Procedure. Ability ratings were obtained from each sample of raters separately. The computerized ratings were obtained from one subject at a time. The paper and pencil ratings were collected in one session with all raters present, but discussion of ratings was not allowed until all the assessments had been completed. For both samples, half the raters estimated ability requirements first for Redeye and then for Stinger; the order was reversed for the remaining raters. The raters in the Scientist Sample completed their estimates on two successive days (i.e., one system on Day 1, the other system on Day 2). The raters in the Instructor Sample completed their rating for one system, took a fifteen minute break, and completed their ratings for the second system.

Instructions for both groups were the same. All raters were asked to think about the soldier abilities that would be required during the successful operation of the system (Redeye or Stinger) under combat conditions. The decision flow diagrams and procedure for using the ability scales were explained. The raters were cautioned that the scales' anchor examples had not been validated for Redeye or Stinger and should therefore be used only as a rough guide. The raters were reminded to answer the decision flow questions based on their knowledge and/or experience with the weapon systems, to think about only one system at a time, and avoid making mental comparisons of the two systems.

Analyses. Given the relatively small sample sizes and exploratory nature of this research study, statistical analyses of variance were not performed on the entire data set from either sample. However, one-tailed correlated t-tests were performed on ratings for selected abilities if all raters in a sample agreed that the ability was required for both systems. Statistical analyses comparing the ratings of the two samples were not conducted since the background of the raters (scientists vs. instructors) was confounded with the rating mode (computer vs. paper and pencil).

## RESULTS AND DISCUSSION

The number of ratings and range of non-zero ratings for each ability and both systems are presented in Table 2. Sixteen abilities were rated for both Redeye and Stinger by at least half the raters in the Scientist Sample. Five abilities were not rated for either system by any of the scientist-raters. There was unanimous agreement in the Scientist Sample that seven of the abilities were required by both weapon systems. In no case was an ability consistently rated as required for one system but not for the other system.

The raters in the Instructor Sample tended to rate more abilities: every ability was seen as required for both systems by at least half of the raters. There was unanimous instructor-rater agreement that six of the abilities were required by both Redeye and Stinger.

Table 2

Ranges of Ability Ratings

| Ability | Scientist Sample (N=8) | | | | Instructor Sample (N=10) | | | |
|---|---|---|---|---|---|---|---|---|
| | Redeye | | Stinger | | Redeye | | Stinger | |
| | n | Range of Non-zero ratings | n | Range of Non-zero ratings | n | Range of Non-zero ratings | n | Range of Non-zero ratings |
| 1. Oral Comprehension | 5 | 5.0-6.0 | 5 | 5.0-6.3 | 9 | 3.0-6.7 | 8 | 4.3-6.7 |
| 2. Written Comprehension | 4 | 3.0-5.0 | 5 | 3.0-5.0 | 10 | 3.3-6.7 | 9 | 3.3-6.7 |
| 3. Oral Expression | 3 | 5.0-5.7 | 3 | 3.0-5.0 | 10 | 3.0-6.7 | 9 | 4.0-6.7 |
| 4. Written Expression | 0 | | 0 | | 10 | 3.0-6.7 | 9 | 4.3-6.7 |
| 5. Memorization | 8 | 5.0-7.0 | 8 | 4.0-7.0 | 10 | 2.3-6.7 | 10 | 1.7-6.7 |
| 6. Problem Sensitivity | 8 | 4.0-5.7 | 8 | 4.0-7.0 | 10 | 2.0-6 | 10 | 4.0-7.0 |
| 7. Originality | 5 | 3.0-4.7 | 7 | 3.0-5.7 | 10 | 1.7-5.7 | 9 | 1.3-5.3 |
| 8 Inductive Reasoning | 1 | 3.3 | 3 | 3.0-4.7 | 10 | 1.3-6.7 | 9 | 2.3-6.7 |
| 9. Category Flexibility | 3 | 3.0-4.7 | 3 | 2.0-6.0 | 10 | 2.0-6.7 | 9 | 3.0-6.7 |
| 10. Deductive Reasoning | 7 | 3.7-6.7 | 8 | 3.7-6.0 | 10 | 2.7-6.7 | 9 | 3.0-6.7 |
| 11. Information Ordering | 5 | 3.0-5.0 | 5 | 3.0-6.0 | 10 | 2.0-5.7 | 8 | 3.7-6.7 |
| 12. Math Reasoning | 2 | 1.0-5.3 | 1 | 6.0 | 9 | 2.3-6.3 | 8 | 1.7-6.7 |
| 13. Number Facility | 3 | 1.0-5.0 | 5 | 1.0-5.0 | 9 | 1.3-6.3 | 7 | 2.3-6.7 |
| 14. Fluency of Ideas | 1 | 2.3 | 1 | 3.3 | 10 | 1.3-6.3 | 6 | 3.3-6.3 |
| 15. Time Sharing | 4 | 3.0-6.7 | 6 | 1.0-6.3 | 10 | 3.3-6.7 | 10 | 2.3-6.7 |
| 16. Flexibility of Closure | 4 | 5.0-6.0 | 6 | 1.7-7.0 | 10 | 3.3-6.3 | 8 | 3.7-6.7 |
| 17. Speed of Closure | 1 | 5.3 | 0 | | 8 | 2.7-6.7 | 8 | 4.3-6.7 |
| 18. Selective Attention | 8 | 2.7-7.0 | 8 | 3.0-5.7 | 9 | 3.7-6.3 | 9 | 3.7-6.7 |
| 19. Perceptual Speed | 8 | 3.3-6.0 | 8 | 4.0-6.7 | 10 | 2.3-6.7 | 9 | 2.7-7.0 |
| 20. Spatial Orientation | 8 | 4.0-7.0 | 8 | 5.0-7.0 | 10 | 3.3-7.0 | 10 | 4.0-7.0 |
| 21. Visualization | 4 | 2.7-6.0 | 4 | 3.7-6.0 | 10 | 2.3-6.7 | 9 | 3.3-6.7 |
| 22. Static Strength | 2 | 2.7-4.7 | 1 | 5.3 | 8 | 3.3-6.7 | 7 | 3.7-6.7 |
| 23. Explosive Strength | 0 | | 0 | | 8 | 2.7-6.7 | 6 | 5.0-7.0 |
| 24. Dynamic Strength | 0 | | 0 | | 8 | 2.3-6.3 | 6 | 1.3-6.3 |
| 25. Trunk Strength | 0 | | 0 | | 7 | 2.7-6.3 | 6 | 1.7-6.7 |
| 26. Stamina | 4 | 3.0-5.0 | 4 | 4.0-5.0 | 9 | 3.3-7.0 | 10 | 1.7-6.3 |
| 27. Extent Flexibility | 5 | 2.7-6.0 | 5 | 2.3-5.0 | 9 | 4.0-6.3 | 10 | 4.0-6 7 |
| 28 Dynamic Flexibility | 0 | | 0 | | 9 | 3.3-6.7 | 9 | 2.3-6.3 |
| 29. Gross Body Equilibrium | 7 | 2.3-3.3 | 6 | 2.0-3.3 | 9 | 3.3-6.3 | 9 | 3.3-6.7 |
| 30. Speed of Limb Movement | 5 | 3.0-5.0 | 6 | 2.0-5.3 | 10 | 3.7-6.7 | 9 | 4.3-6.7 |
| 31. Gross Body Coordination | 1 | 4.0 | 1 | 4.0 | 9 | 3.3-6.7 | 9 | 2.3-6.7 |
| 32. Multi-Limb Coordination | 6 | 1.0-7.0 | 4 | 3.0-5.0 | 8 | 3.3-6.3 | 7 | 2.3-6.7 |
| 33. Wrist/Finger Speed | 1 | 3.0-5.3 | 1 | 4.0 | 10 | 2.7-6.7 | 10 | 2.7-6.7 |
| 34. Finger Dexterity | 3 | 3.0-5.0 | 1 | 4.0 | 10 | 3.0-6.3 | 9 | 2.3-6.7 |
| 35. Manual Dexterity | 3 | 3.0-5.7 | 0 | | 10 | 2.7-6.3 | 9 | 3.3-6 7 |
| 36. Arm/Hand Steadiness | 8 | 2.0-5.7 | 8 | 2.7-6.0 | 10 | 4.3-6.7 | 10 | 3.0-6.7 |
| 37. Control Precision | 1 | 5.3 | 3 | 5.0-6.7 | 9 | 2.3-6.3 | 9 | 2.7-6.3 |
| 39 Rate Control | 1 | 3.3 | 0 | | 9 | 2.3-6.7 | 10 | 3.0-6.3 |
| 139. Reaction Time | 8 | 3.0-7.0 | 8 | 4.0-7.0 | 8 | 4.7-7.0 | 7 | 4.3-6.7 |
| 140. Choice Reaction Time | 0 | | 0 | | 8 | 2 7-6.7 | 8 | 3.3-6 7 |

The mean ability level ratings for abilities rated as required for both systems by all raters in a sample are presented in Table 3. The means pro-

TABLE 3

Mean Ratings for Abilities
Unanimously Rated as Required

| Ability | Scientist Sample | | | Instructor Sample | | |
|---|---|---|---|---|---|---|
| | Redeye $\bar{X}$ | Stinger $\bar{X}$ | t p<.05 | Redeye $\bar{X}$ | Stinger $\bar{X}$ | t p<.05 |
| 5. Memorization | 6.3 | 6.3 | n.s | 5.5 | 5.5 | n.s |
| 6. Problem Sensitivity | 5.0 | 5.3 | n.s | 4.8 | 5.5 | n.s |
| 15. Time Sharing | (4.7)a/ | (3.8) | n.s | 5.3 | 4.5 | 2.26 |
| 18. Selective Attention | 4.6 | 4.7 | n.s | (5.0) | (5 1) | 2.16 |
| 19. Perceptual Speed | 4.4 | 5.2 | n.s | (5.1) | (4.8) | n.s |
| 20. Spatial Orientation | 5.8 | 5.9 | n.s | 5.5 | 5.6 | n.s |
| 133. Wrist/Finger Speed | ( ) | ( ) | | 5.1 | 5.0 | n.s |
| 136. Arm/Hand Steadiness | 4.0 | 4.3 | n.s | 5.6 | 5.4 | n s |
| 139 Reaction Time | 5.6 | 5.7 | n.s | (4.7) | (4 0) | n.s |

a/ Numbers in parenthesis are based on fewer than 8 ratings for the Scientist Sample or 10 ratings for the Instructor Sample.

vide some evidence that Stinger may place more demands on the abilities of soldiers than Redeye, as indicated by higher ability level ratings for some abilities. However, the mean differences are not overwhelming and generally do not approach statistical significance. Nevertheless, the data do suggest that Stinger is certainly not less demanding than Redeye.

Although the data do not fully support the notion that Stinger requires more or higher levels of abilities that Redeye, the mean ratings do help to identify "high drivers" such as memorization, problem sensitivity, and spa-

tial orientation. Relatively high levels of these three abilities are required by both systems. Had this type of information been available early in the system design process, it may have been possible to modify the design of the newer system to reduce the demand for these abilities.

The potential usefulness of ability requirements ratings are limited, however, by the lack of empirical data. The relationship between some estimated level of an ability requirement and actual job or task performance needs to be determined. Without this information it is not possible to establish what effect an individual with a "5" level, as opposed to a "6" or "4" level, of an ability would have on system performance if the requirement was estimated to be at level "6" for that ability. Information about the distribution of abilities in the population is also needed. If most individuals in a given population possess high levels of certain abilities then it may not be necessary to design systems that require low levels of those abilities. On the other hand, if few individuals have high levels of particular abilities, it may be wise to avoid system designs that require high levels of those abilities.

Whether or not an ability rating methodology can be developed into a sensitive psychometric instrument, the rating approach used in the research study presented in this paper appears to have potential applications in the design process. At the least, a procedure which can identify "high driver" abilities before a system has been fully designed and built may be helpful to design engineers and allow them to avoid or modify designs that are likely to be too demanding on the psychomotor and perceptual capabilities of personnel.

ACKNOWLEDGMENTS

This research was conducted while the author was a member of the Task Force for the Reverse Engineering Project. The opinions, views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, expressed or implied, of the U.S. Army Research Institute for the Behavioral and Social Sciences or the Department of Defense or the United States Government.

REFERENCES

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

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

Promisel, D.M., Hartel, C.R., Kaplan, J.D., Marcus, A., & Wittenburg, J.A. (1985). Reverse engineering: Human factors, manpower, personnel, and training in the weapon systems acquisition process (ARI Technical Report 659). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rossmeissl, P.G., Tillman, B.W., Rigg, K.E., & Best, P.R. (1983). Job assessment software system (JASS) for analysis of weapon systems personnel requirements (ARI Research Report 1355). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Shields, J.L. & Baker, J.D. (1981). The Army's personnel problems and the golden spike solution. Paper presented at the National Security Industrial Association First Annual Conference on Personnel and Training Factors in System Effectiveness, San Diego, CA.

Theologus, G.C., Romashko, T., & Fleishman, E.A. (1970). Development of a taxonomy of human performance: A feasibility study of ability dimensions for classifying human tasks (AIR Technical Report 726-5). Washington, DC: American Institutes for Research.

# Computerized Approaches for Estimating Ability Requirements

Paul G. Rossmeissl
U.S. Army Research Institute

The past few years have witnessed a dramatic increase in the power, sophistication, and availability of microcomputers. It seems natural that this new technology be devoted to addressing the recurring problem of determining what human abilities are needed to do what must be done in a given job. This paper discusses some of the ways in which the new technology could be brought to bear on the problem at hand.

One way to use the computer to investigate ability requirements is to take a tried and true method (one that works without a computer) and to computerize it. That is, to display the instrument on a CRT and store the responses on disc rather than using paper and pencil for these functions. The first part of this paper will review an investigation that followed this approach to the issue.

An alternative technique for using computers in this area is to use the machine to administer methods that are either difficult or impossible to administer in any other manner. The final portion of this paper will propose and discuss a few methods of this type.

## Paper-and-pencil vs. Computerized Rating Scales

One of the most commonly used techniques in the determination of ability requirements is the use of ability rating scales (i.e. Fleishman, 1982). Such scales typically include the definition of a human ability and a rating scale anchored with behavioral examples. While the scales are typically presented on paper, they can be easily presented on a computer's CRT as well.

Rossmeissl, Kostyla, and Tillman (1983)[1] conducted an investigation of computerized versus paper-and-pencil modes of rating scale presentation. They used rating scales of 30 human abilities developed by Rossmeissl and Dohme (1982). These rating scales are very similar to those developed by Fleishman (1975, 1982), but contain anchor stimuli that are directly relevant to jobs within Army aviation. The computerized presentation of the scales was implemented by Rossmeissl, Rigg, and Best (1983) using the binary decision flow diagrams developed by Mallamad, Levine, and Fleishman (1980). The two procedures were evaluated by having experienced Army aviators rate the abilities required to perform an Army aeroscout helicopter mission. Half of the aviators estimated ability requirements using paper-and-pencil rating scales, while the other half used rating scales presented on an Apple II microcomputer. All of the ratings were made along a seven point scale.

---

[1]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.

The findings of this research showed that the two methods gave very similar results. Both methods showed very high inter-rater agreement. The inter-class correlation coefficient (ICC) for the paper-and-pencil scales was .91, while the ICC for the computer presented scales was .96. Both methods were also able to produce ratings that discriminated among the various aptitude requirements (both $F$ values > 190 and significant at the .001 level).

This result is perhaps the most important aspect of the research, in that despite the somewhat popular belief of everything being better when its on a computer, the computerized procedure did not show superior psychometric properties in comparison to the traditional paper-and-pencil forms. On further thought, however, this finding really should not be surprising, since after all the same rating scales and much the same definitions were used in both cases.

The only real difference in results between the two procedures was that the computerized method led to ratings of not required for four abilities that were rated as required with the paper-and-pencil forms. These four abilities are given in Table 1. Subsequent analyses revealed that the discrepancy in ratings given to reaction time was an artifact of the

Table 1
Abilities Rated as Not Required with the Computer but as Required with Paper-and-Pencil Forms

| Ability | Mean rating with paper-and pencil forms |
|---|---|
| Stamina | 6.1 |
| Closure flexibility | 5.4 |
| Reaction time | 5.3 |
| Static Strength | 4.8 |

decision diagrams used to program the computer. The reason behind the other differences was not perfectly clear, but was most likely attributable to subtle differences between the two methods in the ability definitions. Such small differences could easily be rectified with future versions of the instruments.

Given that the two procedures produce similar results in a measurement or psychometric sense, a choice between the two instruments would depend upon other consideration Each procedure has advantages for particular situations. The paper-and-pencil forms are inexpensive and can be administered in groups or by mail. These considerations are especially compelling when the instrument needs to be administered to a large number of job experts to get the required data.

The computerized procedure, however, does have advantages when the subject matter experts are in short supply or have limited availability

for testing. One of these advantages is time. Rossmeissl et al. noted that the mission requirements could be evaluated in about fifteen to thirty minutes using the computer, while it would take from thirty to forty-five minutes to complete the task with the paper-and-pencil rating scales. Also, a debriefing of the participants in this research showed that those who rated ability requirements with the computer found the task more enjoyable than those who used the paper-and-pencil scales. Several participants in the computer group offered to perform the task again, perhaps for some other mission. Such a desire seemed to be far from the mind of anyone who had used the paper-and-pencil scales.

Other Computerized Procedures

The approach outlined above where one simply computerizes an already popular paradigm makes minimal use of the power inherent in the microcomputer. An alternative approach is to develop ability requirement assessment procedures that are designed especially for computer administration. The remainder of this paper will discuss some of the procedures of this sort currently be considered by the Army Research Institute (ARI).

One of these approaches would maintain the use of rating scales to estimate the ability requirements, but use Artificial Intelligence (AI) techniques to provide a more powerful and easy to use tool. Current hardware and software capabilities make it possible to develop a system that "learns" from the user as well as one that allows the present user to learn from the responses of previous users. Such a system would select questions and negotiate responses with the user using the responses of the previous users as the basis of negotiation. Rossmeissl, Rigg, and Best (1983) presented an outline of how such a system might work.

The system would be menu driven to elicit information about the boundary conditions of the job being assessed. The highest order menu might contain global job titles like mechanic, driver, instructor, pilot, etc. and ask the user to select the job most like the one being assessed. If the user selected "mechanic," the next menu would present more specific information about mechanics such as automotive mechanic or tracked vehicle mechanic, and ask the user to select that job closest to the job being assessed. The final menu in this sequence might ask the user to select that job closely resembling the one being analyzed. For example if the user had selected tracked vehicle mechanic, he or she would be presented with a menu including jobs or systems in this area, like M113 armored personnel carrier or M1 Main Battle Tank.

A selection from the final menu would initialize scales and data matrices with the boundary conditions for a job of mechanic in that particular system. The user would then proceed through the branching scales "teaching" the software how the new job differs from the one within the reference system. The user would still branch around inappropriate scales as in the earlier procedure, but if he or she skipped a scale which had been a requirement in the reference job, the software would switch into a negotiation mode, much like the Delphi procedure, and seek to determine what accounts for the difference in job ratings. Such a procedure would also occur when the user rated a scale especially high or low in comparison with that scales rating in the reference job or system.

The key aspect in programming a system of this type is the use of AI procedures to control the presentation the material. In this case the software would be constructed within the framework of a work breakdown structure in which the individual functions, procedures programs, and utility programs are all specified in a formative front end analysis. Within this analysis the individual blocks of the system each perform a single, non-redundant function and are controlled by an executive program. As the system is used the executive program adds this information to the data base containing the boundary or reference job, and in this continue to reduce uncertainty about the new job as well as jobs that had been previously entered. A system of this sort would make better use of the power of the computer, while maintaining the traditional rating scale approach to estimating ability requirements.

As noted previously, such a system is well within the capabilities of current hardware and software systems and a prototype could be developed quite easily. The problem in producing an operational system is that a great deal of basic psychological research needs to be accomplished before such a system could be meaningful. Such work would include: determining the proper number and type of boundary conditions at each menu level, scaling the reference jobs or systems, and developing the appropriate type of scale for the final assessment stage. This last step is particularly difficult in that this scale should relate to the reference system but still allow for comparisons across all possible jobs that might be analyzed.

The speed which computers can present and store data has also led ARI to consider ability requirement analysis procedures that are not dependent on rating scales. One such procedure makes use of Thurstone's law of comparative judgement. Thurstonian scaling uses pair-wise presentation of the data to produce an interval scale for the stimuli. Scale values obtained in this manner are directly comparable across both jobs and abilities. Besides being able to make use of the rich statistical theory underlying the law of comparative judgement, such a method has advantages over rating scales in that one need not be concerned with the development of appropriate anchor stimuli and that the procedure is easier to understand and perform than anchored rating scales

The disadvantage of the traditional Thurstonian approach is that it is that it can be time consuming to both administer and to analyze. For example, to obtain scale values for say twenty abilities would require (20!-2!)/2!*(20-2)! = 190 comparisons. If this procedure were executed on paper forms or cards, not only would the comparisons themselves take considerable time, but the raw data would require tedious hand analyses or keypunching into a computer.

Computerizing this process would clearly simplify the job of analysis, but it could also conceivably reduce the number of comparisons required to scale the abilities. This result could be accomplished by having the computer first present all of the abilities asking the user which are required for the job. All abilities estimated as not required could be assigned a scale value of zero. The computer could then arrange the abilities that were thought to be required in pairs for scaling. If

the twenty abilities mentioned above could be reduced to say fifteen the number of pair-wise comparisons required to complete the task could be reduced to (15!-2!)/2!*(15-2)! = 105 or 55% of the comparisons required with the full ability set. In such a case the computer is doing what it does best: inputing information in one form and quickly processing and converting it into another form.

## Conclusion

This paper has outlined a few ways that computer technology can be brought to bear on the problem of estimating human ability requirements within a particular job. An important feature of any such system is that the technology should a be means in arriving at a reliable, valid, and useful analysis tool and not a goal in itself. It is sometimes easy to view a technologically advanced instrument as impressive just because of its technology. The final judgement of any instrument should be how well it does the job at hand, not how it does it. In this case, it is important to remember that a good computerized ability analysis package must be firmly based on empirical research and data before it can be truly useful.

## References


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

Fleishman, E. A. (1982), Systems for describing human tasks. American Psychologist, 37, 821-934.

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

Rossmeissl, P. G., and Dohme, J. A. (1982), Using rating scales to determine the aptitude requirements of Army systems. Proceedings of the 24th Annual Conference of the Military Testing Association, San Antonio, TX.

Rossmeissl, P. G., Kostyla, S. J., and Tillman, B. W. (1983), Initial Test and Evaluation of a Computerized Ability Assessment Technique. Paper presented at the meetings of the American Psychological Association, Annaheim, CA.

Rossmeissl, P. G., Rigg, K. E. and Best, P. R. (1983), Job Assessment System (JASS) for Analysis of Weapon Systems Personnel Requirements. U. S. Army Research Institute for the Behavioral and Social Sciences Research Report # 1355, Alexandria, VA.

Thurstone, L. L. (1927), A law of comparative judgement. Psychological Review, 34, 273-286.

# Chaparral Crew Performance in the Realistic Air Defense Engagement System

Sarli, Gary G., Johnson, David M., and Lockhart, John M.
United States Army Research Institute
for the Behavioral and Social Sciences
Fort Bliss Field Unit, Fort Bliss, Texas

One of the dangers U.S. forces may face in a future conflict is the attack helicopter. These helicopters are capable of approaching their targets at an extremely low altitude, and need unmask (rise) only briefly — 20 seconds or less — in order to engage their targets (Personal communication, Directorate of Combat Developments, April 1985. The Chaparral is a member of the Army's family of Short Range Air Defense (SHORAD) weapon systems which defends against low-flying fixed-wing aircraft (jets) and rotary-wing aircraft (helicopters) at ranges of less than ten kilometers. Chaparral is vehicle-mounted and fires infrared-seeking missiles; the other SHORAD weapons are the man-portable Redeye and Stinger missile systems, and the Vulcan vehicle-mounted gun system (FM 44-1, 1976). All SHORAD weapons employ at least two soldiers: A chief or leader, and a gunner. The leader is responsible for receiving alerts, visually identifying aircraft, and ordering the gunner to fire. The gunner is responsible for activating the weapon, visually acquiring and tracking aircraft, and firing the weapon. Other crew members (in the case of Chaparral, the driver and assistant gunner) help emplace the weapon and then search for aircraft. Actions taken upon detection of an aircraft will vary depending upon many factors, including the aircraft's specific identity and its actions (FM 44-3, 1977; FM 44-6, 1980).

It is vital that the team or crew make the correct engagement decision in a timely manner to avoid destroying friendly aircraft or allowing hostile aircraft to destroy friendly targets. Inasmuch as CONUS Chaparral crewmembers have been shown to misidentify aircraft 18% of the time (TEA 12-81) the Improved Chaparral has been equipped with an Identify Friend or Foe (IFF) electronic interrogation system. When activated by the gunner, the IFF system sends a coded message to the aircraft and classifies the aircraft as a "true friend", "possible friend", or "unknown", depending upon the aircraft's reply. The IFF classification tones are heard by the senior gunner and the squad leader over their headsets. Of course, the IFF system is not perfect; aircraft can be misidentified due to operator error or system damage.

In order to afford friendly aircraft maximum protection, SHORAD crews and teams operate under weapon control statuses. There are three weapon control statuses: weapons hold, tight, and free. In weapons hold, the gunner only fires if attacked. In weapons tight, if an aircraft is identified by IFF or visually as a true friend, it is tracked until it leaves the area. If the IFF classifies the aircraft as a possible friend or as an unknown, the aircraft must be visually identified as hostile before it can be engaged. In weapons free status, true and possible friends are treated the same as in weapons tight; however, if the IFF classifies the aircraft as unknown, it may be engaged by Chaparral unless local directives and/or SOPs say otherwise. Of course, the gunner may always fire in defense of himself or his defended asset (FM 44-4, 1984). It can be seen that the IFF subsystem and the weapons control status complicate the engagement process by providing more data to be processed within the same amount of time and requiring differing responses depending upon the weapons control status.

The present experiment examined Chaparral crew engagements of scale model helicopters under various realistic conditions in an effort to answer questions such as: Is engagement performance different as a function of weapons control status? Does aircraft intent interact with the effects of weapons control status? How does IFF return interact with weapons control status?

The independent variables were 1) Aircraft intent (friendly or hostile), 2) weapon control status (tight or free), and 3) IFF return (possible friend or unknown for hostiles; true friend or unknown for friends).

The dependent variables were total engagement time; times at detection, target handoff (the process by which the squad leader helps the gunner to find the target), visual acquisition, identification, weapon lock-on, and launch; numbers of correct and incorrect target handoffs, identifications, and engagement decisions; and whether or not the missile hit the aircraft. Summary performance measures included the percentage of correct identifications, the percentages of hostiles and friends which were engaged, the percentages of all hostiles and all friends which were killed, the percentages of engaged hostiles and friends which were killed, and the percentage of hostile targets releasing ordnance.

## Method

### Subjects

A total of eight Chaparral crews stationed at Fort Bliss, Texas, served as subjects. The Chaparral may be manned by a crew of four or five soldiers; four-man crews were used in this experiment. Each crew contained the following personnel, all of whom have a Military Occupational Specialty (MOS) Code of 16P: Squad leader (rank E5 or E6); Senior gunner (E4 or E5); Assistant gunner (E2 - E4); and Driver (E2 - E4).

### RADES Equipment

The Realistic Air Defense Engagement System (RADES) uses realistic scale model aircraft which perform realistic maneuvers in an outdoor environment against actual air defense weapons. Live rounds are not fired; rather, a computer determines what the rounds' flight path would have been, and whether or not the aircraft would have been hit. The RADES system is located on a 2 x 2 kilometer site at Condron Field, White Sands Missile Range, New Mexico. This area, located in the Chihuahuan Desert, contains vegetated sand dunes up to about two meters high in the test area, and mountains approximately 10 kilometers to the west and 60 kilometers to the south. Visibility is usually in excess of 60 kilometers.

The portion of the RADES system used in the present experiment consisted of scale model aircraft, a weapon-computer interface, and computers which were used for scenario presentation and data storage. The RADES target aircraft were non-flying 1/5 scale model Soviet Hind-D and U.S. Cobra helicopters. The helicopters were mounted upon three stands which could be raised and lowered five meters under computer or manual control. These stands were hidden behind sand dunes so that the helicopters could not be seen by the weapon crews when the stands were lowered. In this experiment, the stands were positioned 300 to 450 actual meters (1500 to 2250 scale meters) from the Chaparral crew and always faced at least 20 degrees away from the crew. All the helicopters were equipped with a representative infrared

radiation (IR) source so the Chaparral IR seeker would have a realistic probability of locking onto them. To provide feedback to the crews, each helicopter was also equipped with a device which emitted smoke for several seconds if the helicopter was "destroyed".

A computer at RADES HQ obtained information such as azimuth, elevation, and firing from an electronic interface connected to the Chaparral; information such as time of detection was obtained from the crews' communications network. The computer then determined times for each critical engagement event and whether or not the engagement was successful. Another computer at RADES HQ stored each engagement scenario and the data produced during the running of that scenario.

Military Equipment

Military equipment which was used included one Chaparral Missile System equipped with the IR training round MIM-72F guidance seeker (which simulates the seeker on the MIM-72C improved missile), mounted upon an inert M3Ø training missile. One IFF subsystem training set, three 7x5Ø binoculars, two TA312 field telephones with wire, and one 24N tool kit were also used.

During engagements the senior gunner sat in the gunner's compartment of the Chaparral, the driver occupied an observation post 2Ø meters south of the launcher unit, and the squad leader and the assistant gunner occupied a command post 2Ø meters north of the weapon. This distance was 1/5 of the normal separation, to reduce parallax errors which would otherwise have been introduced by the scaled distances to the targets. The weapon system and observation and command posts were fixed during the experiment, with the Chaparral facing south (FM 44-3, 1977; FM 44-4, 1984).

Procedure

Each of the eight crews received two practice trials and eight test trials; one test trial for each combination of intent, weapons control status, and IFF return. Before the engagement trials began, the crews were given a simulated Operation Order specifying the mission, enemy, Primary Target Line (expected direction of attack, which was always due south), and tactical situation. They were told to search a 9Ø degree wide sector centered on the Primary Target Line. The crews were also told that both friendly and hostile aircraft would be attacking ground targets in their area. Specific information concerning weapons control status was passed to the crews on a trial-by-trial basis.

Prior to each trial, the experimenter selected the appropriate IFF response using the hidden Mode Select switch attached to the Chaparral IFF Subsystem Training Set. Each trial began when the crew was placed on alert and began searching for aircraft. Early warning and cueing were always given via telephone 2Ø to 3Ø seconds prior to the appearance of the aircraft. The cueing was given in the form of a clock azimuth, with the Primary Target Line designated as 12 o'clock (3Ø degrees to the left was 11 o'clock, 3Ø degrees to the right, 1 o'clock, etc.). The clock azimuth cue varied from 11 to 1 o'clock; all aircraft first appeared within 15 degrees of this azimuth. On each trial, one helicopter rose into view, hovered for 2Ø seconds, and then descended out of sight. The first crew member to detect the aircraft shouted "Target!" and the target's clock azimuth, and helped the squad leader to find it. The squad leader helped the senior gunner to find the aircraft, then identified the aircraft visually, aided by his binoculars. As soon as the senior gunner detected the aircraft, he centered it in the weapon sight range

ring, electronically challenged it, and listened for the IR tone which indicated that the missile's infrared seeker had acquired it. If the squad leader gave an engagement command and launching was feasible, the senior gunner "launched" the missile. After launching, the squad leader monitored the helicopter and issued another engagement order if necessary (Chaparral engagement, Parts 1 and 2). To prevent the crews from learning the identity of the helicopter on each stand, the direction each helicopter was facing was changed periodically and the type of helicopter mounted upon each stand was changed midway through the trials for each crew.

## Results and Discussion

Table 1 lists the percentage of trials in which various actions occurred and their associated event times. Although the Chaparral crews detected 97% of the aircraft, the crews only destroyed 37% of the hostiles and were only able to destroy 7% of the hostiles prior to ordnance delivery. IFF return had a significant effect upon the percentage of hostiles engaged under weapons free: 100% of the hostiles returning an unknown response but only 50% of those returning a possible friend response were engaged, even though 90% of the hostiles were correctly identified visually (Chi square (1) = 5.33, $p < .05$). The fratricide rate (percentage of friends destroyed) was 16%. The RADES Chaparral crews' visual aircraft recognition performance in the field was similar to the performance of the Chaparral crews studied by TRASANA (TEA 12-81) in the classroom. Aircraft intent significantly affected the correctness of the identification decision; crews identified 90% of hostiles but only 79% of friends correctly (see Table 2).

While the overall engagement time was 17.6 seconds, gunner interrogation and firing occurred rapidly. The longest engagement events were the non-machine-aided detection and visual identification and the IR seeker lock-on events. Lock-on overlaps with visual identification. The only significant effect on any engagement event time was the three-way interaction of intent, weapons control status, and IFF return on identification time (see Table 2). For hostile helicopters, the fastest identification time (14.7 seconds) occurred in weapons free with an IFF return of unknown, and the slowest (19.5) in weapons free, possible friend. For friends, the fastest identification time (15.2 seconds) occurred in weapons free, true friend, and the slowest (21.4) in weapons tight, true friend.

TABLE 1
ENGAGEMENT PERFORMANCE

| PERCENTAGE SCORES | EVENT TIMES (in seconds) |
|---|---|
| Aircraft detected.....97 | Detection.....7.5 |
| Correct handoffs.....87 | IFF interrogation.....8.2 |
| Correct IDs.....88 | Track / lock-on.....16.8 |
| Hostiles.....90 | Visual ID.....17.4 |
| Friends.....79 | Missile launch.....17.6 |
| Friends engaged .....26 | |
| Hostiles engaged.....70 | |
| Friends killed.....16 | |
| Hostiles killed.....37 | |
| Hostiles releasing ordnance..93 | |

Table 2

MANOVA Results

| DEPENDENT VARIABLE: | INDEPENDENT VARIABLE | df | F | p | DIRECTION OF EFFECT |
|---|---|---|---|---|---|
| Detection time | Weapons Control Status (WCS) | 1,56 | 3.40 | .071 | Free > Tight |
| ID Correctness | Intent | 1,56 | 5.62 | .021 | Hostile > Friend |
| Decision to fire | Intent | 1,56 | 3.50 | .067 | Hostile > Friend |
| Decision to fire | WCS x IFF | 1,56 | 3.50 | .067 | Free-UNK > Free-TFPF<br>Tight-TFPF > Free-TFPF |
| Time of Track | Intent x IFF | 1,34 | 3.76 | .061 | Hostile-UNK > Hostile-TFPF<br>Hostile-UNK > Friend-UNK<br>Friend-TFPF > Friend-UNK<br>Friend-TFPF > Hostile-TFPF |
| Time of ID | Intent x WCS x IFF | 1,56 | 4.22 | .045 | Friend > Hostile<br>Free > Tight, TFPF > UNK |
| Decision to fire | Intent x WCS x IFF | 1,56 | 3.50 | .067 | Hostile > Friend<br>Tight > Free, UNK > TFPF> |

The finding that crews were more likely to correctly identify hostiles than friends means that crews are more likely to think a friendly aircraft is hostile and engage it than they are to think an enemy aircraft is a friend and fail to engage it. This may be due to the fact that a 1/5 scale Hind-D is much larger than a 1/5 scale Cobra, and therefore subtends a larger visual angle, making recognition easier. A related idea is that Hind-D's are intrinsically easier to recognize than Cobras because Hind-D's have more distinctive features. Another possible explanation has to do with response bias. The crews are predisposed to identify aircraft as hostile because when they train using the Chaparral, visual identification is not always practiced. Instead, the aircraft presented to them are assumed to be hostile. A third possible explanation for these results is that crews spend more time and/or effort learning to identify hostiles than friends.

It is possible that Weapon Control Status did not have a significant main effect because the crews usually train under Weapons Tight.

Following the Chaparral experiment, more types of helicopters, more stands, and several types of flying jet aircraft have been added to RADES. We are optimistic about the use of RADES to study a wide variety of SHORAD air defense issues in the future.

## References


Baldwin, R. D. (1973). Capabilities of ground observers to locate, recognize, and estimate distances of low-flying aircraft. (HumRRO Technical Report No. 73-8). Alexandria, Virginia: The George Washington University.

Chaparral engagement of aerial targets, part 1 (CL7838F) (undated). Fort Bliss, Texas: U.S. Army Air Defense School.

Chaparral engagement of aerial targets, part 2 (CL7839F) (undated). Fort Bliss, Texas: U.S. Army Air Defense School.

Headquarters, Department of the Army (1976, March). FM 44-1, U.S. Army air defense artillery employment. Washington, D.C.

Headquarters, Department of the Army (1977, September). FM 44-3, Air defense artillery employment, Chaparral / Vulcan. Washington, D.C.

Headquarters, Department of the Army (1984, November 2). FM 44-4, Operations and training, Chaparral. Washington, D.C.

Headquarters, Department of the Army (1980, March). FM 44-6, Operations and training for Forward Area Alerting Radar (FAAR) and Target Alert Data Display Set (TADDS). Washington, D.C.

U.S. Army TRADOC Systems Analysis Activity (1981). TEA 12-81, Chaparral/ Redeye Training Subsystem Effectiveness Analysis, Volume 2, Chaparral Main Report. White Sands Missile Range, New Mexico.

# Command and Control Teams: Techniques for Assessing Team Performance

Merri-Ann Cooper
Advanced Research Resources Organization

Samuel Shiflett
George Washington University

Arthur L. Korotkin
Advanced Research Resources Organization

The Air Force uses tactical command and control ($C^2$) systems to task and manage combat missions and responses to crisis situations. A tactical $C^2$ system provides: (1) central authority and coordination to determine how to use available forces, (2) a structure to send information and decisions between the commander and the forces and between adjacent units, and (3) a mechanism to originate and filter information from a number of sources to the commander. The key to the $C^2$ system is the effective and rapid collection, processing and transfer of information to the commander. Considerable resources, management attention and research have been spent in the topic of $C^2$ systems in recent years. In order to address several still unresolved issues on tactical $C^2$ systems, the Air Force Human Resources Laboratory began a research program focusing on personnel performance and training. This study (with monitor Lawrence Finegold) addressed one of the unresolved issues: team performance. The purposes of the project were to: identify the team characteristics of units operating $C^2$ systems, to study the ways currently used to evaluate $C^2$ operations to see if team aspects are considered, to study $C^2$ teams to determine if team functions could be observed, and to see if rating scales on team functions could be used for the assessment of $C^2$ team performance.

## Method

In order to determine whether a team performance framework was appropriate for studying tactical $C^2$ systems, information was gathered about the basic characteristics of $C^2$ systems, about the team aspects of $C^2$ systems and about the procedures used to evaluate personnel performance in $C^2$ systems. Three methods were used to study tactical $C^2$ systems. First, the project staff reviewed Air Force documentation and research on $C^2$ units. Second, the project staff made field visits to observe three $C^2$ units: The North American Air Defense Command (NORAD)/20th Air Division at Ft. Lee, Virginia; the 728 Tactical Control and Reporting Center (CRC) at Eglin Air Force Base, Florida; and the 507th Wing Tactical Air Control Center (TACC) at Shaw Air Force Base in South Carolina. At these sites, semi-structured interviews were held with operational and training personnel. Third, the project staff observed exercises at two of the units and interviewed exercise participants and the associated exercise staff. The two exercises we observed were war simulations, Blue Flag, in which the TACC participated and a System Training Exercise for the CRC. During the exercises, the ARRO staff both identified team tasks and characteristics as well as observed and evaluated team performance using scales developed to assess team functions.

Because of the size and complexity of the $C^2$ units and the number of teams which make up these units, the observations and interviews focused on two small teams in each $C^2$ unit. Each team was selected using the following criteria: the team has a fixed structure, organization and task; the team participates in exercises; team interactions could be observed, and the teams differ in terms of being proceduralized (restricted in terms of activity options). The Weapons in the CRC represents the proceduralized type of team and the Fighter Duty Officer Team in the TACC was selected to represent a less proceduralized team.

## Assessment of Team Functioning

Three approaches have been used to evaluate team performance. One is the evaluation of the quality of team's product (e.g., Bass, Farrow & Valenzi, 1979; Morgan, Coates, Kirby & Alluisi, 1984). A second is a measure of the performance of the entire system of which a team is a part (e.g., Obermayer, Vreuls, Muckler & Conway, 1974). Neither of these methods is appropriate in this study since they do not differentiate among the contribution to outcomes of individual team members, of teams, and of available resources. We decided, instead, to use a measure of team functions. The measure was based on a taxonomy initially developed by Nieva, Fleishman and Rieck (1978) from a review of the substantial literature on team performance. Draft scales measuring team functions (see Figure 1 for an example) were developed and pretested in a second study by Shiflett, Elsner, Price and Schemmer (1982) in which films of Army combat and combat support teams were observed. Naive raters who had been briefly trained about the functions, observed taped segments and used the rating scales to determine whether or not the functions were present. These raters observations were reasonably reliable.

The team taxonomy is presented in Table 1. The functions in the taxonomy are organized into five general categories: (1) Orientation involves, the generation and exchange of information concerning member abilities, the task, and the environment (resources available, opposition data, and environmental conditions); (2) Resources Distribution concerns matching task requirements to member abilities and the number of members required; (3) Timing involves pacing of activities by the team and by individual team members; (4) Response Coordination concerns the coordination of activities between team members to ensure synchronized performance; and (5) Motivation which involves energizing and directing the group (e.g., developing group norms, reinforcing performance, and resolving conflicts).

## Team Characteristics of $C^2$ System

One of the two $C^2$ teams studied was the Weapons Team in the CRC. The Weapons Team has primarily responsibility for distributing aircraft to intercept unidentified aircraft and attacking hostile aircraft. The Weapons Team is headed by the weapons assignment officer (WAO) who is assisted by a weapons assignment technician. Under the WAO are subteams made up of dyads, each composed of an air weapons controller and an air weapons controller technician. Each of these dyads sits at a radar scope console and has primary responsibility for controlling aircraft in a particular area of the battle zone. The pilot is dependent primarily on that controller to get the aircraft to the target safely and efficiently. In the CRC team, most of the team functions were expressed in terms of communication activities, the

Rate the extent to which you perceived the RESPONSE COORDINATION function occurring indicating the degree to which the team coordination efforts occurred in a complex and detailed manner, requiring careful and continuous monitoring of other team member activities.

7 ____ The response coordination involved very complex and detailed adjustment and sequencing of behavior.

6 ____

5 ____

4 ____ The response coordination involved moderately complex adjustments and sequences of behavior.

3 ____

2 ____

1 ____ The response coordination involved simple adjustment and sequencing of behavior.

Figure 1: Scale measuring response coordination

Table 1

Taxonomy of Team Functions

A. Orientation Functions

   1. Information exchange regarding member resources and constraints

   2. Information exchange regarding team task and goals/mission

   3. Information exchange regarding environomental characteristics and constraints

   4. Priority assignment among tasks

B. Resources Distribution Functions

   1. Matching member resources to task requirements

   2. Load balancing

C. Timing Functions (Activity Pacing)

   1. General activity pacing

   2. Individually oriented activity pacing

D. Response Coordination Functions

   1. Response sequencing

   2. Time and position coordination of responses

E. Motivational Functions

   1. Development of team performance norms

   2. Generating acceptance of team performance norms

   3. Establishing team-level performance-rewards linkages

   4. Reinforcement of task orientation

   5. Balancing team orientation with individual competition

   6. Resolution of performance-relevant conflicts

From Shiflett, Eisner, Price, and Schemmer (1982)

major behavior in the CRC. The Orientation Function, defined in terms of information about intercepting a target, was observed in terms of display plotter board postings of the air situation, that are monitored by team members. A lot of time was spent on Response Distribution, as information about available resources was updated so that resources could be assigned to new targets. The Timing function was most directly seen in the attempts to ensure that the posting of the actual situation was current. Response Coordination functions occurred primarily between teams, as information flowed from the Surveillance Team, which identifies aircraft as friendly or hostile, to the Weapons Teams, to the people who update the plotter boards. Motivation functions were not in evidence, which is not surprising since they typically occur early in the groups development (Shiflett et al., 1982). One function we observed, that is not in the taxonomy, concerned the detection of errors or system monitoring.

The second team selected for study was the Fighter Duty Officer Team in the TACC. The TACC is the tactical air control element with primary responsibility for command and control of theater operations. This team has responsibility for carrying out, monitoring and adjusting tactical air operations as planned the previous day. The team receives and reviews the plan concerning the types of missions, the number of each type of mission, and the resources to be allocated. The team members plan in detail the sorties assigned to them. When the plan is changed, the team is responsible for planning new missions. It's job is to get the appropriate number of the proper type of aircraft, armed with the correct ordnance for the job, to the target and back to base or to an alternate landing site. In addition they are responsible, after the sorties are planned, to post them on the status boards in order of expected time over target. In the TACC, there was some interaction within the team, but most of the interaction observed were between team members and other teams, either in the TACC or in the cooperating units (e.g., the CRC, Army units). The senior members of the Fighters Duty Officer Team were involved in Orientation functions, by updating the team on changes in targets, intelligence and the battle situation. Resource Distribution was observed in the attempts of team members to plan sorties so that the allocated resources achieve the assigned missions. The Timing function was obviously a key aspect of the TACC, since the battle situation is time sensitive. Making sure that aircraft were allocated in a timely manner involved coordination in the team and with other teams. Response Coordination is the most critical of the team dimensions in the TACC since the TACC coordinates the activities of the offensive war. However, most of these functions are with individuals outside the primary team: with ground troops, tankers, reconnaissance, and the CRC. The motivational functions were not directly observable, although the level of motivation could be inferred from the heightened activity at certain periods. Finally, system monitoring or error detection was again observed in the attempts to obtain feedback about whether the missions proceeded as they should.

## Evaluation of $C^2$ Performance

Several types of assessment procedures are currently used to evaluate Air Force personnel who operate $C^2$ systems: assessment of the performance of units during training exercises, assessments of operational readiness, tests of individual performance and knowledge, and reviews of paperwork. In the two exercises that were observed, controllers observed the exercise, identified problems in performance and provided feedback to participants.

These evaluations concerned outcomes (e.g., number of enemy aircraft destroyed) rather than behavior. Individual performance and team interactions were rarely explicitly evaluated. This may be partially the result of the few evaluators, relative to participants, present in the exercise. Since these are primarily training exercises, the focus is also on improving performance, rather than evaluation. The other evaluations (Operational Readiness Assessment, Standardization/Evaluation Report and Management Effectiveness Inspection), despite being more standardized and more intentionally evaluative, do not assess team performance or coordination.

## Use of the Team Functions to Evaluate $C^2$ Functioning

A major purpose for studying tactical $C^2$ teams was to determine the adequacy of the taxonomy of team functions as a basis for evaluating $C^2$ performance. In general, it was found that the team functions were able to capture most team related aspects of tactical $C^2$ teams. All of the functions, except for motivation, defined in the taxonomy appear to be present, to some degree, in tactical $C^2$ teams. It became apparent that one function that occurred was not in the taxonomy. It was system monitoring, which involves the checking of all system elements for errors and ommissions.

A major problem emerging from the observation of $C^2$ teams involved the scales used. The scales describe the level of a function without regard for the situational requirements. Such scales may be less appropriate for evaluation than for description. In addition, in the $C^2$ setting, certain teams functions did not seem to differ in degree and thus were not appropriate to be rated on a seven-point scale. Rather, measures involving presence/ absence, frequency or timeliness might be more appropriate. There were two additional problems in assessing performance in tactical $C^2$ teams. First, the complexity of the $C^2$ situation made understanding and evaluation difficult for an observer who is not knowledgeable about the $C^2$ environment. Second, because many team functions involved communication, an observer not linked into the communication system could not effectively observe many team interactions.

## References


Bass, B. M., Farrow, D., & Valenzi, E. (1979) Analyses of PROFILE data. Miami, Fl: Florida International University.

Morgan, B. B., Coates, G. D., Kirby, R. H., & Alluisi, E. A., (1984) Individual and group performance as a functions of the team training load. Human Factors, 26, 127-142.

Nieva, V. F., Fleishman, E. A., & Rieck, A. (1979) Team dimensions: Their identification, their measurement, and their relationships. Final Report, Bethesda, MD: Advanced Research Resources Organization.

Obermayer, R. W., Vreuls, D., Muckler, F. A., & Conway, E. J. (1979) Combat-ready crew performance measurement system: Final Report. Northbridge, CA: Manned Systems Sciences, Inc.

Shiflett, S. C., Eisner, E. J., Price, S. J., & Schemmer, F. M. (1982). The definition and measurement of team functions. Final Report. Bethesda, MD: Advanced Research Resources Organization.

# Item Factor Analysis of ASVAB 14

Clarence McCormick
HQ U.S. Military Entrance Processing Command

In the Fall of 1984, for a special project, we tested all the students (about 15,000) in about 50 high schools scattered around the country with the Armed Services Vocational Aptitude Battery, Form 14. This study reports the results of some item factor analyses of Form 14A. The results for Forms B and C replicated those reported here. The students were 5,260 9th-12th graders with approximately half male and half female. The analysis involved only the 8 power scales. All analyses used the SPSSX principal axes procedure with squared multiple correlation as initial estimates of the communalities, the greater than unity equivalence criterion and varimax rotations.

All Items. Phi coefficients were used rather than tetrachorics because the procedure available to us for calculating tetrachorics could not handle the 200 items. Forty-three factors were initially extracted. After rotation only five factors showed enough variables loading on them highly enough (loadings equal to or greater than .30) to define the factors. The rest were all either singlets or doublets relative to this criterion. Table 1 presents a summary of the results for these 5 factors. The table presents the number of items in each subtest which showed loadings equal to or greater than 30 on each of the 5 factors; the average and range of the difficulty (p) values for those items; the final solution eigenvalues; and percents of variance accounted for by each factor.

Table 1

Summary of Factor Analytic Results for all 200 Items

| | FACTORS | | | | |
|---|---|---|---|---|---|
| Subtest | I | II | III | IV | V |
| General Sciences | 6 | | | 5 | |
| Arithmetic Reasoning | 4 | 19 | | | |
| Word Knowledge | 15 | | | 14 | |
| Paragraph Comprehension | 7 | | | | |
| Auto and Shop Information | | | 17 | | |
| Mathematics Knowledge | 2 | 8 | | | 3 |
| Mechanical Comprehension | 2 | | 2 | | |
| Eletronics Information | | | 6 | | |
| Average "p" | .75 | .46 | .47 | .44 | .23 |
| Range of p-values | .63-.92 | .30-.62 | .24-.74 | .22-.39 | .20-.26 |
| Eigenvalues | 26.13 | 4.01 | 3.36 | 1.99 | 1.17 |
| Percent of Variance | 13.1 | 2.0 | 1.7 | 1.0 | 0.6 |

The first three factors are clearly the Verbal, Math, and Shop/Technical factors reported for the ASVAB when the subtest scores are factor analyzed (Ree, et al., 1982). Factor IV is another, clearly defined, Verbal factor; and Factor V is another, minimally defined, Math factor.

The two verbal factors are clearly separated on the basis of the difficulty levels of the items. The first factor consists entirely of relatively easy

items while the third factor consists of items at moderate difficulty levels. There is no overlap in the ranges of these values for the two sets. Similarly, for the two

Match factors; however, the distinction now is between moderate and difficult items. Presumably, Factor V might have been better defined if there were more relatively difficult Math items available in the pool.

Factor III, however, presents us with simply a content factor with difficulty values over the full range from easy to hard. In the case of this set of items the common variance attributable to content seems to outweigh common variance on the basis of difficulty level. Perhaps if there were more easy and hard items available in the pool for this content area we might find additional difficulty factors.

With this idea in mind we can look at the 4 Arithmetic Reasoning, the 2 Math Knowledge, and the 2 Mechanical Reasoning items which appear on Factor I. We might hypothesize, as is often done, that these items are primarily verbal in nature and thus show very close relationsips to other verbal items in the pool. Perhaps. However, it is clearly true that these items are very easy items (the p-values range from .64 to .92) like the verbal items on the factor. It seems quite reasonable to hypothesize that these items are allocated to the first factor primarily on the basis of common difficulty level variance rather than common content variance. Again, we might infer that if more easy items for these content areas had been available we might have found separate difficulty factors for them.

One inference, then, which can be made from these results is that item factor analysis can supply information useful for the construction of batteries such as the ASVAB. Given an initial pool of items with relatively large subsets balanced for difficulty level as well as for content we can construct well structured tests which should help eliminate some of the discrepancies found in the factor analytic literature where some items seem to migrate from factor to factor across different studies.

Another important implication of these results is that counselors need to bear in mind the hierarchial nature of the test scores used in the battery. At the item level we find (ignoring differences in difficulty level) three separate, orthogonal sources of variance in the content of the items. When these items are clustered together into subtests and the resultant subtest scores are factored we find the same three factors; however, now these factors are correlated. These intersubtest correlations imply that the subtest scores tend to vary similarly--either up or down--on the average. That is to say that at the level of the subtest scores we find evidence of some general reasoning factor which is not evident among the item responses. It might be useful, as Jensen (1985) points out, to examine the relationships between different levels of general ability and various training criteria. Nevertheless, the counselor should also be aware of the fact that underlying the general factor are several relatively independent sources of variation which can contribute to various differences in individual and subgroup profiles. As an example of this principal we can examine an analysis of the ASVAB composites. The example also serves to underscore the sensitivity of

factor analysis to differences in the samples of items (or scales) which are intercorrelated for the analysis.

Composite Forms. To conserve space we present here only an analysis of the Verbal Composite which is made up of the Word Knowledge, Paragraph Comprehension and General Science subtests (75 items). However, it should be noted that the same results were obtained for the other six composites used in the Student Testing Program.

Six factors were found to be interpretable after rotation. Factor I consisted of 14 WK and 4 GS easy items; and Factor II of 13 WK, 1 PC, and 1 GS moderately difficult items. Factor III consisted of 6 GS and 1 WK moderately difficult items; and Factor V of 4 GS relatively easy items. Factor IV consisted of 8 PC easy items; and Factor VI of 5 PC moderately difficult items.

Thus, when all items were factored together we obtained 3 orthogonal factors reflecting differences in common variance between the three general content areas covered by the ASVAB; and the Word Knowledge, Paragraph Comprehension, and General Science items all loaded, for the most part, on the first factor. However, when we analyse only this set of items we find that the factor breaks up into 3 orthogonal sets representing the 3 separate subtests relative to content plus a matched set representing differences in difficulty levels.

Again we might note that if the subtests had been balanced in terms of equal numbers of hard, moderate, and easy items and in terms of relatively equally high loading items within each content area the results would probably be even clearer. However, we would like to stress here the similarity between these results and those of the multiple regression studies from which these composites were constructed. Certainly, the composite scores are highly intercorrelated. Never the less, within each composite the item sums would seem to be relatively independent and can be expected to vary relatively independently for at least some individuals and for appropriately constituted subgroups. Of course, we might expect to maximize that relative independence by using the results of item factor analyses to help select an appropriately balanced set of items for each subtest.

Subtest Items. The items of each subtest were also factored to obtain evidence on the dimensionality of each of the subtests. Current opinion seems to favor the use of tetrachorics when factoring dichotomous items. It is hard to see why this should be so. The reason often given is that tetrachorics are unaffected by item difficulty levels and so will not produce so called "difficulty" factors. However, it has long been known that difficulty factors can be obtained as readily when tetrachorics are used as when phis are used (Guilford, 1941; Gourlay, 1951). It is also well known that tetrachorics are inappropriate whenever guessing is likely to be involved in the responses (Carroll, 1945; Lord, 1980) as would seem to be the case for many ASVAB items.

At any rate, we calculated item intercorrelations both as phis and as tetrachorics and compared the resulting factor solutions. Both types of correlation coefficients produced the expected "difficulty" factors. In fact, the solutions were so similar that we simply present the summary results for the analysis which used the phi coefficents (Table 2).

TABLE 2
Summary of the Factor Analytic Results for the Subtest Item

| Scale | I No. of Items | I Average p | II No. of Items | II Average p | III No. of Items | III Average p |
|---|---|---|---|---|---|---|
| WK | 14 | .80 | 21 | .47 | | |
| PC | 8 | .67 | 7 | .43 | | |
| GS | 11 | .66 | 4 | .55 | 10 | .34 |
| AR | 5 | .73 | 7 | .59 | 16 | .39 |
| AS | 6 | .60 | 12 | .44 | 6 | .27 |
| MK | 10 | .60 | 3 | .43 | 12 | .35 |
| MC* | 10 | .55 | 4 | .41 | 8 | .36 |
| EI | 7 | .56 | 5 | .42 | 6 | .26 |

*MC also exhibited a doublet with difficulty values in the .20's.

It seems clear from the table all the factors for each subtest seemed to be difficulty factors. WK and PC exhibited two factors each: one for easy, the other for moderately difficult items. The other subtests showed three factors each; one for the easier, one for the moderately difficult, and one for the more difficult items.

Examination of the content of the items on each of the factors failed to reveal any consistent trends for the factors to be related to content differences.

Thus, the major source of variance among the item responses for each of the subtests would seem to be difficulty level. Happily so; for it is just these individual differences which we wish to measure. Each of the subtests, then, can be assumed to be unidimensional varying only along a scale of difficulty. This, seems to be true even where we might not expect unidimensionality. For example GS is made up, basically, of two different items types: those which have content taken from the physical sciences and those with content from the biological sciences. Yet, these items (as we saw in the initial analysis) do not seem to be measuring achivement in these separate areas of study. Rather, they seem to be measuring a more general aptitude: an aptitude for verbal learning.

The same inference as to the basic undimensionality of the subtests can be found by looking at another index of undimensionality the ratio of the first eiginvalues to the second as compared to the ratio of the second to the third. The former ratio ranged from 2.5 to 4 (with most about 3.5) while the latter ranged from 1.2 to 2.2 (with most about 1.5).

Note: Specific statistics obtained in these factor analyses are available from the author upon request.

References

Carroll, J.B. The effect of difficulty and chance success on correlations

between items or between tests. Psychometrika, 1934, 10, 1-19.

Gourley, W. Difficulty factors arising from the use of tetrachorics correlations in factor analysis. British Journal of Psychology: Statistical Section, 1951, 4, 65-73.

Guilford, J.P. The difficulty of a test and its factor composition Psychometrika, 1941, 6, 67-77.

Jensen, A. Test Review: Armed Services Vocational Aptitude Battery. Measurement and Evaluation in Counseling and Development, 1985, 18, 32-37.

Lord, F. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum, 1980

Ree, M.J., Mullins, C.J., Mathews, J.J. & Massey, R.H. Armed Services Vocational Aptitude Battery: Item and Factor Analyses of Forms 8, 9, and 10. AFHRL-TR-81-55. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, March 1982.

# ALTERNATE FORMS RELIABILITY OF THE ARMED SERVICES VOCATIONAL APTITUTDE BATTERY (ASVAB) FORMS 8, 9, AND 10

**Brian M. Stern**
**Human Resources Research Organization**

**Leonard A. White, Hilda Wing, and Sidney A. Sachs**
**U.S. Army Research Institute for the Behavioral and Social Sciences**

The purpose of this research is to document the reliability of examinee scores on alternate forms of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10. The ASVAB Forms 8, 9, and 10 are composed of ten subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto Shop Information (AS), Math Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). Two of these subtests, Numerical Operations and Coding Speed, are considered speeded tests. The remaining eight subtests are power tests with specific time limits for each administration. ASVAB subtests are grouped together in specific ways to form Aptitude Area composites and a measure known as the Armed Forces Qualification Test (AFQT). Scores on the AFQT are used for selection and scores on the ten Aptitude Area composites are used for occupational classification.

The ASVAB 8, 9, and 10 has six parallel forms designated as 8a, 8b, 9a, 9b, 10a, and 10b. For each of the six forms, the four subtests that comprise the AFQT consist of a unique set of items. The remaining six subtests have three parallel forms (8a, 8b), (9a, 9b), and (10a, 10b). For these six subtests, identically numbered ASVAB forms (e.g., 8a and 8b) contain the same items. Since the six forms have been shown to be equivalent (Ree, Mathews, Mullins, & Massey, 1981) the present research did not distinguish among them.

## Reliability of ASVAB Subtest

Wilfong (1980) examined the test-retest reliability of the ASVAB Form 5, the DoD Student Testing Program version of the ASVAB. Test-retest reliability coefficients for six subtests ranged from .61 to .83. More recently, McCormick, Dunlap, Kennedy, and Jones (1983) investigated the stability of ASVAB 8, 9, and 10 subtests and composites using a group of 57 trainees enrolled in a Job Corps center. As part of a research project, trainees were administered parallel forms of the ASVAB on successive days. After correction for range restriction, the coefficient of stability and equivalence for ASVAB subtests ranged from .72 for PC to .88 for MK with a median reliability of .837.

The purpose of the present research was to examine the reliability of alternate forms of ASVAB 8, 9, and 10, in the Army operational environment where the test-retest interval is at least 30 days.

---

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

## METHOD

### Subjects

The sample consisted of 23,129 Army applicants who tested and retested on ASVAB 8, 9, and 10. For this group of applicants retesting occured in FY81. The test-retest interval ranged from 1 to 11 months.

Records of ASVAB scores are maintained by the Military Enlistment Processing Command (MEPCOM). Through coordination with MEPCOM a data file containing the test scores of applicants was made available to the Army Research Institute. The scores for each examinee were the results obtained on the two most recent ASVAB examinations. Applicants who were asked to retest for verification purposes were not included in the sample.

Of the 23,129 applicants who retested, 17,063 (74%) failed to achieve the minimum AFQT scores required for enlistment on previous testing. In the sample, 18,422 (80%) were male and 4,707 (20%) were females; 10,981 (48%) were white, 10,669 (46%) were black, and 1,479 (6%) were other than black or white.

### Procedure

Pearson product moment correlations were computed between subtest scores obtained by examinees on alternate forms of the ASVAB 8, 9, and 10. However, the variance of scores on ASVAB subtests in the sample was subject to restriction in range. As a consequence of this reduction in variance, reliability estimates were reduced relative to the population values. The following formula derived by Kelley (1923) was employed to obtain reliability estimates for the unrestricted population:

$$R_{XX} = 1 - \frac{s_x^2}{S_x^2} (1 - r_{XX}),$$

where

$s_x^2$ = the variance in the restricted sample; $r_{XX}$ = the reliability in the restricted sample; $S_x^2$ = the variance in the unrestricted population; and $R_{XX}$ = the reliability in the unrestricted population.

The reference population was a national probability sample of men and women aged 16-23 (Office of the Assistant Secretary of Defense, 1982).

## RESULTS AND DISCUSSION

Table 1 presents information pertinent to the evaluation of the reliability of ASVAB subtests. The magnitude of restriction in range can be seen by comparing SD's in the population (column 2) to those SD's obtained for the sample (column 1). Correlations between examinee scores on alternate forms of the ASVAB subtests were computed and adjusted for range restriction using the Kelley formula. The alternate form reliabilities of ASVAB subtests in the population ranged from .62 to .87 (column ). Of the power subtests, the lowest reliabilities were obtained for the two shortest subtests, Paragraph

TABLE 1

Reliability Estimates of ASVAB Subtests

| Subtest | Standard Deviation | | Alternate Forms | | Mean Internal Consistencies | IRT Estimates |
|---|---|---|---|---|---|---|
| | Sample | Population | Uncorrected | Corrected | | |
| GS | 3.74 | 5.01 | .62 | .78 | .86 | .86 |
| AR | 4.17 | 7.37 | .58 | .87 | .91 | .87 |
| WK | 5.69 | 7.71 | .71 | .84 | .92 | .86 |
| PC | 2.79 | 3.36 | .47 | .62 | .81 | .68 |
| NO | 9.10 | 10.99 | .65 | .76 | .78[a] | .71 |
| CS | 13.35 | 16.25 | .59 | .72 | .85[b] | .82 |
| AI | 4.75 | 5.55 | .72 | .80 | .87 | .83 |
| MK | 3.09 | 6.39 | .43 | .87 | .87 | .84 |
| MC | 4.04 | 5.35 | .62 | .78 | .85 | .83 |
| EI | 3.21 | 4.24 | .53 | .73 | .82 | .80 |

[a]Parallel form reliability estimate as reported in Sims & Hiatt (1981).
[b]Mean parallel form reliability estimate as reported in Wilfong (1980).

Comprehension and Electronics Information. Estimates of reliability based on the average KR-20 across forms (Ree et al., 1982) and an Item Response Theory (IRT) model (Bock & Mislevy, 1981) are also presented in Table 1. As expected, the reliability of alternate forms separated in time was somewhat less than the reliability of one form in a single testing session.

In the Army operational environment, selection and classification decisions are based on scores on ASVAB composites. Reliability increases when subtests are combined into composites due to the larger number of items in the composite measures. The alternate forms reliability of ASVAB composites in the unrestricted range exceeded .85 in all cases.

A majority of the retesters in the sample were from the lower quartile of the ability distribution. The Kelley formula used to correct reliability coefficients for range restriction is based on the assumption that error of measurement is invariant with respect to ability variations in the groups being tested. One concern here, is that there are reasons for suspecting that error variance in the group of retesters may be higher than the average error variance in the population. Errors of measurement for examinees in the lower portion of the ability distribution are often higher due to the greater prevalence of random response, carelessness, and inappropriate levels of item difficulty. Research by Bock and Mislevy (1981) indicates that ASVAB subtests provide relatively high precision for measuring aptitudes of individuals at the lower end of the ability distribution. However, error of measurement is somewhat higher at the very lowest ability levels. In addition individuals who chose to retest may be among the most highly motivated to improve their previous test scores. Thus, there are several reasons for suspecting that error variance in the sample of retesters is somewhat greater than error variance in the population. Under these conditions, the Kelley formula yields a conservative estimate of test reliability in the unrestricted population (Lord & Novick, 1968).

As a part of this research, Monte Carlo techniques were used to explore further the accuracy of the Kelley correction formula. The case considered was where the error variance in the restricted sample is greater than the error of measurement in the unrestricted population. Bivariate normal density functions were generated with known degrees of correlation. Heterogeneity of error variance was introduced by lowering the reliability of scores in the restricted sample relative to unrestricted population values. The preliminary results of this Monte Carlo work confirmed the arguments presented above. Results indicated that correction for range restriction using the Kelley formula resulted in more accurate reliability estimates. However, even the reliability coefficients adjusted for range restriction were conservative. The magnitude of conservative bias increase to the extent that error of measurment in the restricted range exceeded error of variance in the population.

To sum up, the alternate form reliability of ASVAB subtests and composites was found to be reasonably high. More work is needed to develop correction formulae that can adjust for complexities of distributions encountered in practice. Further investigation into this area is currently being conducted by the authors of this paper.

## REFERENCES


Bock, R.D., & Mislevy, R.J. (1981). Profile of American Youth: Data quality analysis of the Armed Services Vocational Aptitude Battery. Chicago, IL: National Opinion Research Center.

Kelley, T.L. (1923). Statistical Methods. New York: Macmillan.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

McCormick, B.K., Dunlap, W.P., Kennedy, R.S., & Jones, M.B. (1983). The effects of practice on the Armed Services Vocational Aptitude Battery (Tech. Report 602). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Office of the Assistant Secretary of Defense. (1982). Profile of American Youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Office of the Assistant Secretary of Defense.

Ree, M.J., Mathews, J.J., Mullins, C.J., & Massey, R. (1981). Calibration of the Armed Services Vocational Aptitude Battery 8, 9, and 10 (AFHRL-TR-81-49). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Ree, R., Mullins, C., Mathews, J., & Massey, R. (1982). Armed Services Vocational Aptitude Battery: Item and factor analyses of forms 8, 9, and 10 (AFHRLTR 81-55). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Sims, W.H., & Hiatt, C.M. (1981). Validation of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7 with applications to ASVAB Forms 8, 9, and 10 (CNS 1160). Alexandria, VA: Center for Naval Analyses.

Wilfong, H.D. (1980). ASVAB: Technical supplement to the high school counselor's guide. Fort Sheridan, IL: Directorate of Testing, U.S. Military Enlistment Processing Command.

# Psychometric Properties of the Safety Locus of Control Scale

John W. Jones
Chief Industrial Psychologist
The St. Paul Insurance Companies

Lisa J. Wuebker
Industrial Psychology Graduate Student
Georgia Institute of Technology

## Introduction

The purpose of this study was to further validate the Safety Locus of Control Scale (Jones, 1983). The safety scale is based on the locus of control concept (Rotter, 1966), that is, that individuals either perceive themselves as being in control of their behavior and life events (internally controlled) or being controlled by others (externally controlled).

The safety scale discriminates between internal scorers and external scorers. Internal scorers believe that they are personally responsible for their safety and can take preventive steps to avoid accidents and injuries. Conversely, external scorers believe they have little or no personal control in accident prevention. External scorers believe that accidents and injuries are due to forces outside their control, such as chance, fate, or bad luck. Both the reliability and the validity of the safety scale have been documented (e.g., Jones & Foreman, 1985; Jones & Wuebker, 1984, 1985; Wuebker, Jones, & DuBois, 1985). That is, external scorers typically have more accidents and injuries than internal scorers.

In this study, safety scores were correlated with employees' self-reports of their major on-the-job accidents, along with the estimated medical costs associated with these accidents. It was specifically hypothesized that employees exhibiting low levels of safety consciousness (i.e., external scorers) would report significantly more accidents than employees exhibiting high levels of safety consciousness (i.e., internal scorers). In addition, adverse impact analyses were conducted since the safety scale has applications in employment contexts.

## Method

Employees. Two hundred eighty-three (N=283) hospital employees were sampled. The sample included 58 males, 224 females, and 1 uncoded. Two hundred sixty-three employees were non-minorities, 18 were minorities (i.e., Blacks, Hispanics, Orientals), and 2 were uncoded. Finally, 165 employees were less than 40 years old, 117 employees were 40 years of age or older, and 1 employee did not report age.

---

This paper was presented at the 27th Annual Conference of the Military Testing Association, San Diego, California, October 21-25, 1985.

Safety Scale. All employees completed the 17-item Safety Locus of Control Scale (Jones, 1983; Jones & Wuebker, 1985). Safety scale scores can theoretically range from -17 (external scorers) to +17 (internal scorers). Five different studies support the validity of the scale (Jones, 1985). A Spearman-Brown split-half reliability was computed on the odd versus even test items (Jones & Wuebker, 1985). Obtained reliability equals .85.

Accident Criteria. Employees completed the Employee Injury Profile (Gens, Jones, & Nesbit, 1984), a checklist that assesses major on the job accidents and injuries that employees had during the past 12 months. Employees reported how many times they were injured at work, the type of injuries, and their estimate of the total medical costs for their injuries. Types of injuries assessed included: Major cuts (5 or more stiches), serious burns or scalds, fractures, dislocations, crushed parts of body, amputations, strained back and other muscles, ruptured or herniated discs, torn muscles, hernias, eye injuries, ear injuries, and concussions. Both the total number of injuries reported and the total estimated cost of the injuries served as the criterion measures.

Procedure. Employees completed both the safety scale and the injury checklist during working hours. All employees put their names on the questionnaires. Employees were informed that the study was for research purposes only and that information would not be placed in their personnel files.

Research Design and Statistics. Employees were placed in three different groups based on their safety scores. Employees scoring below the 25th percentile (N=63) were operationally defined as the "Low Safety Consciousness Group". This group had scores that would be below the acceptable cut-off score for the safety scale. Employees scoring below standards exhibited attitudes suggesting that accidents are not preventable since they are due to uncontrollable forces such as chance, fate, or bad luck. Employees scoring above the 75th percentile (N=81) were operationally defined as the "High Safety Consciousness Group". These employees exhibited attitudes suggesting a firm belief that accidents are preventable. Finally, the intermediate group of employees (N=136) formed the "Medium Safety Consciousness Group". Both accident frequency scores and estimated medical costs were analyzed as a function of level of safety consciousness. All analyses were computed with the Statistical Package for the Social Sciences-Portable Computer Version (Norusis, 1985).

## Results and Discussion

Descriptive statistics are summarized in Table 1. Inspection of Table 1 shows that the average number of accidents reported per employee equals 0.71 (SD=1.43). Finer analyses show that 71.8% (N=201) of employees reported no accidents at work, 10% (N=28) reported one accident, 8% (N=23) reported two accidents, and 10% (N=28) reported anywhere from three to seven accidents during the past 12 months.

The average estimated medical costs for injuries equals $187.68 (SD=$537.35). More specifically, 71.8% (N=201) reported no medical costs, 4.6% (N=13) reported $100 in medical costs, 4.6% (N=13) reported $200 in medical costs, and 18.9% (N=53) reported anywhere from $300 to $4,300 in medical costs for the year. The accident frequency scores significantly correlated with estimated medical costs ($r[278] = .66$, $p < .001$).

Table 1

Descriptive Statistics

| Variable | N | Mean | SD | Range |
|---|---|---|---|---|
| Safety Scale | 283 | 1.83 | 6.05 | -14 to 15 |
| No. Accidents | 280 | 0.71 | 1.43 | 0 to 7 |
| Medical Costs | 280 | $187.68 | 537.35 | 0 to $4300 |

Reliability. A Spearman-Brown split-half (odd versus even) reliability coefficient was computed on the 283 safety scale scores. Obtained reliability equals .79. This coefficient is consistent with the Spearman-Brown coefficient of .85 that was obtained in earlier research (Jones & Wuebker, 1985).

Validity. The average number of reported accidents was higher for the Low Safety Consciousness Group compared to the Medium and High Safety Consciousness Groups. This between-groups difference was statistically significant ($F[2/277] = 4.92$, $p < .01$). Employees who have an internal safety locus of control orientation reported reliably fewer accidents than employees with an external locus of control orientation. These findings support the concurrent, criterion-related validity of the safety scale. These findings are summarized in Table 2.

Table 2

Accident Frequency as a Function of Safety Consciousness

| Safety Consciousness | N | Mean | SD |
|---|---|---|---|
| Low | 63 | 1.13 | 1.81 |
| Medium | 136 | .71 | 1.41 |
| High | 81 | .38 | .99 |
| Total | 280 | .71 | 1.43 |

The estimated medical costs for injuries was higher for the Low Safety Consciousness Group compared to the Medium and High Safety Consciousness Groups. This between-groups difference was statistically significant ($F[3/277] = 3.62$, $p < .03$. These results are summarised in Table 3. They also support the validty of the safety scale.

Table 3

Injury-related Medical Costs as a Function of Safety Consciousness

| Safety Consciousness | N | Mean | SD |
|---|---|---|---|
| Low | 63 | $346.03 | $805.79 |
| Medium | 136 | $148.16 | $387.70 |
| High | 81 | $130.86 | $473.79 |
| Total | 280 | $187.68 | $537.35 |

Analysis of Table 4 and Figure 1 reveals that 38% of the Low Safety Consciousness Group were involved in one or more major accidents at work, compared to 28% of the Medium Safety Consciousness Group and 21% of the High Safety Consciousness Group. This between-group difference was statistically significant using a directional Chi-squared analysis ($X2\ [2] = 5.13$, $p < .04$). This finding documents that 17% more of the Low Safety Consciousness Group were involved in accidents compared to the High Safety Consciousness Group. Other factors, such as excessive lifting demands of hospital employees, obviously contributed to accidents and injuries among both internal and external scorers on the safety scale. However, the results of this study indicate that a new psychological construct of "safety locus of control" can be assessed and used to predict behavior.

Table 4

Crosstabulation: Safety Group by Accident Group

| | Safety Consciousness | | | |
|---|---|---|---|---|
| | Low | Medium | High | |
| No Accidents | 39 | 98 | 64 | 201 |
| Accidents (1 or more) | 24 | 38 | 17 | 79 |
| | 63 | 136 | 81 | 280 |

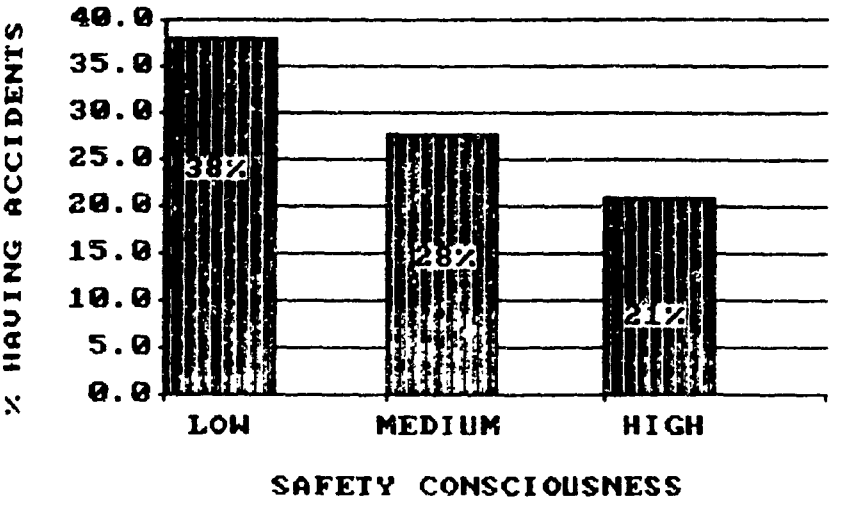



Figure 1. Percent of employees reporting one or more accidents at work as a function of safety consciousness.

Adverse Impact. Inspection of Table 5 shows that safety scores did not significantly differ as a function of sex, race, or age of employees. Analysis of Table 6 shows that the pass-fail rates for employees did not significantly differ as a function of these same demographic variables. For operational purposes, employees in the Low Safety Consciousness Group were considered to have failed the safety scale, while employees in the Medium and High Safety Consciousness Groups were considered to have passed the test.

Analysis of Table 6 shows that 20.6% of males and 22.7% of females scored below standards. In addition, 22% of non-minorities and 22.2% of minorities scored below standards. Finally, 24.8% of the employees who are less than 40 years of age scored below standards compared to 18.8% of the employees who are 40 years of age or older.

Table 5

Adverse Impact Analyses

| | Group | N | Safety Scale Mean | Safety Scale SD |
|---|---|---|---|---|
| SEX: | Males | 58 | 3.14 | 6.33 |
| | Females | 224 | 1.15 | 5.96 |
| | Total | 282 | 1.84 | 6.06 |

F(1/280) = 3.35, p > .05, non-significant

| | | | | |
|---|---|---|---|---|
| RACE: | Non-Minorities | 263 | 1.89 | 6.12 |
| | Minorities | 18 | 1.56 | 5.34 |
| | Total | 281 | 1.86 | 6.06 |

F(1/279) = 0.05, p >.05, non-significant

| | | | | |
|---|---|---|---|---|
| AGE: | <20 | 12 | -1.25 | 6.22 |
| | 20-29 | 62 | 1.97 | 6.34 |
| | 30-39 | 91 | 1.85 | 6.08 |
| | 40-49 | 63 | 2.54 | 5.84 |
| | 50+ | 54 | 1.57 | 5.93 |
| | Total | 282 | 1.84 | 6.06 |

F(4/277) = 1.02, p > .05, non-significant

Table 6

Safety Scale Pass-Fail Rates

| | Group | Safety Consciousness: Low | Medium | High | Total |
|---|---|---|---|---|---|
| SEX: | Males | 12 | 26 | 20 | 58 |
| | Females | 51 | 111 | 62 | 224 |
| | Total | 63 | 137 | 82 | 282 |

Statistic: $X^2$ (2) = 1.03, p > .05, N.S.

| | | | | | |
|---|---|---|---|---|---|
| RACE: | Non-minorities | 58 | 128 | 77 | 263 |
| | Minorities | 4 | 9 | 5 | 18 |
| | Total | 62 | 137 | 82 | 281 |

Statistic: $X^2$ (2) = 0.02, p > .05, N.S.

| | | | | | |
|---|---|---|---|---|---|
| AGE: | <20 | 6 | 3 | 3 | 12 |
| | 20-29 | 15 | 29 | 18 | 62 |
| | 30-39 | 20 | 43 | 28 | 91 |
| | 40-49 | 10 | 34 | 19 | 63 |
| | 50+ | 12 | 28 | 14 | 54 |
| | Total | 63 | 137 | 82 | 282 |

Statistic: $X^2$ (8) = 0.46, p >.05, N.S.

The obtained results are consistent with previous research (Jones, 1985). Employees exhibiting a high level of safety consciousness (i.e., internal scorers) are typically involved in fewer accidents and injuries at work compared to employees exhibiting a low degree of safety consciousness (i.e., external scorers). The Safety Locus of Control Scale provides industry with a standardized evaluation of safety consciousness among employees.

## References


Gens, D., Jones, J.W., & Nesbit, G. (1984) The Employee Injury Profile. St. Paul, MN: The St. Paul Companies.

Jones, J.W. (1983) The Safety Locus of Control Scale. St. Paul, MN: The St. Paul Companies.

Jones, J.W. (1985) The Safety Locus of Control Scale: a review of validation research. St. Paul, MN: The St. Paul Companies.

Jones, J.W., & Foreman, R.J. (1984) Relationship of HFPSI safety scale scores to motor vehicle reports. (Technical Report) St. Paul, MN: The St. Paul Companies.

Jones, J.W., & Wuebker, L.J. (1984) The HFPSI scores of a fatally-injured construction worker. Paper presented at the Fourth Annual Construction Insurance Conference, sponsored by the International Risk Management Institute, Inc., Dallas, Texas, November 13-16.

Jones, J.W., & Wuebker, L.J. (1985) Development and validation of the Safety Locus of Control Scale. Perceptual and Motor Skills, 61, 151-161.

Norusis, M.J. (1985) SPSS/PC+: Advanced Statistics for the IBM PC/XT/AT. Chicago, IL: SPSS, Inc.

Rotter, J.B. (1966) Generalized expectancies for internal versus external locus of control. Psychological Monographs, 80 (Whole No. 609).

Wuebker, L.J., Jones, J.W., & DuBois, D. (1985) Safety locus of control and employee accidents. Proceedings of the Sixth Annual Industrial Psychology Graduate Student Conference, University of Akron, Ohio, April 12-14.


---


Reprints available from Dr. John W. Jones, Chief Industrial Psychologist, the St. Paul Insurance Companies, 385 Washington Street, St. Paul, MN 55102.

# HEALTH LOCUS OF CONTROL BELIEFS AMONG INFANTRYMEN

LAURA W. GRIEGER
U. S. ARMY INFANTRY SCHOOL, FORT BENNING, GEORGIA

Who is responsible for one's health? The answer to this question is a matter of major concern for military leadership at every level. Optimum health and fitness are high priorities among the myriad of concerns and responsibilities of every commander. This study will explore the health locus of control beliefs among Infantrymen to discover whether their beliefs are internal or external and whether demographic factors such as religious preference, ethnic background, age, or level of education account for differences.

Background to support this study begins with a review of the related literature. Development of the Health Locus of Control (HLC) Scale, by Wallston, Wallston, Kaplan, and Maides, was as a unidimensional measure of peoples' beliefs that their health is or is not determined by their behavior (Wallston, Wallston, DeVellis, 1978). The HLC Scale was designed to yield a single score, using a Likert-type scale response, to indicate health-related behaviors as primarily one of the following two types:

a. Health externals: Individuals whose generalized expectancies that the factors which determine their health are such things as: luck, fate, chance, or powerful others, or factors over which they have little control.

b. Health internals: Individuals whose generalized beliefs that control for health is internal and that an individual stays or becomes healthy or sick as a result of his/her behavior.

From both a theoretical and empirical perspective, the exploration of health locus of control is warranted. Rotter's theory that peoples' behavior can be predicted from a knowledge of how they view a situation, their expectancies about their behavior, and how they value the outcomes that might occur as a result of their behavior in that situation (Rotter, 1966) provide the theoretical basis for this study. From the empirical perspective, an increasing number of health researchers have measured locus of control beliefs and have attempted to relate these expectancies to a host of health-related behaviors (Strickland, Wallston, and Wallston, 1978).

Levenson (1974) questioned the conceptualization of locus of control as a unidimensional construct. She reasoned that not only are internal beliefs at right angles to external beliefs, but understanding and prediction could be further improved by studying fate and chance expectations separately from external control by powerful others. Her work further resulted in the decision by K. Wallston and G. Kaplan to reconceptualize health locus of control along multidimensional lines, to develop new scales consisting only of personally worded items, and to create two equivalent forms of the health locus of control scales. Table I presents the items chosen for each of the three scales and form pairs for the Multidimensional Health Locus of Control (MHLC) Scales.

TABLE I

Multidimensional Health Locus of Control (MHLC) Scales Equivalent Forms Internal Health locus of Control Items (IHLC)

| Form A | Form B |
|---|---|
| 1. If I get sick, it is my own behavior which determines how soon I get well again. | 1. If I become sick, I have the power to make myself well again. |

| | |
|---|---|
| 6. I am in control of my health. | 6. I am directly responsible for my health. |
| 8. When I get sick I am to blame. | 8. Whatever goes wrong with my health is my own fault. |
| 12. The main thing which affects my health is what I myself do. | 12. My physical well-being depends on how well I take care of myself. |
| 13. If I take care of myself, I can avoid illness. | 13. When I feel ill, I know it is because I have not been taking care of myself properly. |
| 17. If I take the right actions, I can stay healthy. | 17. I can pretty much stay healthy by taking good care of myself. |

External
Powerful Others Health Locus of Control Items
(PHLC)

| Form A | Form B |
|---|---|
| 3. Having regular contact with my physician is the best way for me to avoid illness. | 3. If I see an excellent doctor regularly, I am less likely to have health problems. |
| 5. Whenever I don't feel well, I should consult a medically trained professional. | 5. I can only maintain my health by consulting health professionals. |
| 7. My family has a lot to do with my becoming sick or staying healthy. | 7. Other people play a big part in whether I stay healthy or become sick. |
| 10. Health professionals control my health. | 10. Health professionals keep me healthy. |
| 14. When I recover from an illness it's usually because other people (for example, doctors, nurses, family friends) have been taking good care of me. | 14. The type of care I receive from other people is what is responsible for how well I recover from an illness. |
| 18. Regarding my health, I can only do what my doctor tells me to do. | 18. Following doctor's orders to the letter is the best way for me to stay healthy. |

External
Chance Health Locus of Control Items
(CHLC)

| Form A | Form B |
|---|---|
| 2. No matter what I do, if I am going to get sick, I will get sick. | 2. Often I feel that no matter what I do, if I am going to get sick, I will get sick. |
| 4. Most things that affect my health happen to me by accident. | 4. It seems that my health is greatly influenced by accidental happenings. |
| 9. Luck plays a big part in determining how soon I will recover from an illness. | 9. When I am sick, I just have to let nature run its course. |
| 11. My good health is largely a matter of good fortune. | 11. When I stay healthy, I'm just plain lucky. |
| 15. No matter what I do, I'm likely to get sick. | 15. Even when I take care of myself, it's easy to get sick. |
| 16. If it's meant to be, I will stay healthy. | 16. When I become ill, it's a matter of fate. |

Descriptive information (means, standard deviations, and alpha reliabilities) for the Multidimensional Health Locus of Control (MHLC) Scales is shown in Table 2.

TABLE 2

Descriptive Data for the Multidimensional Health Locus of Control (MHLC) Scales

| Scale | # of Items | Mean | SD | Alpha |
|---|---|---|---|---|
| IHLC (internality) | | | | |
| Form A | 6 | 25.104 | 4.891 | .767 |
| Form B | 6 | 25.304 | 4.646 | .710 |
| Form A & B | 12 | 50.409 | 9.951 | .859 |
| PHLC (powerful others externality) | | | | |
| Form A | 6 | 19.991 | 5.221 | .673 |
| Form B | 6 | 20.974 | 5.487 | .715 |
| Form A & B | 12 | 40.965 | 10.048 | .830 |
| CHLC (chance externality) | | | | |
| Form A | 6 | 15.574 | 5.751 | .753 |
| Form B | 6 | 15.461 | 5.204 | .691 |
| Form A & B | 12 | 31.035 | 10.204 | .841 |

Conceptualizing locus of control as a multidimensional construct, rather than a unidimensional construct, makes it difficult to think and to talk about types of individuals or situations. Wallston & Wallston (1982) propose a typology of persons based upon possible patterns of scores on the MHLC Scales. The authors readily admit the classification is highly speculative, but balance that with their intentions not to suggest that these are personality types nor do they claim that all possibilities of health locus of control beliefs are captured in the eight patterns displayed in Table 3.

TABLE 3

A Multidimensional Health Locus of Control (MHLC) Scale Typology

Type I
"Pure" Internal

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | X | | |
| Low | | X | X |

Type II
"Pure" Powerful Others External

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | | X | |
| Low | X | | X |

Type III
"Pure" Chance External

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | | | X |
| Low | X | X | |

Type IV
Double External

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | | X | X |
| Low | X | | |

Type V
Believer in Control

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | X | X | |
| Low | | | X |

Type VI
Non-Existent or Rare

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | X | | X |
| Low | | X | |

Type VII
"Yea-Sayer"

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | X | X | X |
| Low | | | |

Type VIII
"Nay-Sayer"

| | IHLC | PHLC | CHLC |
|---|---|---|---|
| High | | | |
| Low | X | X | X |

The usefulness in the typology may be in recognition of the individual differences among health beliefs. Perhaps a revolutionary change needs to take place in the socialization of commanders and health professionals to

the realization that individuals may have to be dealt with differently in preventive health behaviors, adherence to medical regimens, response to symptoms, and interaction with health care settings.

There are no hypothesis statements in this research. The purpose of this research is to gather demographic data in a target population heretofore unstudied and unreported in MHLC Scale literature. The findings will serve as a basis for future theoretical and empirical research: use of the Multidimensional Health Locus of Control (MHLC) Scale as a dependent variable relating health locus of control to demographic variables among Infantrymen.

Some of the considerations leading to the conduct of this research are:

1. What are the demographic characteristics of this target population?
2. What are the results of the MHLC Scales within this population?
3. Do the demographic variables form any recognizable patterns with the MHLC Scales?
4. Do the results of the typology form any recognizable patterns with the demographic variables?
5. What hypotheses can be formulated for future research, based on the differences and/or similarities of patterns?
6. Can recommendations, based on the results of this research, be made to commanders and managers of human resources within the Infantry to promote health-related behaviors?

## METHOD

Subjects:

The target population for this study consists of the annual student load in the subject area of leadership at the U. S. Army Infantry School which includes the courses and number of students displayed at Table 4.

TABLE 4

U.S. Army Infantry School Courses and Projected Student Load

| Course Title | Projected Annual Student Load |
|---|---|
| Officer Candidate School | 1,200 |
| Infantry Officer Basic | 2,200 |
| Infantry Officer Advanced | 1,025 |
| Maneuver Combat Arms Advanced Noncommissioned Officer | 1,230 |
| Total | 5,655 |

The unrefined sample consists of one iteration of each course, 14.32 percent of the projected total annual student load. The subjects are not identified by name nor by class roster number in order to provide complete anonymity. The sample is further refined by identifying foreign students and non-Infantry and eliminating their questionnaires from the data.

The refined sample consists of 577 students, 10.20 percent of the projected total annual student load and 71.23 percent of the total number of questionnaires distributed.

The high rate of useable questionnaires may be attributed to the fact that the questionnaire was administered by instructors during class and although completion was voluntary, the classroom setting may have provided impetus to complete the questionnaire.

The questionnaires discarded were rejected for the reasons shown in Table 5.

TABLE 5

Reason for Questionnaire Rejection

| Demographic Portion | MHLC Scale Position |
|---|---|
| Incomplete | Incomplete |
| Selection of more than one response per item | Handwritten responses |
| Handwritten responses | Selection of more than one response per item |
| Foreign students | |
| Women (Officer Candidate course only) | |

Materials:

Forms A and B of the MHLC Scale are used to gather health locus of control data. Demographic data items displayed at Table 5 are compiled from previously validated items used by the Directorate of Evaluation and Standardization for use with the same target population.

TABLE 6

Demographic Variables

| | |
|---|---|
| Age | Ethnic background |
| Pay grade | Educational level |
| Religious preference | Course attending |

Procedure and Design:

Data collection is conducted in the following manner. During classroom leadership instruction, instructors explain the purpose of the questionnaire as part of research data collection and request the students not provide their name nor roster number to insure complete anonymity. The students are instructed to read and follow the written instructions provided on the questionnaire. Completed questionnaires are collected by the instructors and returned to the researcher for tabulation.

The questionnaires are screened. Foreign students, non-Infantrymen, women candidates in Officer Candidate School, mismarked, and incomplete questionnaires are removed.

Scoring consists of three phases. First, each MHLC Scale is scored. Secondly, median splits are determined for each of the three MHLC Scales. Thirdly, each subject is type classified into one of the 8 "types" according to their pattern of being above ("high") or below ("low") the median splits on the scales. Any score falling on the median is decided as above or below by a coin toss, heads=above and tails=below.

Each subject's scored data and demographic variables are now ready to be recorded on ADP transcript paper for key punch, verification, and duplication of the data deck.

Results and Conclusions:

The results are a compilation of frequency counts. The demographic variables results are:

Age (in years):

| Under 20 | 20-24 | 25-29 | 30-34 | 35 & Over | Sample Size |
|---|---|---|---|---|---|
| 0 | + 172 | + 253 | + 131 | + 21 = | 577 |

Pay Grade:

| E5 | E6 | E7 | 01 | 02 | 03 | Sample Size |
|---|---|---|---|---|---|---|
| 79 | + 87 | + 0 | + 195 | + 172 | + 44 = | 577 |

Religious Preference:

| Protestant | Roman Catholic | Jewish | Other | Sample Size |
|---|---|---|---|---|
| 197 | + 223 | + 0 | + 157 = | 577 |

Ethnic Background:

| Black | White | Hispanic | Asian Amer/ Pac Is | Amer Indian/ Alaskan Nat | Other | Sample Size |
|---|---|---|---|---|---|---|
| 167 | + 390 | + 16 | + 2 | + 0 | + 2 = | 577 |

Highest Civilian Education Degree:

| H.S.Dip/Equiv | A.S. | B.A./B.S. | Master | Ph.D | Other | Sample Size |
|---|---|---|---|---|---|---|
| 117 | + 20 + | 359 | + 54 + | 0 | + 27 = | 577 |

Course Attending:

| Off. Cand | Off. Basic | Off. Adv. | NCO Adv. | Sample Size |
|---|---|---|---|---|
| 81 | + 195 + | 216 | + 85 = | 577 |

Results for the MHLC Scales are displayed next. A subject was classified into a category by their highest score among the three scales.

| Internal | External: Chance | External: Powerful Other | Sample Size |
|---|---|---|---|
| 521 | + 44 + | 12 | = 577 |

If a subject scored the same on two or more scales the decision was made to discard that subject.

The results of the typology are as follows:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Sample Size |
|---|---|---|---|---|---|---|---|---|
| 207 | + 21 + | 53 | + 87 + | 45 | + 73 + | 39 | + 52 = | 577 |

The frequency counts reveal some interesting findings. The next step will be to apply the statistical technique of discriminant function analysis to the data using the Statistical Package for Social Science[X] (SPSS), 1983 to determine if a prediction equation results. Results of the research are programmed for 2d Qtr FY86. You are invited to contact the author if interested. Disclaimer: The content of this paper is the sole responsibility and opinion of the author.

REFERENCES


Levenson, H. (1974). Activism and powerful others: Distinctions within the concept of internal - external control. Journal of Personality Assessment 38; 377-383.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80. (Whole No. 609).

Strickland, B.R. (1978). Internal - external expectancies and health-related behaviors. Journal of Consulting and Clinical Psychology, 46, 1192-1211.

Wallston, K.A., Wallston, B.S., & DeVellis, R. (1978). Development of the Multidimensional Health Locus of Control (MHLC) Scales. Health Education Monographs, 6, 160-170.

Wallston, K.A., Wallston, B.S. (1981). Health Locus of Control Scales. In H. Lefcourt (Ed.), Research with the locus of control construct. NY: Academic Press, 1, 189-243.

Wallston, K.A., & Wallston, B.S. (1982). Who is responsible for your health? The construct of health locus of control. In G. Sanders & J. Suls (Eds.), Social Psychology of Health and Illness. Hillsdale, NJ: Erlbaum, 65-95.

Wallston, K.A., Smith, R.A., King, J.E., Forsberg, P.R., Wallston, B.S., & Nagy, V.T. (1983). Expectancies about control over health: Relationship to desire for control of health care. Personality and Social Psychology Bulletin, 9, 377-385.

Wallston, B.S., & Wallston, K.A. (1984). Social psychological models of health behavior: An examination and integration. Chapter in A. Baum, S. Taylor, & J. E. Singer (Eds.), Handbook of Psychology and Health, Volume 4: Social Aspects of Health. Hillsdale, NJ: Erlbaum.

# Classifying Military Offenders: Application of the Megargee MMPI Typology

Mark L. Paris and Gary E. Brown
United States Disciplinary Barracks

Megargee and Bohn (1979) developed an inmate classification system utilizing the Minnesota Multiphasic Personality Inventory (MMPI). The system was originally designed to enable clinicians to assign individual MMPI profiles to prototypical classifications, to facilitate empirical establishment of relationships between classifications and probable future behaviors, and to enhance treatment of young offenders by targeting specific treatment modalities to appropriate inmate profile types based on the system.

Megargee and Bohn developed their taxonomy in three phases: (1) a statistical analysis of three samples of 100 MMPI protocols to determine the number of naturally occurring subgroups; (2) the matching of subgroups from each sample with each other, and the writing of rules to describe the larger groups resulting from the matching; and (3) the classification of a new sample of protocols according to the new rules.

Eventually, a large, broad-based study of the new system was undertaken as part of an extensive longitudinal research program at the Federal Correctional Institution, Tallahassee, Florida (Megargee, 1974). Of 1214 men subjected to classification by the system, 770 (63.4%) were classified through the use of computer programs which employed profile-type rules. The remaining 394 (32.4%) were made up of ties among classifications (248), possibly invalid profiles (79) or profiles which did not meet the criteria for inclusion in any of the classifications (117). Ultimately, all but 50 cases were classified, and these remaining 1164 individuals were studied along such dimensions as academic and intellectual level, characteristics of the developmental family, measures of interpersonal relations and adjustment, psychologists' objective measures, and various measures of institutional adjustment, including work performance, number of disciplinary infractions, time spent in Maximum custody and detail evaluations. In addition, recidivism data were also studied.

Significant relationships were established between MMPI classifications and various measures of adjustment, recidivism and personality. Furthermore, Megargee and Bohn (1979) proceeded to make treatment recommendations specific to the various classifications, based on their understanding of the implications of the data which had been gathered for each classification.

In an attempt to extend Megargee and Bohn's work to the military correctional setting, an MMPI inmate classification system was begun in February 1985 at the United States Disciplinary Barracks, Fort Leavenworth, Kansas.

## Method

The computer program used to classify the inmates at the United States Disciplinary Barracks (USDB) was the one written by Meyer and Megargee (1977). Inmates were administered the MMPI (all 566 items) on Monday of the second week of their incarceration at the institution. They were informed that testing, as part of their treatment, was required, but that they would receive test feedback and that the results would be seen and used only by mental health personnel, usually their own counselor.

The MMPI's were computer-scored, with the resultant Megargee classifications generated. The program either printed the raw scores, T scores and Welsh code for each individual, as well as the appropriate classification (Able, Baker, Charlie, Delta, Easy, Foxtrot, George, How, Item, or Jupiter), or else, it would print out the profile. The profile would be printed out in the event that (1) the profile was tied among several classifications, (2) it fulfilled the criteria for none of the classifications, or (3) F>99T. In these cases, the program left it open to the clinician's discretion to break ties, and otherwise classify questionable profiles.

Interpretive reports were created for each classification by utilizing the empirical data of Megargee and Bohn (1979). These reports were made available to the individual's future counselor for his/her use in treatment planning.

## Results

Data were crosstabulated, specific offense by Megargee classification, and comparisons were made between frequencies of classification found by Megargee (Megargee, 1974) and by ourselves. By chi-square analysis, three comparisons were found to be statistically significant. In general, however, although our sample size was 1/3 that of Megargee's, Table 1 shows that the pattern of similarity was quite strong.

Table 1

USDB vs FCI inmate classification (percentages)

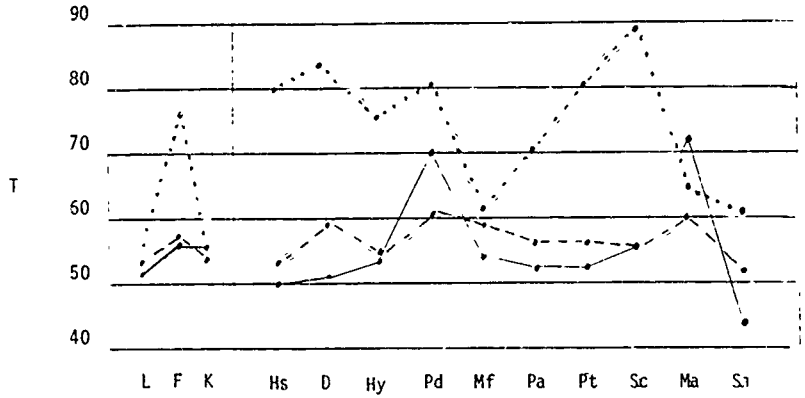
| Classification | USDB (N=390) | FCI (N=1214) |
|---|---|---|
| Able | 12.3 | 13.3 |
| Baker | 1.5 | 1.8 |
| Charlie | 4.9 | 4.6 |
| Delta | 5.1 | 7.0 |
| Easy | 3.8 | 3.2 |
| Foxtrot** | 1.5 | 5.7 |
| George | 5.6 | 5.9 |
| How | 5.9 | 7.7 |
| Item** | 25.6 | 13.0 |
| Jupiter | 1.3 | 1.1 |
| F>99T* | 2.6 | 6.5 |
| No criteria met | 7.4 | 9.6 |
| Ties | 22.3 | 20.4 |

*p<.001
**P<.0001

Inspection of Table 1 reveals that Item, the most psychopathologically benign profile in the typology, is twice as likely to be found in the USDB population as in that of the civilian FCI. Additionally, more than twice as many possibly invalid profiles are found in the FCI group as in that of the USDB. Finally, Foxtrot, the third most psychopathological of the profile types is found more than three times more frequently in the FCI as in the USDB populations.

In addition, comparison of the number of subjects falling into the various classifications revealed a significant effect among the cells, $X^2 = 301.89$, df 9, $p < .0001$. It is noteworthy that the modal profile type, Item, is essentially flat, with the mean T-scores falling in sub-clinical ranges; the next most frequent profile in the sample, Able, is the "4-9" profile type that is traditionally associated with correctional populations (see Fig. 1).

Figure 1
Mean profiles for types Able, How and Item
(after Megargee and Bohn, 1979)




Item ------
Able ———
How . . . .

When classifications were examined with respect to specific offenses (see Table 2), it was found that among three of six offense types studied, Item was found significantly more frequently than the next most frequent type (Item was also found to be the most frequent type for all other offenses, though not to a statistically significant degree). With respect to drug offenses, a bimodal distribution was found: while the frequency of Item was significantly greater than the next most frequent type, Able ($X^2 = 3.93$, $p = .036$), the frequency of Able was found to be significantly greater than the next most frequent type, Charlie ($X^2 = 6.59$, $p = .004$). Similarly, re: rape, a significant difference was found between the frequency of Item and those of either Delta or How ($X^2 = 8.45$, $p = .001$); however, Item did not differ significantly from Able, and Able did not differ from Delta or How. Thus, taken together, Item and Able tend to predominate among the other groups. Among child sex offenders, Item again significantly differed from its closest competitor, How ($X^2 = 5.84$, $p = .008$); How did not differ significantly from the next most frequent group, Delta; however, the data suggest that such a bimodal effect might be revealed as the sample size increases.

Table 2
Breakdown of selected types by offense (percentage)

| | Item | Easy | Able | Delta | Charlie | How |
|---|---|---|---|---|---|---|
| Murder (26) | 23.1 | 7.7 | 15.4 | 7.7 | 3.8 | 3.8 |
| Assault (37) | 16.2 | 5.4 | 2.7 | 5.4 | 2.7 | 13.5 |
| Rape (54) | 25.9 | 3.7 | 14.8 | 5.6 | 3.7 | 5.6 |
| Child (58) | 32.8 | 1.7 | 3.4 | 8.6 | 3.4 | 13.8 |
| Larceny (52) | 21.2 | 5.8 | 13.5 | 5.8 | 3.8 | 1.9 |
| Drugs (126) | 27.0 | 4.0 | 16.7 | 2.4 | 6.3 | 2.4 |

## Discussion

Megargee, et al. set out, with their MMPI classification system, to create a method of predicting behavior within correctional populations. Inherent in this project was the implication that the phenomenon of "classifiability' of prison inmates should be a generalizable one: such an hypothesis could be strengthened by fairly consistent findings of profile type frequency distributions among various correctional institutions. The present pattern of data, with a few exceptions, replicates the findings of Megargee and Bohn (1979) with respect to the relative levels of psychopathology in our inmate population. Collectively, the data indicate that in military as in civilian correctional settings, a huge majority of inmates' personality profiles fall into one of ten basic types, but that in a military institution, one is less likely to find measurable psychopathology or attempts at dissimulation. It is the "flat" or basically normal profile that is the modal finding in military confinement, followed by the expected characterological "4-9" profile of Able. Although this pattern of findings was fairly consistent over several offenses, the child sex offender continues to challenge our expectations; such offenders generally present Item-type profiles, but are increasingly yielding high levels of psychopathology as well, suggesting a complex branching of etiologies going in different directions.

Future research issues will consider therapeutic implications of labeling/classification; that is, whether the process of labeling for therapy is facilitative of a self-fulfilling prophecy would seem of paramount importance if a classification system such as this is to be used for anything other than research, placement or security purposes.

## References

Megargee, E.I. (1974). Applied psychological research in a correctional setting. Criminal Justice and Behavior, 4, 211-216.

Megargee, E.I. & Bohn, M.J. (1979). Classifying Criminal Offenders. Beverly Hills, California: Sage Publications, Inc.

Meyer, J. Jr., & Megargee, E.I. (1977). A new classification system for criminal offenders, II: Initial development of the system. Criminal Justice and Behavior, 4, 115-124.

The views expressed in this article do not necessarily represent those of the USDB, the U.S. Army or the Department of Defense.

# AUTOMATED SCHEDULING OF ARMY UNIT TRAINING

Dwight J. Goehring
Roland J. Hart

U.S. Army Research Institute Field Unit
Presidio of Monterey, California

Computer assistance can reduce the complexity of military training scheduling for planners and lead to better planning of training. Improved training planning can make more efficient use of time and resources, and thus contribute to improving readiness. However, in order to be helpful, computer automated assistance of training scheduling must accommodate the full diversity of the military problem. Our objective is to explore the feasibility of a particular heuristic approach called simulated annealing, for solving the Army scheduling problem.

## I. The Army Training Schedule Problem

The Army training scheduling problem is complex because it occurs across different echelons and involves three different types of calendars/schedules. Priorities of higher echelons almost always predominate over those of lower echelons. In addition, many constraints exist for the Army training scheduling problem. Training and non-training tasks require time and resources. Resources can be either reusable (e.g., ranges) or expendable (e.g., ammunition). Further, activities may have temporal constraints, such as requiring some tasks to be trained ahead of others, or co-occurrence constraints requiring some tasks to be trained simultaneously with others. There are also command priority constraints at each echelon which affect scheduling requirements. In addition, the priority given training varies at different times. Automated scheduling must accommodate these hetereogenous constraints.

## II. Use of Simulated Annealing to Solve the Scheduling Problem

An optimization methodology is needed to solve the Army training schedule problem. An optimization methodology provides (a) a definition of what constitutes a good schedule, (b) a way of measuring numerically how good a schedule is, and (c) a method to search for numerically good schedules. Simulated annealing (Kirkpatrick, Gelatt & Vecchi, 1983) was selected as an appropriate method because of the match between the capabilities of the methodology and the characteristics of the scheduling problem. First, it can handle large problems. The training scheduling problem for an entire Army division is large. In addition, it can handle virtually any type of constraint. This feature is important because the Army training scheduling problem includes a wide variety of constraints. The approach is flexible in the sense that constraints can be changed or added and the method can still work. Flexibility is important because frequent modification is expected in an operational environment. Another feature of simulated annealing that matches the Army training scheduling problem is size of the solution space. Application of simulated annealing is appropriate when multiple "good" solutions are acceptable as opposed to a uniquely "best" solution.

Further, for large combinational problems like the scheduling problem, it becomes impossible to check all possible combinations, in search of the one combination that minimizes the cost function, because of time requirements. The time requirements for such problems increase exponentially with N (i.e., number of activities to be scheduled). Rather than search all combinations, optimization methods use heuristics to minimize cost functions. Such heuristics can be generally classified as either "divide-and-conquer" or iterative improvement strategies (Kirkpatrick, Gelatt & Becchi, 1983). Simulated annealing combines elements of both strategies. As applied to scheduling, the problem is divided into subproblems (involving common levels of "importance") and iterative improvement occurs through the application of the annealing concept.

Finally, the simulated annealing algorithm does not require a set computer time to obtain a solution. The solution will get better on the average the longer the program runs. A computer run can be terminated when a point of diminishing returns is achieved.

The simulated annealing heuristic operates by analogy to annealing in physical systems. Annealing in a physical system involves repeated heating and cooling of liquids or solids. When a liquid is hot there is a rapid random movement of the molecules making up the liquid. Heating a liquid increases the rapidity of the random movement of molecules. By contrast, cooling a liquid slows the random movement of molecules until they reach the point where they are frozen in place. Molecules frozen in place are considered "cold."

As applied to a scheduling system, the molecules are activities or events that must be scheduled. Random movement involves the random assignment of activities to times. The concept of temperature is linked to the cost function, and refers to the degree of random movement permitted relative to the cost function. High temperature is associated with high values of the cost function, low temperature is associated with low values of the cost function. High temperature means that random assignment of activities to times is permitted no matter how large the cost function gets as a result of the assignments. That is, assignments are not constrained by the cost function. On the other hand, cold temperature means that only some random assignments of activities to times are permitted, namely, assignments that reduce or at least do not increase the cost function.

The operation of the simulated annealing heuristic involves successively "heating" and "cooling" the system in search of a "good" schedule defined by a low cost function. The process of successively "heating" and "cooling" the system is accomplished to avoid the problem of getting stuck in local minimums. This problem is a common one for problems based on iterative improvement (e.g., greedy algorithms). The simulated annealing heuristic requires the creation of an appropriate "annealing schedule" involving the extent and frequency of heating and cooling.

In addition to creating an appropriate annealing schedule, an algorithm is necessary to cool the system. (Creating a hot system is not difficult since heating simply entails unconstrained random assignment of activities to times.). Cooling the system, however, requires development of an algorithm, created especially for the problem at hand, that constrains the random

assignments permitted to those producing successively lower values of the cost function, as the system is cooled. In order to "cool" a scheduling system, those tasks that have the greatest potential for increasing the cost function are scheduled first before the schedule gets too full to include them. Then tasks that have less potential for increasing the cost function are successively scheduled. Unconstrained tasks that minimally add to the cost function are scheduled last. This strategy increases the probability that important activities and constraints are accommodated in the resultant schedules.

## III. Illustrative Application of Simulated Annealing

The Army scheduling problem was delimited by focusing on scheduling at the battalion level. This level was selected because it contains scheduling features from all echelons. If the annealing heuristic is effective at this level, one can assume that it can be applied at all levels with the appropriate modifications for each echelon.

All five companies within a single battalion are scheduled simultaneously in the test program. These five companies inherit training activities and events from the Long- and Short-Range Calendar created at higher echelons. The test program uses one-hour units of time for activities and forty-hour weeks. The initial test program has only one annealing schedule for "cooling" the system. This schedule flows from the top to bottom following the activity importance criterion.

The test program employs all of the important categories of scheduling constraints: (a) inheritance of activities from higher echelons and "fixing" activity times, (b) interschedule conflicts of both renewable and expendable resources, (c) company-level training activity priorities, (d) different unit priorities assigned to different companies, and (e) intra-schedule conflicts based on temporal constraints for activities (i.e., before/after, immediately before/immediately after), and temporal ordering of sequences of contiguous activities.

The simulated annealing test program was written in FORTRAN 77 on a VAX 11/780 computer. An overview of the data structures and flow of control of the simulated annealing test program is shown in Figure 1. Blocks A through E in Figure 1 depict five data structures containing the necessary input information.

Initial company-level Training Schedules developed by company commanders are simulated in Blocks F, G, and H of Figure 1. These schedules identify tasks for the week, assign priorities to the tasks ("high" versus "regular" priority), and place activities in temporal order consistent with prerequisite relationships (before/after). These initial schedules represent recommendations from a lower echelon for training schedules that are passed upward. At battalion level, resourcing is accomplished and conflicts are resolved. As depicted in Figure 1, this conflict resolution is accomplished by the passage of the initially recommended schedules to the simulated annealing algorithm in Block I. The simulated annealing algorithm produces the final set of training schedules for the companies within a battalion. In addition, lists of

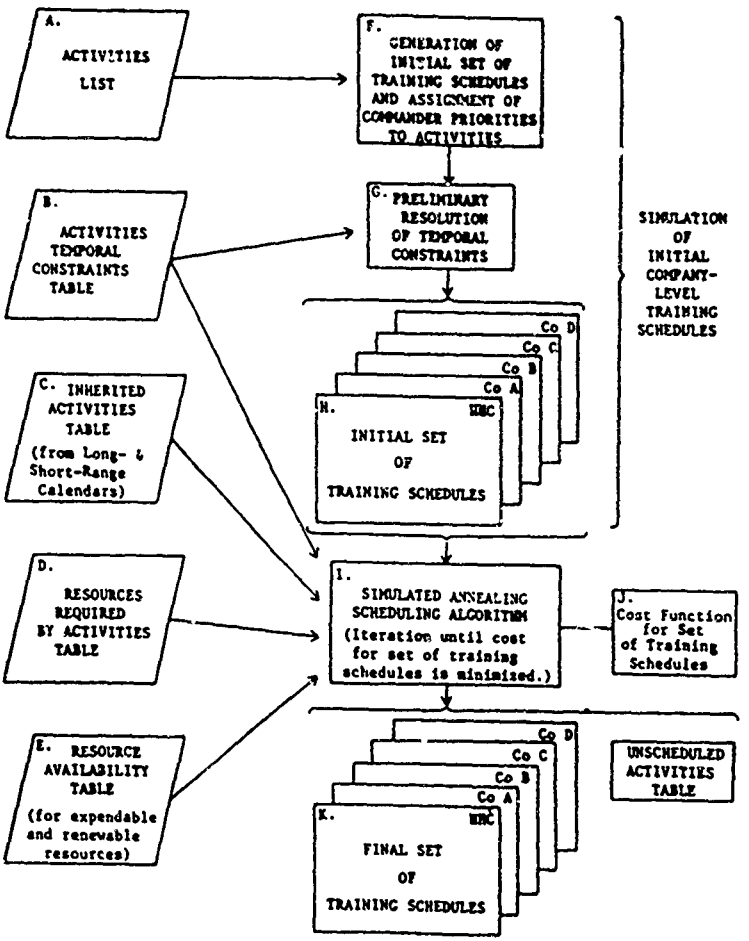



Figure 1. Data Structures and Flow of Control for Simulated Annealing Test Program

activities from the initially recommended schedules that could not be accommodated in the final schedules are given along with their level of importance as defined by their contribution to the cost function.

The cost function for the test program was formulated on the assumption that it is costly to fail to schedule activities. Further, it is more costly to fail to schedule longer than shorter activities and more important than less important activities.

More Specifically, let C represent cost and K represent weights reflecting the relative importance of unscheduled activities. Let U represent the time associated with unscheduled activities. Let i represent the activity number among m unscheduled activities within a Training Schedule, and let j represent the Training Schedule number that varies from 1 to n, the number of Training Schedules. Since there is a Training Schedule for each company, there are five Training Schedules per battalion (n = 5). Given these definitions, the cost function can be written by Equation 1 as:

$$C = \sum_{j=1}^{n} \sum_{i=1}^{m} K_{ij} U_{ij}$$

In order to create the algorithm that "cools" the system, activities need to be partitioned and sequenced for scheduling. In the test program, activities were partitioned and sequenced by the echelon level of the scheduling decision. Since activities appearing on the Long- and Short-Range Calendars are created above the battalion level, they are inherited at the battalion level and are scheduled first. (See Block 2 of Figure 2.). For the test program, it is assumed that resource conflicts and temporal constraints of inherited activities have been previously resolved. Inter-schedule conflicts between company-level Training Schedules are resolved next as shown in Block 3 of Figure 2. These conflicts generally involve conflicts over resources. Company-level training priorities (assigned at company level) are scheduled next as shown in Block 4. Then, temporal constraints yielding intra-schedule conflicts are resolved. Finally, any free blocks of time are filled with unconstrained activities yielding an end set of Training Schedules. (This decision-making sequence is an oversimplification of the real scheduling process in some respects, excluding some details, but it is sufficiently accurate and complete for describing the test program).

The heating and cooling process occurs iteratively for all company-level Training Schedules in a battalion, in search of a battalion set of Training Schedules with a low overall cost function. The weights assigned to the cost function ($K_i$), that determine the importance of unscheduled activities, correspond in order to the "cooling" sequence just outlined.

The operation of the algorithm was sufficiently successful in simulating the performance of Army personnel experienced in training to suggest that the approach is a promising one warranting further research.

## IV. Reference

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220, 671-680.

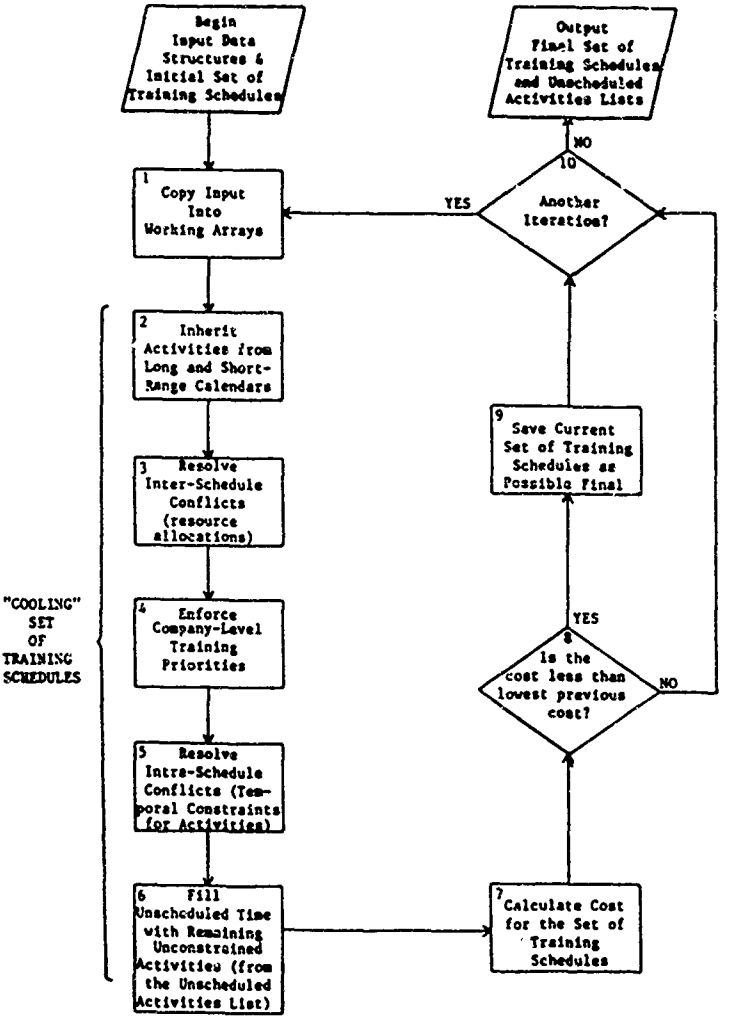



Figure 2. Operation of the Simulated Annealing Scheduling Algorithm

# Supporting the Changing Role of Army Collective Training Developers

Larry L. Meliza
U.S. Army Research Institute Field Unit
Presidio of Monterey, California

Army branch schools prepare Army Training and Evaluation Program (ARTEP) documents as guides to unit collective training. The design of ARTEP documents has evolved considerably in their short, ten-year history to more effectively meet the information needs of ARTEP users. The goal of this paper is to describe 1) the evolution of the ARTEP and 2) efforts to address the growing complexity of the ARTEP developer's job through the careful application of computer technology.

## Evolution of ARTEP Documents

ARTEP documents were initially designed to provide "what to train" guidance by outlining potential training requirements. These documents identify the missions and subordinate collective tasks (i.e., tasks requiring two or more individuals to perform) which a unit (e.g., mechanized infantry battalion) and its subordinate elements (e.g., companies, platoons, squads and teams) might be expected to perform. Each collective task is "defined" by indicating the tactical situation calling for its performance (conditions) and by providing task performance guidance (referred to as Standards). ARTEP users are responsible for selecting the "what to train" guidance which applies to their particular unit and combining this guidance with separately provided "how to train" guidance to develop unit training plans. During three evolutionaly phases, the ARTEP design concept has been expanded to incorporate the application of "how to train" guidance to specific missions and tasks, while leaving ARTEP users with the flexibility to select training requirements appropriate to their particular unit. (TRADOC PAM 310-8, 1981; TRADOC REG 310-2, 1982; TRADOC REG 310-2 (Test-Revised, 1985).

In the first phase of ARTEP evolution, collective/individual task matrices were added to the ARTEP document. These matrices identify individual skills prerequisites to training on specific collective tasks at the smallest unit levels (e.g., squad and platoon). This addition to the ARTEP provided the framework of a progressive unit training plan by beginning to integrate individual skills training and collective training.

The second phase of ARTEP evolution involved adding "Drill" documents to further support a progressive/building block approach to unit training at the smallest unit levels (e.g., squad and platoon) (Hiller, Hardy and Meliza, 1984). Drills are training exercises which address tasks/subtasks warranting repetitive practice to the point of overlearning. Training on drills is intended to prepare a unit for training on the more complex missions/tasks in which the drills are embedded. Since one of the criteria for selecting tasks/subtasks to be addressed by Drills is that they have broad applicability across missions, Drills have high progressive training value. Drill documents provide specific "how to train" guidance for each Drill including "set-up directions", resource requirements and "coaching points". In addition, Drill

documents provide Drill/Mission Task applicability matrices and individual skills/Drill matrices to assist ARTEP users with integrating Drills into unit training plans.

The third, on-going phase of ARTEP evolution involves the addition of ARTEP Mission Training Plan (AMTP) documents. AMTP documents represent a further extension of the progressive training plan concept by defining building block exercises called "Situation Training Exercises (STXs)" for platoon-level and above. STXs address training requirements within portions of missions (i.e., STXs often address two or more collective tasks). Many STXs are applicable to more than one mission and thus these exercises tend to have high progressive training value.

As in the case of "drills", AMTPs provide specific "how to train" guidance for each STX and matrices/figures illustrating descriptive unit training plans. In addition, STXs also provide guidance for training prerequisite leader skills.

To gain a more complete picture of how AMTP documents provide a progressive unit training plan it is important to consider that a particular unit type is intended to have different AMTP documents for various echelons. For example, a mechanized infantry battalion would have separate AMTP documents for battalion, company and platoon level. The platoon-level AMTP is intended to help a unit progress from individual skills training of soldiers to company level training, while the company level AMTP document is intended to help a unit progress from platoon level training to battalion level training.

Incorporating applications of "how to train" guidance into the ARTEP removes certain analytical tasks from the shoulder of the ARTEP user, and it helps to standardize collective training across units. Without the incorporation of such guidance, it had been necessary for each unit trainer to independently develop their own Drills, STXs and progressive unit training plans. This was a large burden for unit personnel and the quality of training programs was therefore uneven.

## The Changing Role of the ARTEP Developer

The current phase of ARTEP evolution is at an experimental stage. Each of eighteen schools has prepared a prototype platoon-level AMTP to support Army-wide field testing of the AMTP concept. The exact design features of the current generation of improved ARTEP documents will be adopted, based upon the results of this field test, and then schools will be expected to begin full implementation of the new ARTEP design concept. Implementation of this concept will involve a substantial increase in the ARTEP development workload.

Under the original ARTEP concept, the analytical work of ARTEP developers was largely complete when tactical doctrine ("how to fight" manuals) and other sources of information had been analyzed through the process of front-end analysis (FEA) to identify the missions and tasks to be addressed by the ARTEP. The final ARTEP document closely resembled the outcome of the FEA process. With the evolution of ARTEP design, ARTEP developers assume the responsibility for analyzing training-relevant features of missions and tasks

and using the results of these analyses to develop training exercises (Drills and STXs) and incorporate training exercises within a descriptive unit training plan.

The evolution of ARTEP greatly expands the ARTEP development audit trail. In effect, the old audit trail involved linking tactical doctrine, FEA results and ARTEP products, in a situation where FEA results and ARTEP were virtually identical. Under the latest ARTEP concept the audit trail involves linking tactical doctrine and FEA results to a variety of ARTEP products. Further, the new audit trail involves linking the various ARTEP products to each other.

The growth in the ARTEP development workload associated with implementing the latest ARTEP design concept comes at a time when many schools are beginning to feel the effects of an ongoing Army Force Modernization effort (i.e., the adoption of new technology and organizational structures). Force modernization, in itself, increases the number of unit types for which ARTEPs must be developed and it increases the frequency with which certain ARTEP documents need to be revised to reflect changes in tactical doctrine. The careful application of computer technology to the ARTEP development process is required to help service schools effectively implement improved ARTEP concepts and deal with changes required by the Force Modernization program.

Mechanisms Being Used to Support the Changing Role of ARTEP Developers

ARI and the U.S. Army Training Board (ATB) are developing a Computer-Based ARTEP Production System (CAPS) to support the preparation and revision of AMTP/Drill documents. The portion of the AMTP/Drill development process to be addressed by CAPS begins with the FEA and continues to the point where camera-ready copies are available for printing. In addition, the CAPS concept is intended to support prompt revision of tactical doctrine literature, FEA and ARTEP products in response to changes in tactical doctrine. In brief, a portion of the CAPS database would include tactical doctrine literature and FEA, and the system would guide collective training developers in using the database to develop AMTP/Drill documents.

A preliminary CAPS concept design study indicated that a Relational Database Management System (RDBMS) might serve as an effective core for a CAPS (Bloedoern, Crooks, Saal, Merrill, Meliza and Kahn, in press). A RDBMS stores data in a flexible, tabular fashion and allows information to be extracted from within or across tables using brief, one-line commands.

A RDBMS appears to be particularly well suited to the ARTEP development process, because it can accommodate the complex ARTEP development audit trail in an economical fashion. For example one RDBMS table might contain listings of the references from tactical doctrine for each collective task identified through FEA, and another table might contain a listing of the collective tasks addressed by specific STXs. Through the use of a one-line command, the information within these two tables can be reorganized to provide a table listing the references from tactical doctrine for each STX, as needed.

## Application of a RDBMS to AMTP/Drill Preparation

A major goal of applying a RDBMS to AMTP/Drill preparation is to reduce the complexity of work. In general, a RDBMS has the potential to facilitate the complex work involved in preparing AMTP/Drill documents in two ways. First, formal training development guidance and less formal job aids (e.g., "rules of thumb" for making decisions and examples of the application of these rules) might be contained within the database in a manner which relates specific job aids/guidance to appropriate jobs in the ARTEP development process. Second, the analytical capabilities of a RDBMS (i.e., rapid reorganization of data to meet specific information needs) might be used to facilitate complex decisions. The decision processes involved in AMTP/Drill preparation were studied in considerable detail with the goal of reorganizing this process, as necessary, to take advantage of potential applications of a RDBMS to ARTEP development.

## Application of a RDBMS to AMTP/Drill Preparation

Work complexity, in this instance, is defined in terms of 1) the number of data elements to be considered and 2) the source of data elements. Certain decisions only require reorganizing information found within tactical doctrine and FEA results. In such cases, the ability of RDBMS to rapidly select and reorganize information can be employed to entirely address decision complexity (i.e., the RDBMS can "make" the decision for the AMTP/Drill developer). Other decisions require consideration of data elements that do not pre-exist in tactical doctrine/FEA. Such data elements are carefully deliberated "judgments" made by AMTP/Drill developers. While reorganization of tactical doctrine/FEA might facilitate human judgments, reorganization does not remove the need for these judgments. Complex decisions of this variety need to be addressed by both "job aids" and information reorganization and retrieval capabilities of a RDBMS.

Three critical patterns become evident when reviewing the entire sequence of decisions to be made by AMTP/Drill developers. First, judgments are required throughout the AMTP/Drill development process. Second, certain judgments serve as input for more than one decision, while other judgments are employed only once. Third, virtually all of the judgments made after the selection of "slices of battle" to be addressed by STXs and Drills are repetitions of judgments made during the selection of STXs/Drills.

Optimum application of a RDBMS to AMTP/Drill document preparation involves minor adjustments in AMTP/Drill development procedures. "Job aids" and the ability of a RDBMS to rapidly select and reorganize information might be employed to facilitate judgmental decisions at the start of the AMTP/Drill development process (i.e., during the selection of STXs and Drills). The results of judgmental decisions relevant to subsequent steps in AMTP/Drill development would then be recorded in the database. RDBMS facilitation of subsequent decisions in the AMTP/Drill development process would be accomplished largely by the ability of a RDBMS to select/reorganize information within the database.

Applications of a RDBMS to AMTP/Drill preparation will be tested and refined by ARI and ATB through the development of a prototype CAPS within the U.S. Army Infantry School. The results of the present effort serve as input for the design of a CAPS database by identifying: judgmental decisions to be addressed by "job aids" embedded within the database; decisions made by collective training developers which, for reasons of efficiency, need to be recorded within the database; data elements (i.e., potential data files) serving as input for key decisions within the AMTP/Drill development process.

## References


1. Bloedorn, G., Crooks, W.H., Saal, H., Merrill, D., Meliza, L. and Kahn, O. Concept Study of the: Computer-Aided ARTEP Production System (CAPS). ARI Research Report. 1403, July 1985.

2. Hiller, J.H., Hardy, G.D., and Meliza, L.L. Guideline for Designing Drill Training Packages, ARI Research Product 84-15, January 1984.

3. TRADOC PAM 310-8, Collective Front-End Analysis for Development of the Army Training and Evaluation Program (ARTEP) and a Method for the Development of Drills, 25 September 1981.

4. TRADOC Reg 310-2 (Test-Revised) Development, Preparation and Management of Army Training and Evaluation Program Mission Training Plans (AMTPs) and Drills, 1 May 1985.

5. TRADOC Regulation 310-2, Development, Preparation and Management of Army Training and Evaluation Program (ARTEP) 1 December 1982.

# Application of Model Aircrew Training System (MATS) to B-52 Combat Crew Training

Conrad G. Bills, Major, USAF
Instructional Systems Development Division
93d Bombardment Wing
Castle AFB CA 95342-5000

Robert T. Nullmeyer, PhD
Operations Training Division
Air Force Human Resources Laboratory
Williams AFB AZ 85224

## INTRODUCTION

### Background

The driving force behind the development of the Model Aircrew Training System (MATS) has been the significant advancement in training simulation capabilities with the advent of the full scale weapon system trainer (WST). Major Miller (1978) concluded that the effective use of this technology would require deliberate integration of the WST, to include: (a) identification of training requirements for the B-52 WST, (b) concurrent training accomplishment report (TAPR) for events in the WST, (c) concurrent scheduling function for both WST and aircraft, and (d) Director of Training development of the WST syllabus to insure training requirements are specified. Since the application of WST technology applies across all phases of training, MATS is to be a total training system which integrates training across the full life cycle of the crewmember in the weapon system, incorporating all media, including the aircraft. Prior to MATS, contracted training system development has been primarily focused on ground training (KC-10, C-5).

The 1980 Board of Visitors report on B-52/KC-135 aircrew training management recommended an investigation of "more effective ways to use training devices and simulators to accomplish training objectives and use available flying time for critical tasks." They noted that Computer-Managed Instruction (CMI) methods would pay large dividends in optimizing the interrelated academic/simulation/flying training components of the total training program. The 1982 Board of Visitors concurred with prior recommendations and emphasized the need for automation of training management, evaluation and delivery.

Also in 1982, the USAF Scientific Advisory Board (SAB) concluded that the Military Airlift Command (MAC) training needed to become more efficient by applying currently available training technology. A major problem identified by the SAB was that the current level of combat training is inadequate, but that additional resources are not available to provide the necessary training within the Air Force. They recommended developing a model training program that would embody state-of-the-art training practices and technology to support reallocation of training resources in a manner that more effectively prepares MAC aircrews for their combat mission. They specified the C-130 weapon system due to the availability of advanced aircrew training devices at the present time. They recommended that this program be a model system for

other programs to emulate since the basic problems in C-130 training are common to other weapon systems. A subsequent MAC aircrew training task force concurred with this recommendation.

About the same time as the SAB study, the follow-on operational test and evaluation (FOT&E) was being conducted for the C-130 WST (Nullmeyer & Rockway, 1984). The study was accomplished in three phases corresponding to initial qualification (Phase I), mission qualification (Phase II) and continuation (Phase III) training. Training in the WST generally transferred positively to the aircraft as measured by proficiency ratings and by sorties to criterion. Substantial differences were shown between the performance of students trained in the WST (with visual cues) over students trained in operational flight trainers (OFT, without visual cues). The primary effect of the WST was putting the right person, the student, at the controls when they went to the airplane, rather than the instructor. The recommendations were: (a) establish clear goals for simulator training, (b) train to a set criterion in the simulator, (c) allow simulator training to accommodate individual differences (among instructors as well as students), and (d) integrate simulator training into the overall training system to take advantage of WST training potential.

In response to these problems, many of which exist throughout Air Force training, AFHRL initiated a contract for a model aircrew training system (MATS). This contract focused on the C-130 aircrew training program where lessons learned from the C-130 WST FOT&E could be used along with other state-of-the-art training technology. Attention was directed towards inefficiencies of the event driven program, heavily labor-intensive and aircraft dependent for all stages of training, resulting in some critical areas (e.g., combat tactics training) receiving inadequate attention because available training resources were allocated elsewhere. MATS could also have application to aircrew training for other weapon systems both within MAC and in other major air commands (MAJCOMS).

The MATS contract specified development of an integrated total training system for the full continuum of C-130 aircrew training. The system would provide for training program development, training delivery, training management, evaluation and training analysis, and training support. The system would include: (a) use of advance training concepts and state-of-the-art instructional technology, (b) proficiency-based instruction with advancement based on demonstrated performance, (c) effective use of existing resources and training media, especially the full exploitation of the combat training capabilities of the C-130 WST and most efficient use of available flight time, (d) improved capabilities for more systematic definition of task training requirements, aircrew performance standards, and training media necessary for proficiency-based training and aircrew quality control, (e) use of computer-supported resource scheduling, record keeping, and reporting, (f) provisions for systematically incorporating advances in training technology and program updates, (g) provisions for both internal and external feedback for evaluation and quality control, and (h) provisions to accommodate variable trainee experience levels, learning rates, and class composition. Emphasis was to be placed on computer-based instruction (CBI). Instructor training would need to reflect a new role for more direct student/instructor contact. This system is to serve as a prototype with generalization not only to other MAC programs, but also to Strategic Air Command (SAC) and Tactical Air Command (TAC) aircrew training.

## B-52 WST FOT&E

B-52 WST FOT&E was initiated in 1983 on the preproduction unit, but was terminated in June 1983 due to the unreliability of the device. With the acceptance of the production unit, FOT&E was resumed in June 1984. The critical issue was transfer of training from the WST to the aircraft. In addition to the transfer of training study, the plan also included comparisons among WST training options (Table 1), initially four options and subsequently followed by option five and option nine (Table 2).

Table 1

Sequence for B-52 WST Training Options

| Option 1 | Option 2 | Option 3 | Option 4 | Option 5 |
|---|---|---|---|---|
| WST-1 | WST-1 | WST-1 | WST-1 | WST-1 |
| WST-2 | WST-2 | WST-2 | WST-2 | WST-2 |
| F-1 | WST-3 | WST-3 | WST-3 | WST-3 |
| WST-3 | WST-4 | WST-4 | WST-4 | WST-4 |
| F-2 | WST-5 | F-1 | F-1 | (WST-5) |
| WST-4 | WST-6 | F-2 | F-2 | F-1 |
| F-3 | WST-7 | F-3 | WST-5 | WST-6 |
| WST-5 | F-1 | F-4 | F-3 | F-2 |
| F-4 | F-2 | F-5 | F-4 | F-3 |
| F-5 | F-3 | F-6 | F-5 | WST-7 |
| WST-6 | F-4 | F-7 | WST-6 | F-4 |
| F-6 | F-5 | F-8 | F-6 | F-5 |
| F-7 | F-6 | F-9 | F-7 | F-6 |
| F-8 | F-7 | F-10 | F-8 | F-7 |
| F-9 | F-8 | F-11 | F-9 | WST-8 |
| F-10 | F-9 | F-12 | F-10 | F-8 |
| F-11 | F-10 | F-13 | F-11 | F-9 |
| WST-7 | F-11 | | WST-7 | WST-9 |
| F-12 | F-12 | | F-12 | F-10 |
| F-13 | F-13 | | F-13 | F-11 |
| | | | | F-12 |
| | | | | F-13 |

Table 2

Sequence for B-52 WST Training Option Nine

| Crew | Pilot | Navigator | Electronic Warefare | Gunner |
|---|---|---|---|---|
| F-9 | Solo | | | |
| F-8 | SACR 60-4 Checkride | | | |
| F-7 | | | | |
| (W-11 | Optional) | | | |
| F-6 | | | | |
| (W-10 | Optional) | | | |
| F-5 | | | | |
| F-4 | First Integrated Crew Flight | | | |
| W-9 | Solo (combat sortie) | | | |
| W-8 | Progress Check | | | |
| W-7 | | | | |
| F-3 | Part Task Training | | | |
| W-6 | | | | |
| W-5 | | | | |
| W-4 | | | | |
| F-2 | | | | |
| W-3 | | | | |
| | F-1 | W-2 | W-2 | W-2 |
| | W-1 | W-1 | W-1 | W-1 |

By August 1984 initial results of B-52 WST FOT&E were available concerning training using the first four options. B-52 WST was showing good reliability, availability, and maintainability. Instructors liked it, particularly for the benefits in improved crew coordination. Transfer of training seemed to be positive. The need for computer assisted scheduling was evident as well as the need to add WST instructor options for individualizing instruction. Additional days were added for WST mission planning/critique. Development of training option five was initiated for a better building block approach in WST training.

In November 1984, the tryout of training option five began. The number of WST missions was increased from seven to ten to include a part task training mission for pilots only. The first two WST missions were abbreviated, one training air refueling only and the other low level only. Subsequent full missions continued development of crew coordination, provided instructor options to meet individual student needs, and then gave opportunities for enhancing crew coordination. Training option five was scheduled for four of the eight crew lines due to limitations of only one operational B-52 WST and insufficient console operator manning.

In January 1985, B-52 WST FOT&E results for the first four training options were reported. In October 1985, training option nine results were added. Training in the WST generally transferred positively

to the aircraft. Time to initial proficiency as measured by instructors on B-52 student progress sheets was significantly shorter for initial qualification students with WST training than for students without WST training (Tables 3, 4, 5 & 6). By the end of training there was little difference between the proficiency of the two groups as measured by standardization/evaluation personnel on SACR 60-4 checkrides. All instructors reported gains in WST training, particularly crew coordination. They indicated that student performance during the first aircraft sortie was comparable to that of students on sortie two or three of the non-WST syllabus. After seven or more front-loaded WST missions, student pilots were proficient enough to fly in the seat low level the first integrated aircraft sortie. Safety was excellent with no mishaps reported.

Table 3

Comparison of Number of Trials to Rating of 3.0/4.0 on B-52 Progress Sheet for Initial Student Crews with Weapon System Trainer (WST) Syllabus and without WST Syllabus

| Position | N | Rating | Q* | p** | JIC*** | p** |
|---|---|---|---|---|---|---|
| Copilot | 38 | 3.0 | .21 | NS | .31 | .05 |
| | | 4.0 | .30 | .05 | .66 | .005 |
| Navigator | 35 | 3.0 | .94 | NS | .25 | NS |
| | | 4.0 | .05 | NS | .39 | .025 |
| Electronic Warfare Officer | 29 | 3.0 | .62 | .025 | .09 | NS |
| | | 4.0 | .28 | NS | .64 | .005 |

*Coefficient of Colligation (consistency of effect)
**Fisher's Table of Correlation Coefficients
***Jenkins Index of Covariation (magnitude of effect)

Table 4

Comparison of Number of Sorties to Rating of 3.0/4.0 on B-52 Progress Sheet for Initial Students Crew with Weapon System Trainer (WST) Syllabus and without WST Syllabus

| Position | N | Rating | Q* | p** | JIC*** | p** |
|---|---|---|---|---|---|---|
| Copilot | 38 | 3.0 | .24 | NS | .40 | .01 |
| | | 4.0 | .43 | .005 | .40 | .01 |
| Navigator | 35 | 3.0 | .29 | .05 | .29 | .05 |
| | | 4.0 | .24 | NS | .61 | .005 |
| Electronic Warfare Officer | 29 | 3.0 | .59 | .005 | .40 | .025 |
| | | 4.0 | .59 | .005 | .69 | .005 |

Table 5

Comparison of Number of Trials to Rating of 3.0/4.0 on B-52 Progress Sheet for Training Option Nine Student Crews with Weapon System Trainer (WST) Syllabus and without WST Syllabus

| Position | N | Rating | Q* | p** | JIC*** | p** |
|---|---|---|---|---|---|---|
| Copilot | 22 | 3.0 | .56 | .005 | .40 | .05 |
| | | 4.0 | .[illegible] | .05 | .64 | .005 |
| Navigator | 21 | 3.0 | [illegible] | .01 | .38 | .05 |
| | | 4.0 | .91 | .005 | .44 | .025 |
| Electronic Warfare Officer | 19 | 3.0 | .49 | .025 | .11 | NS |
| | | 4.0 | .41 | .05 | .67 | .005 |

Table 6

Comparison of Number of Sorties to Rating of 3.0/4.0 on B-52 Progress Sheet for Training Option Nine Students Crew with Weapon System Trainer (WST) Syllabus and without WST Syllabus

| Position | N | Rating | Q* | p** | JIC*** | p** |
|---|---|---|---|---|---|---|
| Copilot | 22 | 3.0 | .83 | .005 | .43 | 0.5 |
| | | 4.0 | .98 | .005 | .57 | .005 |
| Navigator | 21 | 3.0 | .52 | .01 | .33 | NS |
| | | 4.0 | .99 | .005 | .69 | .005 |
| Electronic Warfare Officer | 19 | 3.0 | .15 | NS | .58 | .025 |
| | | 4.0 | .96 | .005 | .77 | .005 |

Generally for each B-52 WST training option, as the number of WST missions increased, the number of aircraft flights decreased without degradation of final proficiency. This was particularly true going to training option five and then again to training option nine. Graduation point was dropped from fifteen to thirteen aircraft sorties and then down to nine.

Training option nine. In February 1985, training option nine was designed to bring each crew member to a desired level of proficiency and pass a required progress check in the WST before going on to integrated flight

training in the aircraft. This training option reduced aircraft flying sorties by approaching proficiency based training. All aircrew training was rehearsed in the WST or another intermediate level training device prior to accomplishment in flight. For events in which the WST was inadequate, the aircraft supported "part task training." Standardization/evaluation personnel performed the progress check in the WST and flew with the crew on the first sortie following the progress check to assess the level of proficiency attained at that stage in training. Again, by the SACR 60-4 check there was no appreciable difference in proficiency between student crews trained with WST and those trained without WST. This was the first training option for which a list of training events was approved for creditation in the WST.

Though training option nine was a leap ahead toward effective use of the B-52 WST, the results of the B-52 WST FOT&E also confirmed the C-130 WST FOT&E results that lessons learned from the C-130 MATS study are needed to maximize the capabilities of WST technology. While monitoring acquisition of the C-130 aircrew training system (ATS), the B-52 program needs to develop improved methods for instructor training and for student preparation in order to take advantage of the enhanced capabilities of the WST. Also, the B-52 program needs continued effort in developing a computer based training system with CMI and CAI. Computer based training (CBT) applications will be particularly computer assisted scheduling and for procedures reinforcement, system simulation, and part task training. This will require a review of the entire training system from the MATS perspective.

## B-52 APPLICATION OF MATS

### Similarity of training.

C-130 aircrew training is sufficiently similar to B-52 aircrew training so that there can be direct spin-offs from C-130 MATS. Both training programs have the three phases for aircrew training, i.e., initial qualification, mission qualification, and continuation.

Initial findings of the C-130 MATS phase I (Fishburne, Williams, Chatt, & Spears, 1984) showed that though instructional systems development (ISD) methodology was purported, the program lacked systems design. Most of the training resources were for the formal school. Instructional strategy was transfer as a product of learning--learn then apply. Instruction was adapted to the "average" trainee. Simulation was not a key element in aircrew training. Flying instruction was primarily in-flight hours. Continuation training was deficient and piecemeal. In comparison with B-52 training, these elements were quite similar.

The C-130 MATS phase I report indicated training program needs which also apply to B-52 training: (a) unify the management of aircrew training, (b) individualize instruction, (c) take maximum advantage of computer-based training technology, and (d) apply modern concepts of learning.

### Increased productivity initiative.

Though the results are not all in from MATS, the cross feed between B-52 and C-130 aircrew training programs has given opportunity to apply lessons learned in developing the B-52 Combat Crew Training School (CCTS) increased

productivity initiative for FY86. B-52 aircrew training has shifted toward the more desired 1:1 student/instructor ratio and yet still increasing student output. A key component will be a firm progress check in the WST with pass/fail criterion and additional training required until desired proficiency is achieved before moving on to integrated aircraft flight activity. Essential elements of a computer managed instruction system will be instituted to keep a pulse on the effects resulting from implementation of the FY86 training plan.

SAC MATS working group.

Headquarters SAC and Human Resources Laboratory, Williams AFB, AZ, are combining forces to review MATS concepts as they apply to SAC aircrew training. They will identify lessons learned from other training programs as well as optimal application of computer based training (CBT). Their goal is to develop a generic SAC model with organizational structure, functions of components in the structure, the information/feedback interfaces of all program elements, the SAC/contractor interface, and the identification of changes needed to Air Force and SAC directives. MATS concepts will initially be applied to the C-130 ATS. As SAC ATS is developed, continued cross feed will facilitate ready application of lessons learned along the way.

## REFERENCES


B-52 weapon system trainer follow-on operational test and evaluation, final report. Castle AFB, CA: Instructional Systems Development Division, October 1985.

Board of Visitors report on B-52 and KC-135 aircrew training management. Castle AFB, CA: Instructional Systems Development Division, February 1980.

Board of Visitors report on B-52/KC-135 aircrew training. Castle AFB, CA: Instructional Systems Development Division, May 1982.

Fishburne, R. P., Jr., Williams, K. R., Chatt, J. A., & Spears, W. D. Design specification development for the C-130 model aircrew training system: phase I report. Seville Training Systems technical report, Irving, TX, January 1985.

Jenkins, W. O. & Hatcher, N. C. The design of behavioral experiments. Auburn University at Montgomery, AL (unpublished manuscript), 1976.

Miller, S. J. Integration of the B-52 weapon system trainer in the combat crew training school program. Air Command and Staff College research report, Air University, Maxwell AFB, AL, May 1978.

Nullmeyer, R. T., & Rockway, M. R. Effectiveness of the C-130 weapon system trainer for tactical aircrew training. Paper presented at the Interservice/Industry Training Equipment Conference, Washington, D.C., October 1984.

Williams, K. R., Marcantonio, A. W., & Fishburne, R. P., Jr. C-130 model aircrew training system: functional design specification. Seville Training Systems draft technical report, Irving, TX, July 1985.

# Sensitivities of Speeded Subtests

Toni G. Wegner
Malcolm James Ree
Air Force Human Resources Laboratory

Most aptitude tests are classified as power tests or speeded tests. Power and speeded tests differ in their emphasis on cognitive capabilities versus speed of processing. The purpose of this paper is to demonstrate that scores on speeded tests can be influenced by a variety of factors, and to discuss the implications of this in the use of speeded tests.

Power tests are designed to tap a designated cognitive capability and are structured to allow a majority of subjects to complete the test. In contrast, purely speeded tests are designed to measure speed without measuring deep cognitive capabilities; these tests consist of fairly easy items so that most of the people who attempt an item will answer it correctly, but few people will complete all items. There is a consensus within the testing literature that standardized testing requires strict control over the test administration procedures, especially such things as controlling the testing environment, time limits, and instructions (see, e.g., Anastasi, 1976; Clemans, 1971). Relatively little research, however, has addressed the deviations of administration procedures, especially as they relate to differential impact on power versus speeded tests.

The Armed Services Vocational Aptitude Battery (ASVAB) is an aptitude battery used for the selection and classification of all military enlisted personnel. Since 1980, the ASVAB has consisted of ten tests--eight power and two speeded. The power tests measure such things as verbal, mathematical, and technical information; the speeded tests, Numerical Operations (NO) and Coding Speed (CS), measure perceptual-psychomotor ability. In research and operational use of the ASVAB, speeded subtest scores have been shown to vary as a function of a variety of factors. These factors, then, produce error variability that is not found to the same extent in the power tests. Data are available to identify three factors that can produce fluctuations of speeded test scores: answer sheet format, test booklet format, and practice.

Answer sheet format effects first became evident after the ASVAB scores obtained from a carefully collected sample of American youth (McWilliams, 1980) were found to deviate from scores of military personnel only on the speeded tests. An initial study was conducted using Air Force basic recruits (Earles, Giuliano, Ree, & Valentine, 1983) to test the hypothesis that differences in the kinds of grids on the answer sheets used by the American youth sample versus military personnel could account for the differences in scores. A large-scale study involving applicants for all the armed services provided confirmation that the answer sheet differences account for the difference in performance between the two groups (Wegner & Ree, 1985).

Further evidence of answer sheet format effects came from another study with Air Force basic recruits. In this study, a third type of answer sheet, one used solely for ASVAB research purposes, was found to produce differences in speeded test scores (Wegner & Ree, 1984). In all three of the above

studies, subjects were administered identical test items using one of two answer sheets in an equivalent groups design. In no case were score differences found between the two groups on any of the power tests. The speeded test score results for the three studies are presented in Table 1. The answer sheets used differed in size, shape, and spacing of the response grids; this was enough to make a significant difference in scores on both the Numerical Operations and Coding Speed tests even at the raw score level.

Table 1. Answer Sheet Format Comparisons

| Study | Test | Operational Answer Sheet | Other Answer Sheet | Difference |
|---|---|---|---|---|
| Earles, et al. (1983)(N = 512) | NO | 41.22 | 37.72 | 3.50* |
| | CS | 52.33 | 51.01 | 1.32 |
| Wegner & Ree (1984)(N = 502) | NO | 39.71 | 34.67 | 5.04* |
| | CS | 50.53 | 47.70 | 2.83* |
| Wegner & Ree (1985)(N = 8598) | NO | 35.83 | 32.64 | 3.19* |
| | CS | 46.93 | 45.59 | 1.34* |

*Significant at the .05 level.

Note. Numbers are cell means. NO maximum is 50; CS maximum is 84.

Like answer sheet format effects, test booklet format effects occur as part of the test materials used. In the case of booklet format effects, differences between booklets were very subtle. Small differences in such things as type font, character pitch, and spacing produced significant variations in speeded test scores. Booklet effects were suspected when systematic differences were found in a single testing period between a form of the ASVAB and a scrambled version of the same form. The original and scrambled forms contained identical items in a slightly different order. The differences in scores were found consistently at monthly testing intervals. Speeded test scores for two months of testing, six months apart, are presented in Table 2 for the original form and the scrambled version of the same form. It is interesting to note that booklet differences did not impact both speeded tests uniformly. The same format that benefited Numerical Operations scores in one type of booklet proved to be detrimental to Coding Speed scores. No substantial systematic differences were found on the power tests.

Table 2. Test Booklet Format Effects

| Test | Time 1 | Original (n = 9119) | Scrambled (n = 7651) | Difference |
|---|---|---|---|---|
| NO | | 38.64 | 37.67 | .97 |
| CS | | 48.60 | 50.99 | -2.39 |
| | Time 2 | (n = 8267) | (n = 7211) | Difference |
| NO | | 38.07 | 37.07 | 1.00 |
| CS | | 47.39 | 50.13 | -2.74 |

Note. Numbers are cell means. NO maximum is 50; CS maximum is 84.

Practice effects, unlike format effects, depend on the test-taker rather than characteristics of the test. If a person takes a test more than once, it is implicit in the theory of testing that his or her score will not be identical each time. This is because a given score is considered to be an estimate of the person's true score, and is expected to vary within a range that specifies the accuracy with which the test can measure. Given this logic, it is reasonable that retests of the ASVAB sometimes show increments and sometimes show decrements in scores on power tests. The same does not hold for speeded tests. When all other conditions are held constant (e.g., motivation), scores on speeded tests are most likely to show uniform increments on subsequent tests. These effects can be attributed to practice, and reveal a condition under which scores on speeded tests can be enhanced relative to scores on power tests in a testing situation. Table 3 shows speeded test results for a single form of the test--one group had previously taken a similar form of the test (and thus had practice), the other group had not. These data were collected as part of the operational calibration of the most recently implemented versions of the ASVAB (Forms 11/12/13), with each score based on data from about 25,000 applicants.

Table 3. Practice Effects

| Test | Practice | No Practice | Difference |
|---|---|---|---|
| NO | 39.73 | 38.61 | 1.12 |
| CS | 52.25 | 48.49 | 3.76 |

Note. Numbers are cell means. NO maximum is 50; CS maximum is 84.

The three factors described above clearly do not exhaust the conditions under which speeded tests show sensitivity. Obvious problems can result, for example, when timing discrepancies occur. Consider the Numerical Operations test in which subjects are given three minutes to attempt 50 items. With subjects answering 35 items correctly on the average, it takes approximately five seconds to answer each item. Sloppy timing of even a few seconds can have significant effects.

Experience with the above issues has shown that the way to deal with the sensitivity of speeded tests depends on which of the factors is in question. If score comparisons are to be made at all on the basis of a test (which is a major reason to administer standardized tests), both answer sheet format and test booklet format need to be considered. It is preferable that tests be administered using identical booklets and answer sheets. If this is impossible (for example, due to obsolete answer sheets or the use of existing data), an adjustment should be developed to make the scores comparable (see Wegner & Ree, 1985, for one way this can be done). Practice effects, if they cannot be controlled, should be noted in the interpretation of test scores. In the use of tests likely to be affected by practice, one might consider incorporating practice into the instructions for the test so that outside practice effects are minimized.

In summary, speeded tests are sensitive to a variety of internal and external factors that do not similarly affect power tests. These effects on speeded test scores can be minimized by avoiding, controlling, or taking into consideration these factors.

## References


Anastasi, A. (1976). Psychological testing (4th ed.). New York: Macmillan.

Clemans, W. V. (1971). Test administration. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed) (pp. 188-201). Washington, DC: American Council on Education.

Earles, J. A., Giuliano, T., Ree, M. J., & Valentine, L. D., Jr. (1983). The 1980 youth population: An investigation of speeded subtests. Unpublished manuscript, Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

McWilliams, H. A. (1980). Profile of American youth: Field report. Chicago, IL: National Opinion Research Center.

Wegner, T. G., and Ree, M. J. (1984). Comparison of ASVAB scores on operational versus research answer sheets. Unpublished manuscript, Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Wegner, T. G., and Ree, M. J. (1985). Armed Services Vocational Aptitude Battery: Correcting the speeded subtests for the 1980 youth population (AFHRL-TR-85-14). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

# Initial Operational Test and Evaluation of Armed Service Vocational Aptitude Battery (ASVAB) Forms 11, 12, and 13: Data Quality Analysis

John R. Welsh
Toni G. Wegner
Air Force Human Resources Laboratory

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple test battery that is used by the Department of Defense (DOD) and the United States Armed Forces to select and classify applicants for military service. Operational versions of the test are replaced periodically. As new versions are implemented, tables must be generated to convert scores from the new tables to the metric of the reference population. Temporary tables are developed before the test is used operationally through an operational calibration. Once the versions are put in operational use (using the temporary tables), an Initial Operational Test and Evaluation (IOT&E) is conducted to finalize the tables. This paper describes data quality issues surrounding the IOT&E of ASVAB Forms 11, 12, and 13.

Since October 1980, the ASVAB has been composed of ten subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). Combinations of two or more subtests are used by the services to determine the specialty and occupational classification of applicants. The Armed Forces Qualification Test (AFQT) is a composite used by all the services for determining enlistment qualification of applicants; it consists of AR, WK, PC, and half-weighted NO.

Operational versions of the ASVAB are replaced by DOD and the joint services every three years to update obsolete items, take advantage of advances in the field of psychometrics, and minimize exposure of the battery. Replacement of the battery generates the requirement to equate new forms of the test to a reference test. The accurate establishment of the new tests on the score scale of the reference test is essential for meaningful year-to-year comparison of the distribution of abilities of applicants and recruits, as well as providing consistent meaning of scores used to make classification decisions.

In October 1984, the reference population was changed from a 1944 population based on men under arms to a representative sample of 1980 American youth (males and females, ages 18-23). This change was made to provide manpower and personnel planners with a more relevant normative base from which to make manpower decisions. The representative sample of 1980 American youth was tested on ASVAB Form 8a, which was implemented for operational testing in October 1980.

The most recent versions of the ASVAB are Forms 11/12/13. These batteries were designed to be parallel to ASVAB Form 8a. An operational calibration of ASVAB Forms 11/12/13 was conducted in May and June 1982 to develop tables to convert scores to the 1980 American youth reference

population (based on Form 8a scores). Forms 11/12/13 were scheduled to be implemented in October 1983. Subsequent issues surrounding the sensitivity of the ASVAB speeded subtests (NO and CS) to differences in answer sheet format caused a one-year delay in the planned implementation of the new forms (See Ree, Welsh, Wegner, & Earles, 1985). Corrections for answer sheet format effects were made (Wegner & Ree, 1985), and Forms 11/12/13 were implemented in October 1984 on the 1980 metric.

The IOT&E of ASVAB Forms 11/12/13 was conducted in Oct and Nov 84, immediately following implementation. Data were collected at all Military Entrance Processing Stations (MEPS) on each of the two versions of the three new forms (i.e., 11a, 11b, 12a, 12b, 13a, and 13b) and the reference test (Form 8a, labeled 13c for the IOT&E study) in an equivalent groups design. The purpose of the IOT&E was to establish the defining relationship of the six new ASVAB forms to the reference test under operational conditions. The data were to be used to check the operational calibration of the new forms, and to generate the final operational conversion tables.

All applicants who tested for the services between 1 Oct 84 and 30 Nov 84 had their answer sheets scanned at the Air Force Human Resources Laboratory; the results were provided to a contractor to be analyzed. During the data collection phase of the IOT&E, anomalies in mean AFQT raw scores of applicants taking the six new forms and the reference test were noted. This report will describe the data quality issues surrounding the IOT&E in the investigation of these anomalies. A more complete description of the development of the newly implemented ASVAB forms is contained in Prestwood, Vale, Massey, and Welsh (1985).

## The Problem

Anomalies were noted when the AFQT means for the six new forms and the reference test based on the operational calibration (OPCAL) were compared with the same means based on IOT&E data. In the OPCAL, all seven versions were administered to service recruits; 11a and 8a/13c were also administered to service applicants. In the IOT&E, all data were obtained from service applicants. Whereas the 13c AFQT mean is about the same as the means for the other versions in the OPCAL, it is clearly higher than the means of the other versions in the IOT&E. These data are presented in Table 1.

Table 1. Mean Raw AFQT Scores by Test Form

| Form | OPCAL | | IOT&E |
|---|---|---|---|
| | Applicants | Recruits | Applicants |
| 11a | 72.26 | 75.21 | 73.63 |
| 11b | | 75.31 | 73.42 |
| 12a | | 73.20 | 71.51 |
| 12b | | 74.90 | 73.49 |
| 13a | | 75.11 | 73.82 |
| 13b | | 75.20 | 73.49 |
| 8a/13c | 72.17 | 75.02 | 74.49 |

Because the IOT&E is the defining relationship between the new forms and the reference test, these data are valid as long as there is no reason to suspect the data are faulty. The anomalies between the AFQT means from the OPCAL and IOT&E raised suspicion and warranted investigation of the anomalies to verify the appropriateness of the use of the IOT&E data. The data quality analyses investigated four possible explanations.

First, it was possible that the groups taking the seven batteries were not equal in ability and that, by chance, the group that took Form 13c was slightly smarter. The second possibility was that the reference test was compromised to some extent, especially since Form 8a had been used operationally for two years. Evidence for nonequivalent groups or compromise in the IOT&E would invalidate these data as the defining relationship between the new forms and the reference test. The third possibility existed that there might have been a system-wide scoring error (key-error or conversion table error) in the automated MEPS scoring system. An error of this kind would be correctable. Finally, it was possible that format or printing differences between OPCAL and the IOT&E caused the observed mean score differences. The impact of these would depend on the nature of the differences.

Analyses

The analyses led to the elimination of the first three possibilities. The equivalence of the groups was examined by comparing the raw score performance of all seven groups on the General Technical composite. This composite contains the same subtests as the AFQT, with the exception of NO. Table 1 shows the groups were generally quite equal on ability. This also suggests that the raw score elevation of AFQT on Form 13c was due to NO.

Table 2. Mean Raw General Technical Composite Scores by Test Form

| Form | 11a | 11b | 12a | 12b | 13a | 13b | 13c |
|---|---|---|---|---|---|---|---|
| GT | 55.01 | 54.34 | 54.29 | 55.72 | 55.33 | 55.48 | 55.32 |

Note: Data are based on 18,000 plus applicants per form.

The second explanation was ruled out because there was independent evidence that compromise was at a low level at the time of the IOT&E (Sims, Truss, & Curia, 1985). A spot check of the MEPS scoring system revealed no error in the test scoring or in the use of appropriate conversion tables.

Given that the raw score AFQT differences could be traced to NO, format or printing differences were the most plausible explanation for the elevation of the 13c mean. This is consistent with other research (Wegner & Ree, 1985) showing speeded subtests to be sensitive to answer sheet format effects. Inspection of the IOT&E test booklets for the seven versions of the ASVAB revealed slight differences between the style used to print Form 8a/13c and that used to print the other six tests (which were identical). These differences can be characterized by the closer distance between letters and

numbers on the reference test. The type facing was also somewhat bolder. It should be noted that the format used for Form 8a/13c was identical to that used to collect data on the reference population.

The spacing on the NO subtest was visually more compact on Form 8a/13c, suggesting the format could account for Form 13c being slightly easier or faster. A study was conducted with Air Force basic recruits to test the hypothesis that format effects could account for the originally observed raw score anomalies. Identical NO items were printed in two formats: the 8a/13c format and the format used for the other six versions. These were randomly administered to two groups, each containing about 120 recruits. The mean NO raw scores were 41.4 for the group with the Form 8a/13c format and 40.1 for the group with the Forms 11/12/13 format. The results indicate that the Form 8a/13c format used in the IOT&E was slightly faster than the format of the other versions. Because of ceiling effects on NO with Air Force recruits, the magnitude of the difference is smaller than would be obtained with a broader range applicant sample; however, this difference still accounts for most of the mean AFQT difference between Form 8a/13c and the other six versions in the IOT&E (see Table 1).

## Conclusions

It was concluded from these analyses that the difference in mean AFQT raw scores between Form 8a/13c and Forms 11/12/13 was probably due to the slightly "faster" print format used for the reference test. For comparison purposes, the print formats for the seven versions used in the OPCAL were also inspected. These booklets used a format that was different than either of those used in the IOT&E, but all seven versions were printed with the same format. This explains why the pattern of AFQT means was different for the OPCAL and the IOT&E.

Because the format used for Form 8a/13c in the IOT&E is the same as that used for the 1980 American youth reference population, and there is no reason to suspect the IOT&E data set is faulty, the IOT&E data are appropriate to use for the development of final conversion tables. These data define the relationship between Forms 11/12/13 and the reference test.

## References


Prestwood, J.S., Vale, C.D., Massey, R.H., & Welsh, J.R. Armed Services Vocational Aptitude Battery: Development of forms 11, 12, and 13 (AFHRL-TR-85-16). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Ree, M.J., Welsh, J.R., Wegner, T.G. & Earles, J.A. (1985). The equating and implementation of the Armed Services Vocational Aptitude Battery (ASVABs 11, 12, and 13) in the 1980 youth population (AFHRL-TP-85-21). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sims, W. H., Truss, A. R., Curia, M. (1985). Extent of cheating on ASVAB (CNA 84-117707). Alexandria, VA: Center for Naval Analyses.

Wegner, T. G. & Ree, M. J. (1985). Armed Services Vocational Aptitude Battery: Correcting the speeded subtests for the 1980 youth population (AFHRL-TR-85-14). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

# DEVELOPMENT OF AN INTEGRATED PILOT SELECTION SYSTEM

Jeffrey E. Kantor
Air Force Human Resources Laboratory

## INTRODUCTION

The Air Force Human Resources Laboratory (AFHRL) is conducting a multi-year research and development (R&D) program to improve procedures for selecting candidates for USAF Undergraduate Pilot Training (UPT). The principal goal of selection procedures for UPT is to screen out those candidates with low chances of completing training and/or becoming successful operational pilots. This is more important now than ever because the costs of each UPT eliminee have risen dramatically (approximately $64,000) and the operational mission in which USAF aircrews are employed has increased in complexity and difficulty. Therefore, improvements in identifying candidates who have the requisite abilities for success will both reduce the wasted costs associated with eliminations and help ensure that operational pilots are capable of meeting the rigorous demands of today's military aerospace environment.

The first step in this R&D program was to review the current pilot candidate selection procedures. Since 1955, the principal components used in USAF pilot selection have included a physiological fitness examination; the paper and pencil Air Force Officer Qualifying Test (AFOQT); certain personal factors such as age and the type of college degree obtained by the candidate; and actual flying experience, assessed either by the possession of a civilian pilot's license or performance in a short Air Force light aircraft flying program given to promising candidates without a license. While these components do significantly predict success in UPT, literature reviews (Imhoff & Levine, 1981) and discussions with flight training and operational personnel indicated that additional candidate testing in the areas of psychomotor (hand-eye coordination) performance, cognitive abilities such as information processing speeds, and personality factors would have considerable potential for making improvements to the current system. To accomplish testing in these areas, the Basic Attributes Tests (BAT) were developed.

The BAT is a totally automated testing system designed around a commercially available supermicro-computer (Alcyon). The BAT system encompasses all the necessary hardware and software for high speed graphics, millisecond interval response timing, multiple joystick control, and keypad response entry. Software sub-routines in the BAT provide internal diagnostic checks, perform auto-recoveries in the event of systems failures, control test presentations, score test performance, conduct immediate data audits and checks for out-of-range responses and perform data file handling chores. The current version of the BAT has 15 subtests which are being evaluated for aircrew selection. These include measures of perceptual and mental encoding speed, mental rotational abilities, decision making speed, time sharing ability, figure-ground separational ability, risk taking, decisiveness, personality factors, and two tests of psychomotor ability. While the majority of these tests are in the early stages of evaluation, the psychomotor tests have been extensively analyzed and their usefulness in the effort to improve the selection of Air Force pilot candidates is the main topic of this paper.

CLEARED FOR PUBLIC RELEASE

## METHOD

### Predictive Validation

The two psychomotor tests were administered to 1725 subjects prior to their entry into UPT during fiscal years 79 through 82. These subjects had been selected by the normal system and their psychomotor scores were held in confidence until after they either completed or were eliminated from pilot training. Air Force UPT is a 49 week course with approximately 175 flying hours conducted in the T-37 and T-38 jet training aircraft. Following training termination, graduate or elimination status was obtained on these subjects and used as the criterion for the predictive validation of the psychomotor tests.

### Concurrent Validation

As an additional validation of the psychomotor tests, test units were sent to Williams AFB, Arizona to collect psychomotor scores on 95 graduating UPT students. For this group, the results of the Advanced Training Recommendation Board (ATRB) were obtained as criteria. The ATRB is convened towards the end of UPT and consists of instructor pilots, training squadron commanders and wing staff. The purpose of the ATRB is to decide which of the graduating students should receive a Fighter-Attack-Reconnaissance (FAR) follow-on assignment recommendation and which should receive a Tanker-Transport-Bomber assignment recommendation. Only the better students receive a FAR recommendation and the ATRB decisions were used as the criterion for the concurrent validation of the psychomotor tests.

### Psychomotor Test Description

The first of the two tests-TWO HAND COORDINATION-was a pursuit tracking task in which a triangular target moving in a circle had to be tracked with a cross-shaped cursor. The movement of the cursor was controlled by two joysticks. One joystick controlled the left-right axis (X1) movement of the cursor while the other joystick controlled the up-down axis (Y1) movement. The second test-COMPLEX COORDINATION-was a compensatory tracking task in which the subject was required to keep a cursor as close as possible to the intersection of a vertical and horizontal line while also keeping a short bar of light as close as possible to the vertical line. Test generation software included a biasing function which increased the difficulty of this task. The movement of the cursor in the left-right axis (X2) and up-down axis (Y2) was controlled by one joystick while the movement of the short bar of light in the left-right axis (Z2) was controlled by foot operated rudder pedals (later versions of this test, developed after 1983, have eliminated the rudder pedals and use both joysticks for the second test as well).

For both tests, scores are obtained by accumulating the absolute displacements from the cursor to the target point and, for the second test, from the bar of light to the vertical line. Each test has a three minute practice period followed by five minutes of scored performance. Five scores are produced, one of each of the control axes (X1, Y1, X2, Y2, Z2). Since these scores reflect tracking error, lower test scores indicate better performance.

RESULTS

Predictive Validation

For the psychomotor tests to be useful, at least some of the five scores must significantly differentiate between candidates who graduated and those who eliminated from UPT. In addition, it is reasonable to expect some differences among the scores for different types of eliminees. Reasons for elimination from UPT are flying deficiency, academic deficiency, medical, self-initiated, and fatality. If psychomotor ability, as measured by these tests, is related to flying aptitude, then the greatest differences in test scores should be found between the graduates and those who eliminated for flying training deficiency reasons.

These hypotheses were evaluated by comparing the means of the psychomotor scores among three categories--UPT graduates, all UPT eliminees, and UPT flying deficiency eliminees. The mean psychomotor scores for these groups and the probabilities of the differences among these means occurring by chance alone is presented in Table 1. All five scores showed significant ($p \leq .001$) differences in the expected direction between graduates and either category of elimination. In addition, for all three of the COMPLEX COORDINATION scores, the means of the flying deficiency eliminees were significantly ($p \leq .01$) worse than those of the eliminees for all other reasons. These results indicate that both psychomotor tests can identify eliminees but that the COMPLEX COORDINATION test may be a better test to identify those candidates who will eliminate for flying deficiency. These results validate the use of the psychomotor scores as predictors of success in USAF UPT.

TABLE 1

| | Psychomotor Scores | | | | |
|---|---|---|---|---|---|
| | $X_1$ | $Y_1$ | $X_2$ | $Y_2$ | $Z_1$ |
| Means by UPT Outcome* | | | | | |
| Graduate Means | 14315 | 16341 | 3559 | 2858 | 4725 |
| All Elim Means | 15829 | 17621 | 4936 | 4173 | 6678 |
| Flying Deficiency (FD) Elim Means | 16302 | 18007 | 5593 | 4702 | 7580 |
| Probabilities of Psychomotor Score Differences Occurring by Chance Alone | | | | | |
| Grads vs All Elims | .001 | .001 | .001 | .001 | .001 |
| Grads vs FD Elims | .001 | .001 | .001 | .001 | .001 |
| FD vs Other Elims | .103 | .182 | .006 | .010 | .005 |

*Psychomotor scores reflect errors and therefore, lower scores mean better test performance.

## Concurrent Validity

The psychomotor testing of UPT students nearing graduation at Williams AFB permitted the concurrent validation of the tests against the criterion of the ATRB decision. Only the better students received a FAR recommendation. It was found that the FAR students had significantly ($p \leq .01$) better scores than the TTB students on two of the five psychomotor measures (a third score was significant at $p \leq .05$). Also, using multiple linear regression, the five psychomotor scores produced a significant ($p \leq .01$) multiple correlation against the FAR/TTB criteria ($R = .43$). These results show that, in addition to identifying candidates with low probabilities of UPT graduation, psychomotor scores also relate to superior performance in UPT. Taken altogether, these results indicate that the quality of UPT students can be improved with psychomotor screening.

## Development of a Psychomotor Screening Equation

To obtain the maximum prediction accuracy from the psychomotor scores, a weighted equation or linear model was developed to predict UPT outcome. This equation provides a screening score based on psychomotor ability. Because the criterion was dichotomous and was coded 0 for eliminees and 1 for graduates, the screening score can be roughly interpreted as a probability of success in UPT. To determine the weights for the model, multiple linear regression was used with all five psychomotor scores as predictors and UPT outcome as the criterion variable.

While all five psychomotor scores were significant predictors of UPT outcome, the scores within each test were highly interrelated. Accordingly, the most useful linear model would contain the fewest psychomotor scores which still accounted for as much of the criterion variance as all five scores together. However, that does not mean that the tests could be changed to present only the most predictive axes because that would change the nature of the tasks in the tests. After several model comparisons using the F-ratio, the final screening equation ($R = .196$) contained only the $X_1$ and $Y_2$ scores. Using this equation, psychomotor screening scores were computed for the 1725 subjects in the predictive validation analysis and the mean screening score for those who graduated (79.0) was found to be significantly ($p \leq .001$) less from those who eliminated (75.1).

The practical usefulness of this psychomotor screening model was assessed using hit/miss tables produced for the 1725 case sample, comparing the predicted UPT success scores from the final regression equation versus actual UPT outcomes. To illustrate the effectiveness of the psychomotor screening model, two points have been chosen from the hit/miss tables. These two points represent the screening effectiveness of the psychomotor scores had they been used to rank order the 1725 cases and had only the best 90% and 80% of them been chosen. These examples are given in Table 2. For example, using the 10th percentile cut score (taking only the best 90% of the subjects) on this sample would have resulted in correctly rejecting 19.6% of all the eliminees while incorrectly rejecting only 8.0% of the graduates.

Development of an Integrated Pilot Selection System. While this psychomotor-based system appears useful, a potentially more effective approach would consider all valid screening information simultaneously. The next step in the research was to determine how to integrate the psychomotor information into the current selection process. Again, a multiple regression analysis framework was used. A subsample of 268 UPT cases was identified for whom data were available on age, AFOQT scores and performance in the Flight Screening Program (FSP). The FSP is a 14-hour, light-aircraft flying course used to screen applicants on basic flight skills. These are the major components used for selection of non-Air Force Academy pilots candidates. Through regression techniques, an integrated model was developed which captured the unique predictive information available from the experimental psychomotor tests and operational selection factors. The experimental psychomotor tests and every other source of information contributed unique prediction in a final regression model which had a multiple R = 0.45 ($p \leq .001$). Another hit/miss analysis was conducted using this integrated model. The effectiveness of the integrated approach, when used to select the best 90% and 80% of the 268-case sample, is also illustrated as part of Table 2. The results were quite impressive. Had the integrated system been used to select the best 90% (a 10 percentile cut score), then 28.9% of all the UPT eliminees in the sample would have been correctly rejected while only incorrectly rejecting 2.6% of the graduates. The integrated approach would have done substantially better than the psychomotor only system.

TABLE 2. Examples of Screening System Effectiveness

| | 10th Percentile Cut | 20th Percentile Cut |
|---|---|---|
| Percent of All Elims Rejected | | |
| Psychomotor Only Model | 19.6% | 31.8% |
| Integrated Model | 28.9% | 47.4% |
| Percent of All Grads Rejected | | |
| Psychomotor Only Model | 8.0% | 15.7% |
| Integrated Model | 2.6% | 10.4% |

DISCUSSION

The results of the analyses conducted demonstrate that the two BAT psychomotor tests produced scores which were valid predictors of UPT performance. Candidates who graduated from UPT were differentiated by their scores from those who eliminated. Also, superior UPT students (FAR recommended) were differentiated from weaker students. This differentiation of pilot candidates could be used through the implementation of the psychomotor only screening model, however, a much more effective use was found through the development of the integrated pilot selection system. The integrated system made use of the valid information available from the current operational selection system as well as the psychomotor test scores. Because of the potential improvements in selection accuracy inherent in the integrated

system, in January 1985, the Air Force Air Training Command decided to implement the integrated selection system following an operational trial which has already begun.

Ongoing R&D is continuing with the other BAT subtests. As these tests become validated, their use in the integrated system approach will be evaluated. If it is found that they significantly add to the predictive accuracy of the current integrated system, they will be included in a fashion similar to the psychomotor tests. Because all BAT subtests can be administered on the same computerized testing system, their inclusion in the integrated selection system will entail software changes only. In this regard, the Air Force UPT selection system can continue to mature and become more accurate over time. Additionally, new BAT subtests, measuring other relevant characteristics, can be developed, validated, and added to further improve the way in which candidates are selected for Air Force pilot training.

REFERENCE

Imhoff, D. L. & Levine, J. M. Perceptual-motor and cognitive task battery for pilot selection. AFHRL TR-80-27, AD-A094 317. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, Jan 1981.

# SELECTION OF SKILLED MAINTENANCE EMPLOYEES

# IN THE U.S. POSTAL SERVICE

Henry A. Mueller
Senior Psychologist
U.S. Postal Service
Washington, D.C.

## Abstract

The Maintenance Selection System is a comprehensive computer based evaluation system designed to improve the assessment of applicant abilities and knowledge in key qualification areas. The skilled maintenance craft positions filled by this system are in the areas of mail processing equipment maintenance, in addition to building and building equipment maintenance.

The Maintenance Selection System provides procedures, instruments, and techniques to improve the efficiency and quality of maintenance selection and promotion activities in order to ensure that the best qualified candidates are selected for placement into vacant positions. This system integrates valid and reliable independent measures into a comprehensive selection rating of an applicant's total qualifications. Moreover, system generated information is used by maintenance managers to diagnose deficiencies and to provide training suggestions for individual repairmen and technicians in these technical areas.

There are five major elements in the Maintenance Selection System: Qualification Standards, Written Examinations, Review Panel Evaluations, Supervisor Evaluations, and the Scoring and Reporting Program. These elements and the supporting structure and procedures are described.

## Introduction

The mechanization and automation of mail processing activities in the Postal Service is increasing at an astounding rate. We now have equipment which can sort more than 32,000 pieces of mail per hour. Building equipment in our facilities is also being upgraded. Complex electronic equipment is used to monitor and control essential environmental systems. The proper, cost effective preventive and corrective maintenance of this intricate equipment is crucial to the operations involved in moving the mail and, indeed, to the continued survival of the Postal Service.

The introduction of this complex electronic equipment has substantially changed the necessary qualifications of those who monitor, maintain and repair this equipment. A number of performance audits which have been completed by both Postal Operations and Inspection Service personnel have indicated that the competencies of the maintenance support staff are less than optimal to meet our expanding maintenance requirements, both in terms of current equipment maintenance activities for which they are given job training, and basic knowledge, skill and ability (KSA) requirements which they should bring with them to the job.

Whenever possible, skilled maintenance craft positions[1] are filled by promotion of qualified applicants from within the maintenance department. If our maintenance operations are to function effectively, we must have a system to determine what the minimum qualifications for a position are; we cannot simply select the best applicants from all those available in the overall labor market.

For most maintenance craft promotions where there are qualified applicants, the best qualified applicant is selected. Qualifications considered include the ability to perform the job, merit, experience, knowledge, and physical ability. However , if there is no appreciable difference in the qualifications of the best of the qualified applicants seniority is the determining factor. If we are to obtain maximum utility we must differentiate between applicant qualifications to the greatest extent possible. This requires that we obtain a comprehensive picture of an applicant's total qualifications for selection purposes.

Hiring unqualified people is costly in terms of both money and time. Productivity can drop. Morale problems can arise throughout the organization. An unacceptably high amount of training may be necessary to develop the basic skills that the new employee was erroneously assumed to bring to the new job. Maintenance managers tell us that in some cases no amount of training can enable the person to achieve even a minimally satisfactory level of performance.

---

[1] In the Postal Service a position refers to the duties and responsibilities performed by one or more individuals doing the same tasks.

The purpose of the maintenance selection development project was to develop the procedures, instruments, and techniques required to select and promote individuals who are qualified and capable of maintaining and learning to maintain our expanding inventory of technologically advanced equipment. In an effort to combine measures covering an applicant's total qualifications into a single selection rating, it was first necessary to identify valid and reliable selection methods which have been found to be predictive of future job performance for similar maintenance positions. A review of the literature looked at previous studies of written examinations as well as alternate selection procedures.

The usefulness of standardized tests for personnel selection is supported by Ghiselli (1966, 1973) in his reviews of published tests. In his 1973 article he states, "As would be expected from the characteristics of the jobs which constitute the trades and crafts, tests of intellectual abilities and of spatial and mechanical abilities have fairly substantial validity for the prediction of trainability." In the same article he goes on to say, "When it comes to measuring aptitude for performing the actual job itself, tests of intellectual abilities, spatial and mechanical abilities, perceptual accuracy, and personality traits are all found to be equally effective, having moderate validity." Ghiselli concludes his article stating, "It will be recalled that single tests are being considered here, and that judiciously selected combinations of tests would have been [sic] higher validity."

Schneider (1976) states "Achievement tests are also used in industry as predictors. Because the nature of an achievement test is to assess the outcome of a process of relatively formal education or formal experience, such tests are useful in predicting immediate job performance." Ghiselli (1966) has demonstrated that an interview rating based on discussion of specific aspects of an individual's previous work and educational history had reasonably high validity, even under very unfavorable circumstances. Schmidt, et. al. (1979) describe an alternative evaluation method which requires job applicants to use structured guidelines in assembling a portfolio of verifiable accomplishments relevant to the job. These accomplishments are then submitted to a panel of job experts who rate the candidate. The rationale for "unassembled examinations" is behavioral consistency: past performance is the best predictor of future performance in similar jobs.

These findings were used to develop a new maintenance selection system which integrates independent selection measures (i.e., written examinations and evaluations of training, education, experience, and job performance) into a comprehensive computer based evaluation system. The developmental process used in the study was designed to ensure that the selection process would be valid, i.e., that it would provide ratings that could appropriately be used to select qualified applicants who would be able to learn and to perform the important tasks in the skilled positions. To assure their usefulness, reliability, and defensibility, the new selection procedures were developed to conform with both professional and federal standards. (APA, 1980; Uniform Guidelines, 1978) This system also enhances uniform application of qualification requirements to ensure the integrity of selection and promotion decisions and to help us meet the Postal Service's merit and equal employment opportunity objectives.

System Development

## Job Analysis

A comprehensive occupational study of 54 Postal Service trades and crafts positions was undertaken to establish a data base for the development of a selection program. The initial phase of developing job analysis questionnaires began with a search of existing inventories and previous job analysis studies for similar jobs. Relevant task inventories yielded thousands of elemental tasks which were subsequently refined into 688 more general task statements and grouped into 24 functional job dimensions, called work behaviors. This effort was assisted by the contributions of eight groups of ten initial- and second-level supervisors from diverse postal installations in all five postal regions.

Using self-administered job analysis questionnaires information was collected from approximately 1300 incumbents and the original 80 supervisors. The survey included information about: (1) the tasks and work behaviors performed; (2) KSAs which are required for minimally satisfactory performance; (3) the tools and instruments used; and (4) equipment repaired and types of repair performed.

## Measures

### Qualification Standards

New qualification standards were developed to describe the minimum KSAs which are required on entry and which are necessary for the satisfactory performance of the duties and responsibilities of each position. They also contain additional KSAs which are not absolutely essential but which contribute to improved job performance by enabling the applicant to perform a wider variety of tasks and/or learn to perform a wider variety of tasks with less training. These are referred to as Desirable Qualification Factors. The Qualification Standards also include examination, physical, training, and licensing requirements.

### Written Examinations

Two multiple-choice paper and pencil instruments were developed to measure basic abilities such as mathematics, reading comprehension, and following oral directions. Three multiple-choice job knowledge tests were designed to assess knowledge of such subjects as mechanics, electricity, digital electronics, etc.

### Review Panel Evaluation

A Review Panel, established for each position, is composed of these members: (1) the senior installation Maintenance Manager or designee; (2) a supervisor or manager who is knowledgeable in the functional area of the position, and (3) an Employee and Labor Relations representative. The Review Panel evaluates written and oral documentation concerning the candidates education, training, and experience accomplishments pertaining to required and desirable KSAs. Performance Level Scales with behavorial examples, or benchmarks anchor four performance levels on the nine-point scale for each KSA.

Supervisor Evaluation

Nine-point Performance Level Scales with behavorial examples, or benchmarks, are used by the immediate supervisors to describe the applicant's level of qualification on 19 common KSAs.

## Major System Elements

Qualification Standards

Based on the job analysis information, new qualification standards were developed for each of the skilled maintenance positions. These qualification standards describe the minimum qualifications as well as additional KSAs which are not absolutely essential but which contribute to improved job performance.

Written Examinations

With the assistance of well qualified incumbents and supervisors with previous maintenance experience, five new written examinations were developed to measure the KSAs required for each of the skilled maintenance positions. An analysis indicated that a majority of the jobs required basic abilities such as math, reading comprehension, and mechanical comprehension. Specific job knowledge is also necessary for satisfactory performance in all of these positions.

Each of the examinations requires approximately three and one-half hours to administer. Once an examination is completed, the KSA scores obtained for all current Postal Service employees are retained in a computer file for use in selection to other positions which require the same examinations. Only the examination questions which cover the required or desirable KSAs for a position are used in the scoring algorithm for that position register.

Except under special circumstances, applicants may not retake an examination for a period of one year. When a Postal employee retakes an examination, the highest score ever achieved on each KSA, on that examination, is retained by the system. There is no penalty for retaking an examination even if a lower score is achieved.

Review Panel Evaluation

A Review Panel is established for each position to evaluate the qualifications of all current Postal Service employees for each position for which they apply. Written applications, supplemental applications and interviews are used to assess the applicant's past training, education and experiences and the effectiveness of the applicant in those situations. The purpose of the assessment is to determine whether the applicant actually possesses the required and desirable KSAs and to what degree.

After reviewing the accumulated evidence each panel member independently assigns a rating to each required or desirable KSA, on the qualification standard, that is evaluated by the Review Panel. The Panel members then

discuss their ratings and assign a composite rating which is the consensus of the Review Panel for each job element. Ratings are recorded on the optically scannable Review Panel/Supervisor Evaluation Form.

Supervisor Evaluation

Immediate supervisors have a major role in deciding who to promote or transfer within the maintenance craft. The supervisor of each maintenance employee who applies for a promotion or transfer evaluates and rates the subordinate's level of qualification on 19 common KSAs which are required for a majority of the positions. Evaluations are made on the basis of the applicant's job performance during the previous twelve months. Performance Level Scales are provided to the supervisors to standardize the evaluation process. If the supervisor cannot rate the subordinate on a particular KSA, due to insufficient information, the supervisor rates the KSA "Not Evaluated" for that individual. This rating does not penalize the applicant because both KSA and register scores are the weighted average of the available information. Ratings are recorded on an optically scannable Review Panel/Supervisor Evaluation Form.

The next-higher level supervisor or manager reviews the ratings and discusses any differences of opinion with the originating supervisor. When concurrence is reached and the forms have been verified they are forwarded to the National Test Administration Center for processing.

A maintenance craft applicant who requests to be re-evaluated by the Review Panel or to retake an examination is also required to have a new supervisor evaluation. This ensures that the most recent information about an employee's work performance is used for the promotion decision.

Scoring and Reporting Program

Application Procedures

Interested Postal Service employees are required to complete a Personal Data/Job Selection PS Form 2518-A (i.e., application) for all positions applied for. All applications for an installation are assembled and submitted to the National Test Administration Center (NTAC) for review. NTAC optically scans the applications and an automated warehouse system returns to the office a complete set of examination materials and a listing of all applicants, indicating those who have examination or evaluation data on file. In addition to examination materials, pre-printed scannable rating forms for the Review Panel and supervisor evaluations are sent to the office. Completed examinations and evaluations are sent back to NTAC for scoring.

When the PS Form 2518-As are received at NTAC a permanent record is created for each employee. Identifying information on this record allows the system to print each applicant's Social Security Number code and other background information on the pre-printed evaluation forms. Evaluation ratings on the forms are automatically matched with the employee's record when the forms are returned to the NTAC and electronically scanned. The employee's record identifies when complete information is received for each position register.

Applicant Results

After the scoring of all materials is completed, a Numeric Eligibility Register (NER) is provided for each position. The NER is essentially a sheet register, or listing of all applicants and their respective register scores. The register score is the arithmetic mean of the KSA composite scores required for the register (i.e., both proficiency and desirable KSAs). The KSA composite scores are the weighted average of the examination score and/or evaluation rating(s) for the KSA. In the KSA composite score the examination score is worth twice the weight of any single evaluation rating. Evaluation ratings on the same KSA are equally weighted.

Every proficiency requirement (i.e., required KSA) for a register has a specific qualifying score which must be achieved before the applicant is eligible for that particular register. All of the applicant's KSA composite scores must equal or exceed the respective qualifying scores in order for the applicant to meet the minimum qualifications and be considered eligible for selection or promotion.

Applicants who do not meet the minimum qualifications are reported as ineligible and are provided with a listing of the KSAs on which they are deficient. In addition, all Postal Service employees, regardless of eligibility status, are provided with a listing of the KSAs on which they scored below the 50th percentile for the incumbent population in the position.

Promotion Eligibility Registers (PERs)

When a maintenance position vacancy becomes available, it is filled first from a PER of current maintenance craft eligibles. PERs are created based on information from written examinations, supervisor evaluations and Review Panel evaluations. The NER score is used to rank order maintenance craft candidates for promotion on the PERs. Candidates who desire to improve their standing on a register may recompete when they have demonstrated the acquisition of new or additional training, education, or experience pertinent to the qualifications for the position.

Inservice Selection Registers

If the PER for a maintenance position is depleted, an office will normally select a Postal Service employee from outside of the maintenance craft. Inservice applicants outside of the maintenance craft are placed on selection registers in the order of their NER scores which are based on their written test scores and Review Panel evaluation ratings. A local Selection Committee (separate from the Review Panel), composed of an Employee Relations representative, and a technical maintenance supervisor in the functional area of the position, is responsible for ensuring that all qualifications, not measured by the written tests or Review Panel, are satisfied before selection or promotion. This is normally accomplished by interviewing previous supervisors, and other knowledgeable individuals to determine whether any negative information exists which would prevent the applicant from performing the requirements of the position satisfactorily.

### Entrance Hiring Registers

When no current Postal Service employees are eligible for promotion or reassignment, positions are filled by entrance applicants. Applicants from outside of the Postal Service (i.e., entrance applicants) are ranked on the basis of their written test scores, plus veterans points. Using the rule of three, qualified applicants are considered on an "as-needed" basis in relation to their position on the hiring register. At the time of hiring consideration, applicants must meet the total qualifications for the position. This is determined by a Selection Committee which reviews and evaluates a written and supplemental application for the individual as well as any additional information which they can obtain from applicant interviews and information from previous employers.

### Promotion Counseling

The system also provides information which can be useful for promotion counseling. A general comparison can be made between an applicant's test performance and that of the incumbent population. Based on this information, informal meetings with supervisors or training personnel to discuss applicant deficiencies are encouraged. There are many mutual benefits to be obtained by informing applicants on how they can improve and by encouraging them to obtain appropriate training and experience to increase their promotion potential.

## Summary

There have been problems selecting qualified maintenance personnel who are capable of maintaining and learning to maintain our expanding inventory of technologically advanced mail processing and building-related equipment. In a coordinated and cooperative effort the Office of Human Resources and Office of Maintenance Management in the Postal Service have developed the Maintenance Selection System in order to ensure that we can select and promote qualified maintenance personnel.

An effective and fair selection system requires accurate identification of current selection requirements, reliable, objective, and valid measurement instruments, and procedures and a timely method of distributing qualification ratings. It is believed that the Maintenance Selection System, using an integrated systems approach and constant surveillance, meets this need for the Postal Service. Current research efforts will monitor system performance and provide a comprehensive understanding of the interrelationships between the measurement methods within the system. Future refinements may enable us to revise and streamline the system and, at the same time, increase the validity and utility in the promotion process.

In conclusion, Ghiselli, Schneider, Schmidt and others have all proposed independent methods for measuring applicant qualifications. The Maintenance Selection System integrates written examinations, Review Panel Evaluations and supervisor evaluations into a composite rating which presents a comprehensive evaluation of an applicants total qualifications for promotion to skilled maintenance craft positions.

## REFERENCES


Division of Industrial-Organizational Psychology, American Psychological Association. Principles for the validation and use of personnel selection procedures. Washington, D.C.: Author, 1980.

Ghiselli, E. E. The validity of occupational aptitude tests. New York: John Wiley & Sons, 1966.

Ghiselli, E. E. The validity of aptitude tests in personnel selection. Personnel Psychology, 1973, 26, 461-477.

Schmidt, F. L., Caplan, J. R., Bemis, S. E., Decuir, R., Dunn, L., & Antone, L. The behavioral consistency method of unassembled examining. Washington: U.S. Office of Personnel Management, 1979.

Schneider, B., Staffing Organizations (Santa Monica, California: Goodyear, 1976).

Uniform Guidelines on Employee Selection Procedures (1978). Federal Register, 1978, 43, No. 166, 38290-38309.

# THE EFFECTS OF THE FLIGHT SCREENING PROGRAM ON ATTRITION IN UNDERGRADUATE PILOT TRAINING

By

JOHN C. QUEBE
Air Force Human Resources Laboratory

## INTRODUCTION

Air Force commissioned officer and Officer Training School (OTS) pilot training candidates who do not have a Private Pilot's License are required to complete a Flight Screening Program (FSP). FSP is a 14-hour flying program in the T-41 (Cessna 172). All student sorties are graded by the Instructor Pilot (IP). Students performing especially poorly may be eliminated before completion of the program. After 12 flying hours, students are administered Final Evaluation Flight covering the basic flying skills taught. Students failing this evaluation may repeat it one time. Students achieving a satisfactory Final Evaluation Flight grade proceed to Undergraduate Pilot Training (UPT). All other students are eliminated from the pilot training program for flying training deficiency (FTD). Thus, the program acts as a screen for entry to UPT.

In 1980, an experimental evaluation of the FSP was begun with the aim of determining the effects of the program on UPT and particularly its effectiveness in reducing attrition in UPT below the level which would be expected without FSP. The research plan used to achieve this aim may be regarded as essentially addressing four questions:

1. Does the FSP have any effect on UPT attrition rates?
2. If FSP does affect attrition rates is the effect from screening, training or both?
3. Does the FSP confer a flight training and/or experience benefit?
4. If there is a training/experience effect, would a longer program of FSP flying significantly increase the training/experience benefit?

## APPROACH

### SUBJECTS

Different FSP treatments were given to different groups of pilot candidates who were then followed through UPT. Five groups were defined as follows:

Group I (No FSP). Consisted of 123 entrants who would normally have been required to complete the FSP but were allowed to enter UPT directly. These cases, therefore, were unscreened and untrained.

Group II (Extended FSP). Consisted of 57 entrants who were given an extended FSP of 20 hours instead of the normal 14 hours, although screening was still applied at the 14-hour point. These cases, therefore, had 6 extra hours of FSP training and experience.

Group III (Normal FSP). Consisted of 514 students who passed through the normal 14-hour FSP and were screened at the 14-hour point. These cases, therefore, had normal FSP and were screened. This group may be regarded as a control group.

Group IV (Unscreened). Consisted of 266 students who were given 14 hours at FSP but, regardless of performance, were sent on to UPT. In effect, these subjects were not screened.

Group V (FSP Failures). Was defined as a sub-group of Group IV and consisted of 34 of the 266 (13%) who were judged by the FSP IP's to be FSP failures for FTD reasons. Nevertheless, they were allowed to proceed to UPT. The members of this group, therefore, had received 14 hours flying experience but were considered to be unsuitable for UPT.

## PERFORMANCE CRITERIA

The FSP treatment effects were evaluated for impact on the following criteria:

1. Pass/fail for FTD reasons at the end of UPT.

2. Pass/fail for all reasons at the end of UPT.

## RESULTS

Comparisons were made between different groups to provide the answers to the four primary questions posed in the research plan. Results are based primarily on chi-square analyses. Final UPT training outcomes for each of the five groups are given in Table 1.

TABLE 1. UPT Outcome Data for FSP Groups

| | FSP Experimental Group | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | I | | II | | III | | IV | | V | |
| | N | % | N | % | N | % | N | % | N | % |
| Final UPT Outcome | | | | | | | | | | |
| FTD | 34 | 28 | 2 | 4 | 78 | 15 | 42 | 16 | 13 | 38 |
| Non-FTD Eliminees | 20 | 16 | 5 | 8 | 56 | 11 | 39 | 15 | 11 | 32 |
| All Eliminees | 54 | 44 | 7 | 12 | 134 | 26 | 81 | 30 | 24 | 71 |
| Graduates | 69 | 56 | 50 | 88 | 380 | 74 | 185 | 70 | 10 | 29 |
| Total N | 123 | | 57 | | 514 | | 266 | | 34 | |

## OVERALL EFFECTS OF FSP: DOES FSP ACHIEVE ANYTHING?

If the FSP achieves nothing in terms of reduced UPT attrition, it is not acting as a screening device, and is conferring no training/experience benefit.

To determine whether FSP has any affect on UPT attrition, the UPT outcomes of trainees who had entered UPT without previous FSP experience or screening (Group I) were contrasted with those of trainees who had taken the normal 14-hour FSP (Group III). The results of these comparisons are given by phase of training. Table 2 shows that final UPT attrition for all reasons was significantly lower in the group which had been through FSP (26%) than in the group which had not (44%) ( $p \leq .001$). Attrition for FTD reasons was also significantly lower (17% vs 33%; $p \leq .001$).

Conclusions. FSP has a significant effect on UPT attrition rates. Students who had been through FSP had lower attrition rates for both FTD and all reasons combined.

TABLE 2. OVERALL (SCREENING AND TRAINING) EFFECT (GROUPS I & III) UPT FINAL OUTCOME

| | OVERALL ATTRITION | | | FTD ATTRITION | | |
|---|---|---|---|---|---|---|
| | PASS | FAIL(ALL) | TOTAL | PASS | FAIL | TOTAL |
| GROUP III (FSP) | 380 | 134 | 514 | 380 | 78 | 458 |
| (%) | (74) | (26) | (100) | (83) | (17) | (100) |
| GROUP I | | | | | | |
| (NO FSP) | 69 | 54 | 123 | 69 | 34 | 103 |
| (%) | (56) | (44) | (100) | (67) | (33) | (100) |
| Total | 449 | 188 | 637 | 349 | 112 | 561 |
| | (70) | (30) | (100) | ( 80 | (20) | (100) |
| $X^2 = 14.33$, df = 1, $p \leq .001$ | | | | $X^2 = 12.4$, df = 1, $p \leq .001$ | | |

SCREENING EFFECT OF FSP: DOES FSP SCREEN EFFECTIVELY FOR UPT?

If the FSP screens out probable UPT eliminees, by eliminating them for FTD in FSP, it is fulfilling its primary purpose. Whether the FSP screens effectively for UPT may be approached in two different ways. The first question which may be asked is: "Can individuals who are likely to fail in UPT be identified at FSP?" The second question is: "Does the screening which takes place at FSP significantly reduce attrition rates in UPT?" The first question is concerned primarily with the validity of the FSP as a method of identifying potential UPT failures. The second question is more complex in that the answer depends on organizational factors such as the cut-off standards applied in FSP and the consequent rejection ratios.

Identification of Potential UPT Failures During FSP

To determine whether potential UPT failures could be identified after 14 hours flying at FSP, it was necessary to compare attrition rates in UPT between the 34 FSP 'failures' admitted to training (Group V) and the students who went through FSP alongside them, their contemporaries who were the FSP graduate element of the complete unscreened group (Group IV). (Contempories of Group V were examined to minimize the effects of any changes in FSP and UPT over time.) Through these comparisions, it was found that the overall

attrition rate of the Group V "FSP failure" (71%) was significantly ($P \leq .001$) higher than the overall attrition rate of the Groups IV "FSP Graduates" (25%). Similar results were found with respect to FTD elimination rates in UPT (see Table 3).

Conclusions. These results indicate that some high UPT failure risks can be identified at FSP with a good degree of accuracy, (only 29% of those identified as FSP 'failures' finally graduated from UPT).

The Effects of FSP Screening on UPT Attrition. Comparison of Screened and Unscreened FSP Groups.

To determine whether FSP screening overall had a significant effect on UPT attrition, the UPT outcomes of cases in Group III, who had been through the normal 14-hour FSP and had been screened, were compared with the UPT results of Group IV, who had been through FSP but had not been screened; the latter group contained the 34 'FSP failures' identified as Group V. These analyses revealed no significant differences in UPT results between the screened and unscreened groups. For overall attrition, the rates were 26% for the screened group and 30% for the unscreened group (Table 3). For the FTD attrition criterion, the rates were 17% for the screened group and 19% for the unscreened group (Table 3).

TABLE 3. SCREENING EFFECT (GROUP III vs IV)
UPT FINAL OUTCOME

| | OVERALL ATTRITION | | | FTD ATTRITION | | |
|---|---|---|---|---|---|---|
| | PASS | FAIL(ALL) | TOTAL | PASS | FAIL(ALL) | TOTAL |
| GROUP II' | | | | | | |
| GROUP III (SCREENED) | 380 | 134 | 514 | 380 | 78 | 458 |
| (%) | (74) | (26) | (100) | (83) | (17) | (100) |
| GROUP IV (NOT SCREENED) | | | | | | |
| (%) | 185 | 81 | 266 | 185 | 42 | 227 |
| | (70) | (30) | (100) | (81) | (19) | (100) |
| TOTAL | 565 | 215 | 780 | 565 | 120 | 685 |
| (%) | (72) | (28) | (100) | (82) | (18) | (100) |
| $X^2 = 1.47$, df = 1, P: NS | | | | $X^2 = .14$, df = 1, P: NS | | |

Conclusions. The implication of these findings is that the 14 hour FSP, with screening at the twelfth lesson did not achieve a significant screening effect for entry to UPT. However, before finalizing such a conclusion two aspects of the research should be noted.

First, the difference in UPT attrition between individuals identified in FSP as high UPT risks and those judged to be better risks has shown that the FSP examiners could discriminate with a resonable degree of accuracy.

Second, while these analyses were unable to show that FSP had a significant screening effect on UPT attrition, there are indications that performance in FSP is related to performance in UPT.

TRAINING/EXPERIENCE EFFECTS OF FSP. DOES THE FSP GIVE A TRAINING/EXPERIENCE BENEFIT IN UPT?

The analyses conducted to identify a screening effect of FSP were not conclusive. The next step was to examine whether FSP provides training/experience which lowers UPT attrition. Possible training and/or experience benefits of the normal 14-hours FSP were examined by comparisons between the UPT results of the group which was not required to attend FSP (Group I) and those of the group that went through the FSP but was not screened (Group IV). Higher UPT success rates for Group IV would be attributable to the training and experience received by this group in the FSP. Overall attrition was significantly ($p \leq .05$) higher in the no-FSP group than in the FSP-experienced group (44% vs 30%; Table 4). Also, attrition for FTD reasons was significantly ($p \leq .01$) higher in the no-FSP group (33% vs 19%; Table 4).

Conclusions. The FSP confers a significant training/experience benefit reflected in the UPT attrition rates. Attrition rates are lower among students who have passed through FSP (even though no screening was applied) than among those who have not been to FSP.

TABLE 4. TRAINING/EXPERIENCE EFFECT (GROUPS I vs IV)
UPT FINAL OUTCOME

| | OVERALL ATTRITION | | | FTD ATTRITION | | |
|---|---|---|---|---|---|---|
| | PASS | FAIL(ALL) | TOTAL | PASS | FAIL(ALL) | TOTAL |
| GROUP III | | | | | | |
| GROUP III (TRAINED) | 185 | 81 | 266 | 185 | 42 | 227 |
| (%) | (70) | (30) | (100) | (81) | (19) | (100) |
| GROUP IV (NOT TRAINED) | | | | | | |
| (%) | 69 | 54 | 123 | 69 | 34 | 103 |
| | (56) | (44) | (100) | (67) | (33) | (100) |
| TOTAL | 254 | 135 | 389 | 254 | 76 | 330 |
| (%) | (65) | (35) | (100) | (77) | (23) | (100) |
| | $X^2 = 6.14$, df = 1, $p \leq .05$ | | | $X^2 = 7.61$, df = 1, $p \leq .01$ | | |

TRAINING/EXPERIENCE EFFECTS OF FSP. WOULD A LONGER FSP GIVE GREATER BENEFIT IN UPT?

If an effect of FSP on UPT attrition is due to training/experience, would a longer FSP course provide greater benefit? This possibility was evaluated next. One group of students (Group II) was given 6 hours extra flying experience at the FSP. Those who had not reached a satisfactory standard by the Final Evaluation Flight were screened out at that point, but the remainder entered UPT with a total of 20 hours of FSP flying experience instead of the normal 14 hours. Differences in UPT attrition rates favorable to this group in comparison to the group which had received the normal 14-hour FSP with screening at the 12-hour point, would be attributable to the extra 6 hours flying experience in the FSP. At the end of UPT, overall attrition was

significantly lower in the 20 hour FSP group (12%) than in the normal-FSP group (26%); $p \leq .05$; Table 5). The difference in FTD attrition was also significant ($p \leq .05$; 4% and 17%, respectively; Table 5).

TABLE 5. EFFECT OF 6 HOURS EXTRA FSP TRAINING (GROUPS II vs III) UPT FINAL OUTCOME

| | OVERALL ATTRITION | | | FTD ATTRITION | | |
|---|---|---|---|---|---|---|
| | PASS | FAIL(ALL) | TOTAL | PASS | FAIL(ALL) | TOTAL |
| GROUP II (20 HR) | 50 | 7 | 57 | 50 | 2 | 52 |
| (%) | (88) | (12) | (100) | (96) | (4) | (100) |
| GROUP III (14 HR) | | | | | | |
| (%) | 380 | 134 | 514 | 380 | 78 | 458 |
| | (74) | (26) | (100) | (83) | (17) | (100) |
| TOTAL | 430 | 141 | 571 | 430 | 80 | 510 |
| (%) | (75) | (25) | (100) | (84) | (16) | (100) |
| $X^2$ = 4.53, df = 1, $p \leq .05$ | | | | $X^2$ = 5.18, df = 1, $p \leq .05$ | | |

SUMMARY OF RESULTS FOR THE EFFECTS OF EXPERIMENTAL FSP TREATMENTS ON UPT ATTRITION

(1) The analyses showed that the current 14-hour FSP had significant effects on attrition in UPT. Attrition rates were lower in the group which had undergone the FSP (Group III) than in the group which had not (Group I).

(2) High UPT-attrition risks could be identified in FSP. However, with the data available for analysis, no difference in attrition rates in UPT was apparent between the group which had been screened in the FSP (Group III) and the group which had not (Group IV).

(3) There was clear evidence that the FSP conferred a significant flight training and experience benefit. Attrition rates in UPT were lower among pilots who had taken the 14-hour program without being screened than among students who had been allowed to enter UPT directly.

(4) Extension of the FSP to 20 hours of flying gave an additional training benefit. Attrition rates in UPT were significantly lower in a group which had received the extended FSP than in the group which had undergone the 14-hour FSP.

# WHAT PERFORMANCE DOES A PERFORMANCE TEST TEST?

T. M. Ansbro
Staff, Chief of Naval Education and Training
Naval Air Station, Pensacola, Florida

Performance testing addresses performance of work by job incumbents and work planned for prospective incumbents, worker requirements for standing or certification in occupational fields, and/or skills profiles for personnel and manpower models. Our attempts to measure or otherwise evaluate these performances in productivity assessment, performance evaluation, or other process/product measurement subsystems don't always measure appropriately, completely, or expertly.

An American executive recently said, "Nobody in management likes, comfortable having to deal forthrightly with embarrassingly true, painfully accurate, detailed, and technically verifiable performance evaluation until the stakes are so high that there is no alternative." He then cited the highly successful landings of the Space Shuttles, ever one performed before the world public, and frequently a first-time effort by each pilot. Consider the "consequence of inadequate performance" (favorite job/task questionnaire item): unacceptable risk or loss of life, astronomically expensive equipment, and almost irreversible damage to US technology prestige. Performance requirement: near-perfection, passing score 100% (minus perhaps 2% margin for error), failure inconceivable, and job-performance environment "go, no-go".

The majority of the world of work enjoys a job-performance environment devoid of most of such harsh realities and dangers, more than a bit less demanding, and certainly more relaxed. When consequence of error is less than stringent, we may relax a bit too far, lose some perspective in establishing and applying performance standards to language of evaluation, or lose a skirmish in the battle of bias and ignorance versus informed objectivity. Consider these examples of executive judgement:

US Federal Civil Service --- "His performance is outstanding in every respect" --- rater awarded to 50% of his staff each year for several years.

US Navy printed personnel rating form (50th percentile) --- "Typical Outstanding 4.0 Chief Petty Officer". Possibly an inspiration for this comment from a Commanding Officer of an activity --- "I don't like to think of these occupational standards (OCCSTDS) as 'minimum' requirements of the rate. What we want in the Navy is performance ABOVE the minimum. I want everybody in this outfit to be 'Outstanding'. 'Satisfactory' is 'High Lousy'. Everybody knows that."

Last year's College Board's specification of needed learning outcomes in English Literature cited "ability to read critically, read literary text analytically, read with understanding a range of literature rich in quality and representative of different



Copy available to DTIC does not permit fully legible reproduction

literary forms and cultures" --- specifically, "interest in and a sense of inquiry about written words, ability to respond intuitively and imaginatively to literature."

A conference critic of Instructional Systems Development (ISD) cited "ISD Overkill" --- an ISD analysis of "skills" required to drive a car turned up 4,000 "discrete behaviors." The presenter opined that there are "only a handful of general competencies" needed to master driving a car "successfully;" therefore, "these are the ones to define and teach." Precisely how to get to that handful of competencies was not presented.

More examples of this kind might tempt a cynical observer to wonder how we ever evaluate with reasonable accuracy what we do or assess the product of our endeavors; a concerned one might concede the mess and ponder how to cull through, examine our world of work, and come up with an appropriate matrix of performance measurement. He might determine that we need to straighten out an evaluation vocabulary of such terms as (above) "generic learning outcomes, general competencies, quality, typical, outstanding" with such disparate quantities as several thousand versus "a handful," or percentile ratings not clearly supported by concrete component scaling. Such terms must beg for clarification if not uniformity.

What are the objectives of work evaluation in the first place? Design a job, describe it, hire/train a prospective incumbent, transfer/promote/retrain an incumbent, assemble an occupational field? Construct a ship/squadron/batallion manning document? Certify or license members of an occupational/career field? Confer credentials? Develop curricula/training programs? Whatever the objective, there should be a definitive test to establish that it has been or can be met. Methodologies and a variety of instruments abound, within fields and subdivisions of fields, accompanied by job-knowledge data, also, categories: current US armed forces job classfications (MOS, AFSC, Rating, NEC) number some 4200, with additions pending. We enjoy an embarrassment of riches, but not a uniform system to put all these riches together in a most meaningful way.

And that suggests that in terms of job performance evaluation there should be a test for each identifiable element of work that we can logically assemble within the categories with which we are dealing: occupational field, job, duty, task, skill, certification/paygrade requirement, and circle of standardization. These ingredients should support any testing program with objectives to determine or establish:

Performance requirements for a prospective jobholder to meet (job entry), satisfactory on-job task/skill performance;

Skill mastery levels, task/skill ranking criteria;

Job/skill certification/licensing criteria.

Needed are coded occupational-field and job-knowledge data to support the above; also, work-behavior inventories for projected or incoming technologies, systems, equipment (and task/skill performance deltas between them and contemporary counterparts).

Choosing the ingredients to provide a performance test for such objectives requires having these ingredients at our disposal: appropriate inventories, fleshed out, current, and with requisite analysis mechanisms to provide appropriate data outputs --- a job for the computer. Fortunately, many of the requisite mechanisms are either in development or in use. Logistics Support Analysis (LSA) is a massive and ambitious systems/equipment acquisition and occupational data base already in at least incipient function under a DOD mandate. HARDMAN, which should adjunct to LSA, is also under way in the Navy, and an adaptation ('MANPRINT') is in the works in the Army.

What work behaviors are germane to all these objectives? Of what use are specific skill elements as determinants of task (eventually, job) performance capability? Is there a clear line from skill demonstration to task performance; an "audit trail"? Can task performances be amalgamated to indicate job performance, job performance scientifically grouped to determine competency requirements for an occupational field? Information recall (of fact, rule, or principle), frequently labelled "remembering", is a conventionally required test-item mechanism. How often is it really the work behavior most appropriate for testing? Calling Morse Code out of memory is an enabling mechanism (skill) for sending/receiving/encoding/decoding messages (tasks), which also employ other computational and manipulative skills; it may be buried deeply among the suite of corporate skills that in the aggregate demonstrate worker capability to perform these tasks. Potential to predict performance capability would be relative to the skill's prominence among the aggregate; and only where skill prominence is appropriately reflected in a test of performance would its contribution to capability predictability be valid.

Figure 1 is an attempt at representing flow, and in a sense, metamorphosis of basic information ("job knowledge") through skills toward job task performance. Where the evidentiary links between components in the task-internal hierarchy are unbroken, there would be an audit trail establishing (or at least tracing) the connection, and the contribution, of static and processed information to skills performance and on up to job task performance (Ansbro, Hayes, 1981). Without such an audit trail, the mere existence of knowledge items in an inventory of task-supporting references is not sufficient to justify their selection for use in prospective test items to predict or reflect performance; nor is random selection among task-underlying skills. Appropriate reflection of the skill most closely associated with task performance is the keyword.

It may appear satisfactory to test worker speed and accuracy in sending and receiving, since these are the job tasks, and we can deduce that task-component skills have then been inferentially tested. But, for a proper profile of the task and skill performance suite of a classic worker in a model job in a recognized occupational field (for training program development,

position description, or manning document construction), we need performance information on more than obvious terminal tasks performed on a specific job. We may have to test for skills identified as transferable. In the aggregate, it is the skills that are transferable among and across tasks. Task requirements are specific, therefore fixed; skill requirements tend toward the generic, are therefore fluid elements in a performance matrix.

In a rough graphic representation of a single-job task/skill/knowledge inventory, a job may be depicted as in Figure 2. At this point, task subdivisions (Omnibus, Embodied, Unique) in Figure 1 acquire significance. Omnibus tasks contain all component work behaviors (skills, etc) of action-associated tasks of lesser complexity and smaller compass in the inventory; therefore "embody" them. "Unique" tasks have no such direct connections of commonality among at least the majority of their component work behaviors (Davis, Perry, 1980; Ansbro, 1984).. Illustrated is a single job encompassing its duties and tasks, and their interrelations. Omnibus tasks encompass embodied tasks, and in turn, support performance of duties; the whole constitutes a job.

Skills underlie and support many tasks, but assume varied importance or criticality to task performance; they extend through multiple layerings of task performance and transfer not only across tasks but also across jobs and occupational fields. Soldering and welding skills, for instance, provide extensive task support throughout metalworking, electricity/electronics, and fabrication occupational fields. Much the same characteristics attend skills associated with use of general purpose test equipment. Skill layerings (or mastery indexes) within these skills apply variously among task hierarchies; and to depict this increasingly complex environment of task-skill-occupational field interrelationship requires a matrix.

Expand the single-job illustration by replicating it in a single lateral line as far as established or assumptively associated job-task-skill boundaries extend, and we have a picture of an occupational field. Not every component of that field will be equally typical, but a circle of standardization can be established that will encompass what is most typical or representative. Relatively esoteric, extra-high-tech, or rarely performed jobs or tasks (sometimes including skill support) will fall outside the circle, but still within the field. It's a short step from this depiction to laying out a matrix of field/job/duty/task/skill relationships and skill-mastery layerings from which arrays of work-behavior groupings may be assembled to provide performance test action items and evaluative criteria. With adequate occupational data base support, the computer can either be called in to assist in the make-up of the testing, or it can be programmed to do the job itself.

With respect to testing in an occupational field, Navy

custom has been to examine for job-knowledge competence in a field rather than in a specific job within that field (or rating), therefore, generally within the circle of standardization. This exercise, however, becomes more difficult as the fields expand with the constant advance of technology, which, as it becomes more platform/system/equipment-specific, may be narrowing the circle to the extent that it represents less of the field than can still be considered authoritatively representative. Such almost-amoebic reproduction of jobs and tasks can periodically (or at least eventually) require redefinition of the field. But it is the specifics of job and task descriptions that tend to broaden and fractionate the occupational fields, while the common component (and therefore transferable) skills tend to hold the fields in still-recognizable form. However, increasingly expanding application of the computer to large data fields should tell us that it is no longer necessary to hold tightly to established structures when they are expanding or changing to differing configurations.

Establishing job performance requirements for testing or the objectives discussed herein necessitates a massive, up-to-date occupational data base of sufficient descriptive detail and inventory depth to store, analyze, and rank data input, sort for job/duty/task/skill commonality across and among worker classifications, and establish task and skill complexity and competency indexes (Ansbro, 1984). Such a data base and analysis mechanism must function across and among job classifications and be internally blind to occupational specialty insignia. Task and skill ranking should depend upon complexity and competency; commonality and uniqueness should apply across and among classifications, for definition or redefinition of occupational field boundaries. Job/duty/task/skill/knowledge data should be cross-coded within, among, and across occupational fields and training programs, and should be retrievable elements in hardware and systems/equipment acquisition data systems, involving training/learning objectives in courses of instruction and the structure and components of performance testing.

REFERENCES

Ansbro, T. M., The Case For A Common Data Base. Proceedings, Second Annual Air Force Conference on Technology in Training and Education (AFTITE). Sheppard AFB, TX: 1984

Ansbro, T. M. and Hayes, W. A., The Job Task Analysis/Skills and Knowledge Marriage. Proceedings, 23rd Annual Conference of the Military Testing Association. Arlington, VA: 1981

Davis, D. D. and Perry, N. N., Determining Task Commonality in Navy Training. Proceedings, 22nd Annual Conference of the Military Testing Association. Toronto, Ontario, Canada: 1980

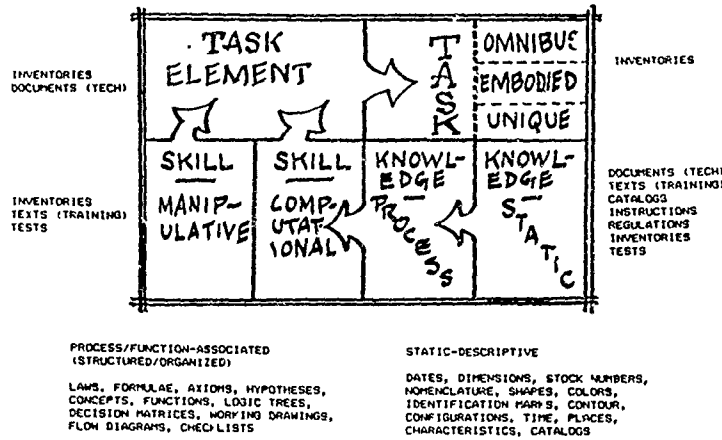



Figure 1.

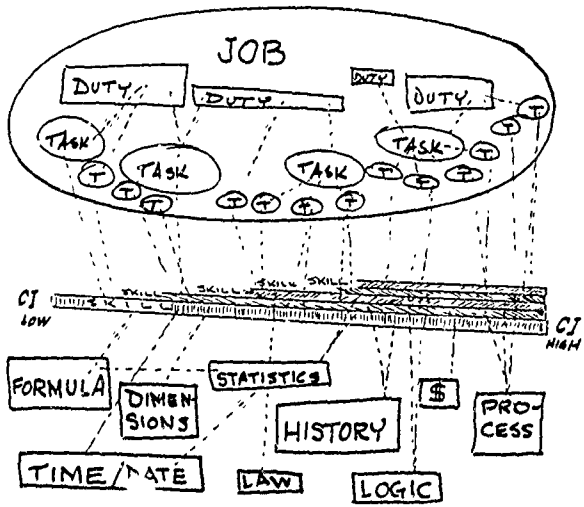



Figure 2.

How Qualitatively Informative are Test Items?:
A Dense Item Analysis

William M. Bart
Educational Psychology
University of Minnesota

The concept of the "dense item" was first introduced in a paper presented by W. Bart at the 26th Annual Military Testing Association Conference held in Munich in 1984. This concept is central in a framework directed to the goal of improved diagnostic and prescriptive testing. The dense item framework and its associated methodology of refined item digraph analysis has subsequently and unexpectedly provided a way to evaluate cognitive theories and instructional theories. This paper describes two quantitative indices that can be used to tell the degree to which a test item has the first two qualitative properties of a dense item.

## The Dense Item

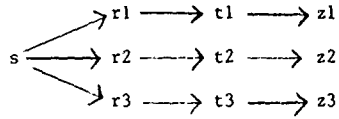
A dense item is any test item that indicates exactly why subjects provide the responses they give and exactly what instructional sequences should be provided the subjects to correct any faulty rules or procedures they may be using. Thus a dense item is an ideal diagnostic and prescriptive instrument and serves as a bridge between the field of learning and cognition and the field of instruction and teaching. A dense item has ten properties and each property has its own quantitative index. The first two properties are response interpretability and response discrimination. These two properties and their indices are featured in this paper.

## The Refined Item Digraph

Bart (1984) introduced a methodology termed "diagram analysis" and intended to investigate test items. This methodology has been refined and relabelled as "refined item digraph analysis". The refined item digraph is a graphic way to depict the associations between an item stem (e.g., "2 + 2 = __") and the responses to the item stem (e.g., "4") and the inferences among the responses, the rules that relate the item stem to the responses, and the instructional sequences that relate the rules, whether defective or not, to the correct rule. Psychologically, a rule is a sequence of one or more cognitive operations that permits an individual to generate a response from an item stem and an instructional sequence is a sequence of one or more instructional experiences that permits a student to learn the correct rule from an initial mastery of a defective rule.

A refined item digraph of a test item with three responses, if the item were dense, would use the following notation: (a) the item stem is termed "s" and constitutes a set S; (b) the responses are termed "r1, r2, and r3" and constitute the set R; (c) the rules are termed "t1, t2, and t3" and constitute the set T; and (d) the instructional sequences are termed "z1, z2, and z3" and constitute the set Z. In this case, let us assume that r2 is the correct answer, then t2 is the correct rule and z2 is the identity instructional sequence which, when employed, maintains the knowledge and usage of the correct rule. The refined item digraph of this item if it were a dense item would then be the following:

Figure 1: Refined item digraph of a dense item with three responses.




This refined item digraph indicates the inferences an instructor could make from a consideration of the responses of a student to the item. For example, if a student generated r3 as his response to item stem s, the teacher would know not merely that the response was wrong, but also that r3 results from usage of defective rule t3. The teacher could also infer that instructional sequence z3 should be provided to the student so that he/she can learn t2, the correct rule.

A refined item digraph has only between-set inferences and no within-set inferences interrelating the sets S, R, T, and Z for an item. An item digraph has, however, both between-set inferences and within-set inferences interrelating the sets S, R, T, and Z for an item and being indicated by arrows. A refined item digraph and an item digraph are both digraphs, because they both are arrays of points interconnected by arrows (Harary, Norman, & Cartwright, 1965).

In the explication of the dense item and its properties, it is sufficient to consider only between-set inferences for an item and that is why it is parsimonious to analyze only refined item digraphs and not item digraphs in general for this exposition on the dense item.

The Index of bod

One important concept in digraph theory that is useful in refined item digraph analysis is that of between-set outdegree. The between-set outdegree of point x, bod(x), is the number of arrows emanating from x to other points in other constituent sets in the item digraph. As an example, bod(r1) = 1 for Figure 1, because only one arrow emanates from r1 in set R to points in the other sets S, T, and Z. That one arrow indicates an inference from response r1 to rule t1. This index of bod will be crucial in defining the quantitative indices associated with the first two properties of the dense item.

Two Properties of the Dense Item

The first two properties of a dense item are response interpretability and response discrimination. In the subsequent sections, each property and its quantitative index will be described. In addition, an example will be provided regarding the computation of the index.

Response Interpretability

A test item has response interpretability if each response to the item is interpretible by at least one rule. An index of response interpretability, $I1$, indicates the degree to which an item has response interpretability.

In order to define $I1_i$, the index of response interpretability of item i, it is necessary to define $I1_{ij}$, the index of response interpretability for response j to item i. $I1_{ij}$ is defined in the following manner:

$I1_{ij} = 1$ if $bod(r_{ij}) > 0$ and (1)

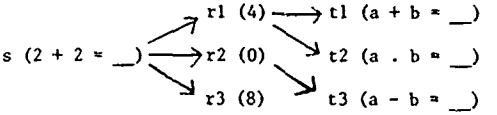
$= 0$ otherwise.

$I1_i$ is then defined in the following manner:

$$I1_i = \sum_{j=1}^{k} I1_{ij}/k \text{ with k being the number of responses to item i.} \quad (2)$$

To exemplify the computation of these indices, let us consider the following item involving the item stem "2 + 2 = __", the responses "4, 0, and 8", and the rules "a + b = __, a . b = __, and a - b = __." With this item the correct answer "4" is interpreted by the correct rule "a + b = __" and by the incorrect rule "a . b = __", the incorrect answer "0" is interpreted by the incorrect rule "a - b = __", and the incorrect answer "8" is uninterpreted. Figure 2 depicts the refined item digraph for this item.

Figure 2: Refined item digraph of an item with three responses.




The index of response interpretability for this item can be calculated in the following manner. In this item, k = 3, because there are three responses to this item.

$$I1_i = \sum_{j=1}^{k} I1_{ij}/k$$

$$= (I1_{i1} + I1_{i2} + I1_{i3})/3$$

$$= (1 + 1 + 0)/3$$

$$= 2/3$$

$$= .67.$$

The index of response interpetability for this item is .67 which indicates that 67% of the responses to the item are cognitively interpreted. This value is quite high and would very likely be superior to the indices of response interpretability of most other items. Most items have little response interpretability and thus have little diagnostic value.

Response Discrimination

A test item has response discrimination if each response to the item is interpretible by only one rule. The index of response discrimination, $I2_i$, indicates the degree to which an item has response discrimination.

In order to define $I2_i$, the index of response discrimination of item i, it is necessary to define $I2_{ij}$, the index of response discrimination for response j to item i. $I2_{ij}$ is defined in the following manner:

$$I2_{ij} = 1/\text{bod}(r_{ij}) \text{ if } \text{bod}(r_{ij}) > 0 \text{ and} \qquad (3)$$

$$= 0 \text{ otherwise.}$$

$I2_i$ is then defined in the following manner:

$$I2_i = \sum_{j=1}^{k} I2_{ij}/k \text{ with k being the number of responses to the item.} \quad (4)$$

To exemplify the computation of these indices, let us apply them to the item depicted in Figure 2.

$$I2_i = \sum_{j=1}^{k} I2_{ij}/k$$

$$= (I2_{i1} + I2_{i2} + I2_{i3})/3$$

$$= (1/2 + 1 + 0)/3$$

$$= 1/2$$

$$= .50.$$

The index of response discrimination for this item is .50 which indicates a modest level of response discrimination for this item. If a response is either uninterpreted or is interpreted by more than one rule, then that response does not discriminate well among cognitive rules. As the number of rules which interpret a response increases, so does the degree to which the response discriminates among cognitive rules decrease. That condition is reflected in the formulation of equation (3). The index of response discrimination is an item discrimination index which is concerned with the degree to which an item discriminates among cognitive rules and not ability groups as is the concern of item discrimination indices of conventional psychometric theory.

The Promise of Refined Item Digraph Analysis

Refined item digraph analysis provides a psychometric system which can be used to evaluate the extent to which an item has the dense item properties and thus has diagnostic and prescriptive value. This approach to item analysis involves ten dense item properties and has been developed by W. Bart. Refined item digraph analysis also provides a way to evaluate cognitive theories and instructional theories. Although only two dense item properties are discussed in this paper, it is hoped that these two properties and their quantitative indices can be used to evaluate the diagnostic values of existing items and to suggest areas of item research which could lead to improvements in the diagnostic values of test items.

References

Bart, W. (1984). The dense item: a bridge between learning and instruction. Proceedings of the 26th Annual Conference of the Military Testing Association. (pp. 391-396). Munich, Federal Republic of Germany.

Harary, F., Norman, R., & Cartwright, D. (1965). Structural models: an introduction to the theory of directed graphs. New York: Wiley.

Acknowledgment

The author wishes to express his appreciation to the following parties for their financial assistance and support in this research: (a) Prof. Dr. F. E. Weinert and the Max Planck Institute for Psychological Research; (b) the Fulbright Kommission; and (c) the Wilson/ U. of Minnesota College of Education Alliance.

# Test-Item Readability: How the Variables Work

R. Eric Duncan, Captain, USAF
University of Texas at Austin

## Introduction

In 1981, Duncan proposed a model of test-item readability which incorporated item, examinee, and environmental characteristics. These characteristics interact to produce the readability level of an item. Duncan (1981), however, lacked a solid theoretical framework from which to relate the semantic and syntactic components of an item to its readability. To correct that deficiency Duncan adapted Kintsch's (1974) propositional approach to text-based readability to multiple choice test items. This paper will briefly describe this adapted approach, describe the most critical syntactic and semantic variables which predict item readability, and present initial results in the attempt to predict item readability.

## Semantic Variables and Kintsch's Propositional Theory

The variables described here can be grouped into four distinct areas: semantic variables, syntactic variables, a cognitive load variable, and measures of prior knowledge. The semantic variables include propositional density, operator density, argument density, and propositional [illegible]. The variables in this area are directly related to Kintsch's (1974) propositional approach to the description of memory in semantic memory and so a brief explanation of that theory is in order.

Kintsch's (1974) semantic approach to reading and processing textual material focuses on the proposition. Kintsch and Keenan (1973) point out that sentences read from text are not stored verbatim, but rather as propositions. Propositions are word concepts combined to form a logical set of lexical items and contain a relation (usually a verb) and n arguments (nouns, adjectives, pronouns). These propositions are put together in a logical manner, establishing a representation of the text, known as a text base, in memory. A text base is simply "an ordered list of propositions" (Kintsch, 1974, p. 13). To obtain this list of propositions, the text must be semantically analyzed. Propositions, which include relation and argument(s), are then abstracted beginning with the first sentence.

Before describing the propositional construction process, relations and arguments need to be described. A relation is a word concept (not necessarily a word) which describes some action or state of being and normally appears as a verb, adjective, adverb, or noun.

| | | |
|---|---|---|
| John sleeps | (SLEEP, JOHN) | (1) |
| Mary bakes cake | (BAKE, MARY, CAKE) | (2) |

In example (1) from Kintsch (1974), SLEEP, a verb, is the relation describing some action that is being performed by the argument, JOHN. Fillmore (1971) established semantic rules for arguments. Arguments must be an agent, experiencer, instrument, object, source or goal and are structured in that order of importance. In example (2), MARY is the agent that BAKE(s) the object CAKE. Propositions can serve as arguments for other propositions, as well. In example (3),

| | | |
|---|---|---|
| If Mary trusts John, | (TRUST, MARY, JOHN) = a | (3) |
| She is a fool | (FOOL, MARY) = b | |
| | (CONDITION: If, a,b) = c | |

the proposition "c" has propositions embedded in it as arguments. This function is important when building text bases, since it is more economical and requires less memory space than recreating new propositions that had been processed earlier.



Copy available to DTIC does not permit fully legible reproduction

## Levels and Operators

Two important features in the d[illegible] of a text base are presented [illegible] (3). The first feature is that of the level of propositions. "A proposition is said to be subordinate to another if it contains an argument that also appears in the first proposition" (Kintsch, *et al.*, 1975). Subordination can occur, as is most often the case, immediately after the superordinate proposition in the text base, or can occur much later in the text base, as in the case of propositions being used as arguments for other propositions. In example (3), subordination is indicated by the indentation of propositions "b" and "c". Indentation of propositions is a convention established by Kintsch and was used in creating item protocols.

The second important feature shown in example (3) is that of operators. Operators are mechanisms which require inference on the part of the reader. Simply, operators do not state an explicit relationship among propositions, but rather challenge the reader to obtain an implicit meaning [illegible] information. Operators can also be described as taxing memory space and requiring memory searches [illegible] and manipulate previously encountered propositions. They include: Causality, Contradiction, Part, Time, Location, Condition, Conjunction, and Purpose. The operator would normally appear in the listing of the text base as presented in the "c" proposition. One additional operator (MATCH) has been created that was not included in Kintsch's operators. This operator identifies the cognitive operation of matching the propositions in the item alternatives to those propositions stored in memory.

## Empirical Evidence in Support of Kintsch's Theory

Now that the basic components of Kintsch's theory have been described, this section provides the empirical evidence which supports the contention that test items are broken down into logical semantic units (propositions) before storage in memory.

Kintsch and Monk (1972) demonstrated that experimental subjects stored text material in the same manner, regardless of the syntactic complexity of the text. They found that [illegible] syntactically complex paragraphs took longer to read but that there was no significant differences in the number of propositions recalled. Kintsch and Monk suggested that text is not represented syntactically in memory, but rather that it is represented semantically in propositional form. This evidence supports the contention [illegible] broken into propositions and is stored in memory.

Kintsch and Keenan (1973) examined the effect on reading time and recall of the number of propositions and the level in text of the propositions. The length of sentences (total number of words) and the number of propositions they contained were covaried. This approach is better known as propositional density, i.e., number of propositions/number of words. The levels of propositions were also varied. Kintsch and Keenan found that, if reading time was unlimited, propositional density significantly affected recall. They also demonstrated that superordinate propositions were recalled better than subordinate propositions. This finding was later supported by Kintsch, *et al.* (1975). This result can be more easily explained by referring back to example (3). Proposition "a" is a superordinate proposition while propositions "b" and "c" are subordinate to "a". Kintsch and Keenan, and Kintsch, *et al.* have shown that proposition "a" has a greater probability of recall than propositions "b" or "c".

In addition to supporting the results of Kintsch and Monk (1972) [illegible] and Keenan (1973), Kintsch, *et al.*, (1975) also examined



Copy available to DTIC does not permit fully legible reproduction

the effects that the number of different word concepts (arguments) would have on recall. Results indicated that the more frequently a word concept (argument) is repeated in the text, the better it is remembered. The authors also showed that, as the number of different arguments increases, reading time increases and recall decreases. Kintsch, *et al.* suggested that the history paragraphs may be easier than the science paragraphs they used because they contain propositions that "are already part of the subjects' general knowledge" (p. 209).

**Variables of Interest**

Experimental evidence has shown that propositional density, (# of propositions/# of words), operator density (# of operators/# of propositions), argument density (# different arguments/# of propositions), and propositional level contribute to reading comprehension. The hypothesized relationship between propositional density (PD) and reading comprehension scores in test items was that as PD increased (more propositions per word), the reading score necessary to understand the item would also increase. This is also true for operator density (OD), argument density (AD), and propositional level (PL).

Syntactic variables from Duncan (1981) were included in the estimate of item readability. Centerembeddedness, a modifying word or phrase located between the subject and predicate of an item, has been shown to make text less comprehensible than present (Lambert and Siegel, 1971). A modifying phrase that precedes the subject of an item is known as a left-branched phrase. A phrase which follows the predicate of an item is known as a right-branched phrase. Schwartz, *et al.*, (1970) showed that left-branched phrases reduced the comprehensibility of text while right-branched phrases had no appreciable effect on comprehensibility. These three variables, then, constitute the syntactic element of test-item readability.

There are two other elements that contribute to the prediction of item readability: (1) measures of prior knowledge, and (2) cognitive load. Kintsch and Vipond (1977) suggest that prior knowledge could enhance a person's ability to extract meaning from text. They indicate that the best method to assess prior knowledge is with a vocabulary test containing words used in the text material being read. This direct method, in the present study, is impossible to apply since all test and subject data had been taken from historical data files. Substitute measures include jargon and uncommon words. The number of uncommon words (UW) is determined by comparing all words in an item to the *Common Word List* compiled by Kincaid, *et al.*, (1980). This list contains the 20,000 most common words used by enlisted Navy personnel with a 9th grade reading ability. After comparison, those words not appearing in the list are compared to the text reference material to determine if they are sufficiently explained. If there is no explanation or definition given, the word(s) are counted as uncommon. The second variable used to assess prior knowledge is jargon. Jargon is based on the judgment of the item writers. Jargon words, such as CBPO, grade, and MAJCOM are not common to the general public and are specific to Air Force personnel. The jargon variable is expected to enhance readability since the use of jargon is commonly used in the military to communicate frequently complex names or phrases in a succinct manner. The variable used to measure cognitive load is Bloom's Taxonomical Level (Bloom, *et al.*, 1956). This variable indicates the cognitive activity necessary to read and answer a test question and includes the following levels: (1) rote memory, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6)



Copy available to DTIC does not permit fully legible reproduction

evaluation. Raters will evalu[illegible] this variable and indicate t[illegible]

Th[illegible] of interest, then include four semantic variables adapted from Kintsch's propositional approach, t[illegible] syntactic variables, two measures of p[illegible] knowledge, and a variable to assess cognitive load.

**Experimental Approach**

One hundred multiple-choice items from a commonly used Air Force test were evaluated on the variables of interest by seven trained raters in a counter-balanced fashion. The data was reviewed by the author to insure accuracy. Two thousand examinees were randomly selected and their Armed Services Vocational Aptitude Battery scores were used to determine each examinee's reading grade level (PGL). This was accomplished using regression equations provided by Madden and Tupes (1966). The PGL of the 25th percentile of all examinees answering each item correctly was assigned as the PGL for that item.

The PGL and *p* value for each item were used as the dependent variables in further analyses. Regression analyses were performed to determine if statistically significant equations to predict test-item readability and item difficulty could be produced. Two regression approaches were attempted - all possible subsets regression and stepwise multiple linear regression. The independent variables were the item variables previously described.

**Results and Discussion**

Frequency analyses showed diversity among predictor variables with the syntactic variables providing little variability. This occurred because the items evaluated were relatively syntactically simple. Eighty-three percent of the items had no center-embedded phrases and seventy-eight percent had no left-branched phrases. PB did show a bit more variability with a range of 0 to 8 right-branched phrases, a mean of 1.92, and a standard deviation of 1.73. Seventy-nine percent of the items were evaluated at the rote-memory level of Bloom's Taxonomy and ninety-six percent of the items were at or below the application level. Forty-seven percent of the items contained at least one jargon word with one item containing six different jargon words. Since these items were written at a relatively simple level, only sixteen percent of them contained one uncommon word, three percent with two, and one with three uncommon words.

The semantic variables seemed to have more normal distributions with acceptable means and standard deviations. The [illegible] shown below provided the [illegible], standard deviations, variances, and numbers of cases. It should be remembered that, except for propositional level, these data are proportions, with a range of 0.00 to 1.00.

| Variable | X | S.D. | Variance | N |
|---|---|---|---|---|
| Propositional Density | .609 | .160 | .026 | 100 |
| Operator Density | .501 | .100 | .010 | 100 |
| Argument Density | .583 | .116 | .013 | 100 |
| Propositional Level | 5.430 | 1.183 | 1.399 | 100 |

These variables provide the most variability for use in [illegible] and regression of all the independent variables.

The 25 percentile RGL criterion variable had less range and variability than anticipated. This variable had a mean of 8.916, standard deviation of .123, and a range of 8.691 to 9.356. The item *p*-value criterion had a mean of .470 and a standard deviation of .2068, with the most difficult item being .08 and the easiest .93.



Copy available to D[illegible] A permit fully legible re[illegible]

Evaluation of the intercorrelation matrix is made with a 0.10 level of significance in mind. None of the correlations among the syntactic variables (CE, LB, and RB) were significant. This is most likely due to the lack of variability in CE and LP. The relationships of the syntactic variables and the measures of prior knowledge were also insignificant. However, the correlation of -.19 between jargon (JAR) and uncommon words (UW) was significant at p=.06. If the theoretical relationships between these variables and comprehension are true, the negative relationship indicates that jargon aids in item comprehension, while uncommon words hinder.

The correlations between syntactic and semantic variables tended to support anticipated relationships. The correlation between propositional density (PD) and RB (r=-.42,p<.001) indicates that as the number of propositions per word increases the frequency of right-branched phrases decreases. This finding is logical, in that, high PD values usually occurred in items with few words. The propositional level (PL) of an item often shows modification within it. RB also indicates modifying phrases. This [illegible] relationship should and does exist. Surprisingly, operator density (OD) has a positive correlation with LB (r=.23,p<.05). This result indicates that as the number of operators increase so does the number of left-branched phrases. The only explanation the author can provide is that OD, like LB may [illegible] item comprehension as its frequency increases.

The [illegible] correlations among the semantic variables [illegible] that argument density (AD) has stronger relationships with the other semantic variables than does PD. This may be a function of the lack of linearity in PD. The OD variable relates negatively to AD which reinforces the previous observation about OD (r=-.21,p<.001).

The regression analyses were disappointing in that the $R^2$ values accounted for little variance. In the stepwise multiple linear regression analysis with the 25th percentile PGL as the criterion, only three variables contributed to a meaningful change in the $R^2$ - RB, LB, and Bloom's Taxonomy. The regression equation with these variables produced a multiple R of .231 and an $R^2$ of .053 with an F of 1.75 (nonsignificant). When stepwise regression was used to predict item difficulty, different independent variables contributed to explained variance. These variables, jargon, PD, and OD, produced a multiple R of .232 and an $R^2$ of .054, with an F of 1.77. The all possible regressions analysis, using $R^2$ as the selection criterion, produced almost identical results.

While neither of the analyses with either criterion variable produced a significant $R^2$, they do point out some possible trends. One of the more interesting trends is that syntactic variables tend to predict PGL better than semantic variables and prior knowledge, while semantic variables and jargon tend to predict item p-values better than the other variables. The author believes that measured PGL (or estimated in the case of the criterion variable used in this research) may reveal how an item is put together and not what it is asking. For example, a person with a higher PGL may be better able to grapple with complex syntax and decide what type of information is being sought (rote, application, etc.) more easily than an individual with a lower PGL. This then allows the better reader to get to the meat of the question and start matching the propositional makeup of the item with what is stored in memory. The prediction of item p-values by the semantic variables and jargon indicates that item difficulty results from the closeness of the match between the propositions in the item and propositions stored by the [illegible] that was obtained through



Copy available to DTIC does not permit fully legible reproduction

the study reference material. Test items, then, should be syntactically simple and provide propositions as similar to those in the reference material as possible.

There is, however, more to test-item readability than just the item characteristics. It is possible that the examinee controls the majority of the variance in item readability and that the items themselves can be massaged only so much. More items with greater variability in their characteristics need to be analyzed to determine if this conclusion is valid.

Bibliography


Bloom, B.S., M.D. Englehart, G.J. Furst, W.H. Hill, and D.R. Krathwohl, 1956. Taxonomy of educational objectives (subtitle: The classification of educational goals): *Handbook I, The cognitive domain*. New York: David McKay Company, Inc. Portions reprinted in G.H. [illegible], W.D. Hopkins, and J.C. Stanley (eds.) 1972. *Perspectives in educational and psychological measurement*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., Selection 12.

Duncan, D. [illegible] *Theory and Model of Item Readability*. Published in the Minutes of the 23rd Annual Conference of the Military Testing Association, 1981.

Fillmore, C. J. Some problems in case grammar. Cited in Kintsch, W., *The Representation of Meaning in Memory*. Hillsdale, N.J.: L. E. Earlbaum Associates, 1974.

Kincaid, J.P., Aagard, James A., and O'Hara, John W. *Development and Test of a Computer Readability Editing System (CRES)*. Training Analysis and Evaluations Group Report No. 83, March, 1980. TAEG; Orlando, Florida, 321813.

Kintsch, W. *The representation of meaning in memory*. Hillsdale, N.J.: L. E. Earlbaum Associates, 1974.

Kintsch, W. and Keenan, J.M. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 1973, 5, 257-274.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., and Keenan, J. M. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 1975, 14, 196-214.

Kintsch, W. and Monk, D. Storage of complex information in memory: Some implications of the speed with which inferences can be made. *Journal of Experimental Psychology*, 1972, 94(1), 25-32.

Kintsch, W. and Vipond, D. Reading Comprehension and Readability in Educational Practice and Psychological Theory. In Lambert, J.V. and Siegel, A.I. Psycholinguistic determinants of readability. In a.I. Siegel and J.P. Burlett (Eds.) *Application of structure-of-intellect and psycholinguistic concepts to reading comprehensibility measurement*. AFHPL-TP 74-49, Lowry AFB, Colo., 1974.

Madden, H.L. and Tupes, E.C. *Estimating reading ability level from the AQE General Aptitude Index*. Lackland Air Force Base, Texas: Personnel Research Laboratory, Aerospace Medical Division, PPL-TR-66-1, AD-632 182, February 1966.

Schwartz, D., Spartman, J.P., and Deese, J. The process of understanding and judgments of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, 1970, 9, 87-93.



Copy available to DTIC does not permit fully legible reproduction

# A COMPARISON OF SERVICE JOB STANDARDS FOR FOUR MILITARY SPECIALTIES

Brian K. Waters

Human Resources Research Organization

The U.S. Military Services (Army, Navy, Marine Corps, and Air Force) enlist over 300,000 new active duty recruits each year. The Services use numerous criteria, primarily aptitude composite scores, to assign a recruit to one of the hundreds of Service specialties.[1] It would seem reasonable that aptitude entry requirements for similar jobs across Services would be somewhat similar. Armed Services Vocational Aptitude Battery (ASVAB) composite score entrance requirements for four different kinds of jobs across the four Military Services, (Radio Operator, Jet Engine Mechanic, Artilleryman, and Administrative Specialist) are analyzed in this paper. Entry standards for each specialty are shown as well as the proportion of the national youth population who meet minimum scores for each specialty/Service. Service recruiting goals for FY 1986 for these jobs, in terms of desired distributions of AFQT scores, are discussed along with their relationship with the minimum standards. Similarities and differences in aptitude level requirements across Services and specialties are examined for their implications for manpower policy.

## SERVICE DIFFERENCES

It is important that the reader understand at the outset that the Military Services have distinctly different, Congressionally-mandated functions which affect the kinds of skills, knowledges, and abilities that they need to fulfill their respective missions. The Army and Marine Corps, for example, have extensive ground combat responsibilities which are quite different from most Air Force and Navy specialties which tend to be more technically oriented. Certainly the environment of a ship is very different from an aircraft, tank, or infantry foxhole. Even for what is ostensibly the "same" job, such as aircraft mechanic, the particular weapon system may dictate a quite different mix of abilities for job incumbents. These job differences, and the historical, cultural, and organizational traditions of the Services dictate that each one set its own classification standards. The Services have had long-standing research programs to validate their standards, generally based upon the statistical relationship between aptitude test scores and performance in military training. More recently, the Service personnel research laboratories have been trying to validate their selection and classification procedures against actual on-the-job performance -- a far more difficult, though also more desirable basis for the establishment of minimum aptitude standards.

---

[1] A generic term "specialty" is used in this paper to denote occupations across Services, including Navy ratings.

[2] Paper presented at the 27th annual meeting of the Military Testing Association in San Diego, C , October 24, 1985. The opinions expressed in this paper are those of the author, and do not represent official DoD or Service policy.

## SELECTION OF SPECIALTIES FOR ANALYSIS

It was desired that a sampling of military specialties be used to provide a broader examination of Service classification processes beyond a single job. Several criteria were used to select the specialties for analysis from the hundreds of Service specialties. First, candidate specialties had to be part of the 32 occupational specialties being investigated in the Joint-Service sponsored job performance measurement (JPM) project. Second, specialties with cross-Service equivalent jobs were sought. Third, a representation of mechanical, administrative, electrical, and general specialties was desired. Fourth, at least one combat specialty would be included, and finally, specialties which had a relatively large number of annual accessions in each Service would be most useful.

The DoD Occupational Conversion Table, which categorizes Service specialties into common DoD occupational groupings in terms of similar tasks, provided the basis for cross-Service specialty matches. Four DoD Occupational Codes were selected. Table 1 shows the four primary specialties and their equivalent jobs in the other Services.

**TABLE 1 Occupational Specialties Selected**

| DoD Occupational Code (Type) | Specialty Data | Military Service | | | |
|---|---|---|---|---|---|
| | | Army | Navy | Marine Corps | Air Force |
| 041 (Combat) | Specialty | Cannon Crewman | Gunner's Mate | Field Artillery[b] Batteryman | [a] |
| | Service Code | 13B | GM | 0811 | -- |
| | # | 24,925 | 7,462 | 2,821 | -- |
| 201 (Electrical) | Specialty | Radio Teletype Operator | Radioman[b] | Field Radio Operator | Ground Radio Operator |
| | Service Code | 05C | RM | 253X | 2935X |
| | # | 8,691 | 16,181 | 5,109 | 1,749 |
| 510 (Administrative) | Specialty | Admin Specialist[b] | Yeoman | Admin Clerk | Admin Specialist |
| | Service Code | 71L | YN | 015X | 702X0 |
| | # | 24,418 | 11,956 | 4,243 | 27,528 |
| 601 (Mechanical) | Specialty | Aircraft Power-Plant Repairman | Aviation Machinist's Mate | Aircraft Power Plant Mechanic | Jet Engine Mechanic[b] |
| | Service Code | 68B | AD | 602X | 426X2 |
| | # | 840 | 13,340 | 1,122 | 10,620 |

[a]No Air Force equivalent specialty to DoD Occupational Code 041
[b]Primary JPM project Service Specialty

The four jobs selected include about nine percent of Army enlisted positions, ten percent of Navy billets, eight percent of Marine Corps jobs, and eight percent of Air Force enlisted positions. Although certainly not statistically representative of the hundreds of separate specialties in the four Services, they do suggest how a relatively large number of new recruits are matched with their jobs.

## STUDY LIMITATIONS

The target population for this study is limited to the U. S. Military Services' active duty enlisted personnel policies in March, 1985. Information provided during interviews with Service and DoD policy personnel may not reflect official positions or policies. Data shown reflect only the four specialties examined and do not generalize to other occupations.

## SPECIALTIES ANALYZED IN THIS STUDY

Table 1 depicts the four specialties chosen for analysis across the four Services. DoD has arrayed all Service specialties into categories which are similar to one-another based upon detailed task analyses of the jobs by occupational analysts. Clusters of the specialties analyzed in this study are identified by the DoD Occupational Codes shown on the far left of the table. Service-specific aptitude composites are made up of three or more ASVAB subtest scores as shown in Table 2. It should be noted that the Services use different score scales for their composites. Thus, comparisons of composite

**TABLE 2 Minimum Entry Standards for Four Analyzed Specialties, by Service**

| DoD Code | Military Service | Specialty Title | Specialty Code | Classification Composite | Minimum Score | Percent PAY[a] Qualified |
|---|---|---|---|---|---|---|
| 041 | Army | Cannon Crewman | 13B | AR+CS+MC+MK | 85 | 71 3 |
| | Navy | Gunner's Mate | GM | AR+EI+GS+MK | 204 | 54 6 |
| | Marines | Field Artillery Batteryman | 0811 | AR+CS+MC+MK | 90 | 67.4 |
| | Air Force | -- | -- | -- | -- | -- |
| 201 | Army | Radio Teletype Operator | 05C | AS+CS+NO+PC+WK | 100 | 51 6 |
| | Navy | Radioman | RM | AR+CS+NO+WK | 149 | 55 7 |
| | Marines | Field Radio Operator | 253X | AR+EI+GS+MK | 90 | 66 5 |
| | Air Force | Ground Radio Operator | 2935X | CS+NO+PC+WK | 50 | 42.8 |
| 510 | Army | Administrative Specialist | 71L | CS+NO+PC+WK | 95 | 56 0 |
| | Navy | Yeoman | YN | AR+CS+NO+WK | 165 | 30 4 |
| | Marines | Administrative Clerk | 015X | CS+NO+PC+WK | 100 | 43 3 |
| | Air Force | Administrative Specialist | 702X0 | CS+NO+PC+WK | 35 | 54 1 |
| 601 | Army | Aircraft Power Plant Repairman | 68B | AS+EI+MC+NO | 100 | 54 2 |
| | Navy | Aviation Machinist's Mate | AD | AR+GS+EI+MK | 190 | 64 4 |
| | Marines | Aircraft Powerplant Mechanic | 602X | AS+EI+MC+NO | 100 | 54.1 |
| | Air Force | Jet Engine Mechanic | 426X2 | AS+GS+MC | 30 | 58 9 |

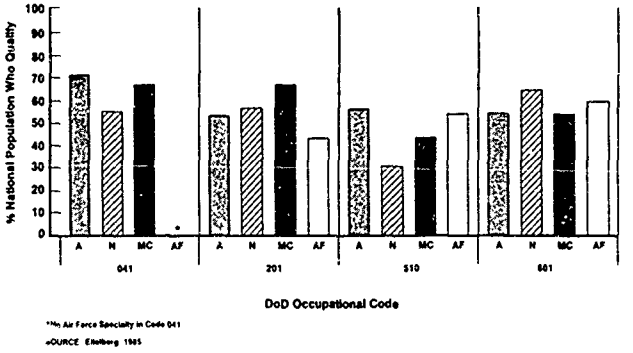
SOURCE Eitelberg, M J., with Lathrop M E., & Laurence, J H *Manpower for Military Occupations*, (1985) Naval Postgraduate School Monterey, CA

[a]18-23 year old males qualified for specialty (From Profile of American Youth Population)

NOTE ASVAB Subtest abbreviations are as follows

| | | | | | |
|---|---|---|---|---|---|
| AR | = | Arithmetic Reasoning | MC | = | Mechanical Comprehension |
| AS | = | Automotive-Shop Information | MK | = | Mathematics Knowledge |
| CS | = | Coding Speed | NO | = | Numerical Operations |
| EI | = | Electronics Information | PC | = | Paragraph Comprehension |
| GS | = | General Science | WK | = | Word Knowledge |

cutting scores should not be made across Services. Figure 1 displays the proportion of the 18-23 year old national population who meet minumum aptitude composite score standards for the 041, 201, 510, and 601 DoD Occupational Code specialties analyzed in this study.




*No Air Force Specialty in Code 041

SOURCE Ellenberg 1985

Figure 1. Percent of National Population of 18-23 Year Old Males Who Qualified for Service Entry and Specialty Entry by DoD Occupational Code and Service

Table 2 and Figure 1 contain a great deal of data about the effects of formal Service standards on qualification for these four jobs. Differences within similar jobs across Services and across the four specialities are striking. For example, in the administrative specialties (DoD Code 510), 56 percent and 54 percent of youth would qualify for the Army 71L and Air Force 702X0 specialties respectively, but only 30 percent for the ostensibly comparable Navy rating of Yeoman. Either the jobs are not really similar as the DoD Occupational Code suggests, or other policy variables are entering the standard setting process. This type of inter-Service difference is common, and will likely complicate the cross-Service use of JPM project results. Similarly, looking at these four specialities across the four Services yields inconsistent results. The most difficult of the jobs to qualify for in the Navy is Yeoman; while for the Army, it is Radio Teletype Operator; for the Marines, Administrative Clerk; and for the Air Force, Ground Radio Operator. Of the four specialties examined, lowest aptitude composite scores are required to qualify for Cannon Crewman, Aviation Machinist's Mate, Field Artillery Batteryman and Field Radio Operator, and Jet Engine Mechanic for the four Services, respectively. Certainly there are quite different Service policies reflected in these data.

Table 3 displays the Services' AFQT category recruiting goals for the four specialties from a recent report to Congress. (DoD, 1985) A remarkable diversity exists between Services in their desired aptitude distributions for new recruits in ostensibly the same jobs. A comparison of these Service requirements is also surprising when considered along with the previously discussed qualification standards for these jobs. For example, the Navy goal is 100 percent of its Yeomen recruits to be in AFQT Categories I to IIIA (the upper half of the population). The implications of these data on recruiting costs and supply/demand are obvious, though the rationale appears less so. An excellent discussion of these issues is available in DoD, 1985, Volume I. Other similar comparisons of Tables 2 and 3 data are equally provocative. In some cases, it appears that goals may be driven by a Service's recruiting market rather than by Service personnel requirements.

**TABLE 3 Military Service FY 1986 AFQT Recruiting Goals for Four Specialties**

| DoD Code | Military Service | Specialty Title | Specialty Code | FY 1986 Percent AFQT Category Goals I-IIIA | IIIB | IV |
|---|---|---|---|---|---|---|
| 041 | Army | Cannon Crewman | 13B | 58 | 30 | 12 |
| | Navy | Gunner's Mate | GM | 60 | 28 | 12 |
| | Marines | Field Artillery Batteryman | 0811 | 54 | 37 | 9 |
| | Air Force | -- | -- | -- | -- | -- |
| 201 | Army | Radio Teletype Operator | 05C | 64 | 27 | 9 |
| | Navy | Radioman | RM | 32 | 48 | 20 |
| | Marines | Field Radio Operator | 253X | 61 | 36 | 3 |
| | Air Force | Ground Radio Operator | 2935X | 90 | 10 | 0 |
| 510 | Army | Administrative Specialist | 71L | 94 | 29 | 7 |
| | Navy | Yeoman | YN | 100 | 0 | 0 |
| | Marines | Administrative Clerk | 015X | 71 | 28 | 1 |
| | Air Force | Administrative Specialist | 702X0 | 90 | 10 | 0 |
| 601 | Army | Aircraft Powerplant Repairman | 68B | 70 | 24 | 6 |
| | Navy | Aviation Machinist's Mate | AD | 73 | 20 | 8 |
| | Marines | Aircraft Powerplant Mechanic | 602X | 74 | 26 | 0 |
| | Air Force | Jet Engine Mechanic | 426X2 | 79 | 21 | 0 |

SOURCE DoD 1985

## CONCLUSION

Although it is clear that the individual Services must retain the prerogative of setting their own classification standards based upon their own personnel requirements, the analysis in this paper opens many questions in the mind of the military manpower analyst. Why are there such diverse entry standards for what are ostensibly the same jobs across Services? Why do the Services require radically different aptitude distributions for personnel fulfilling the specialties with the same DoD Occupational Codes?

The questions raised seem to point toward the methodologies used by the Services to define their personnel requirements. As shown by their submissions to the Congress (DoD, 1985, Vol. 1), their methods for determining their manpower requirements are not consistent. There appears to be a high priority need for research into establishing a methodology for determining the aptitude needs of the Services based upon a defensible way of specifying minimum requirements for each specialty. Such a methodological breakthrough would increase the efficiency of the assignment of personnel to specialties in all of the Services, and would provide a sound basis for requests for recruiting and training resources to the Congress.

## REFERENCES


Department of Defense
1985 Report to the House and Senate Committees on Armed Services: Defense Manpower Quality, Volumes I, II, & III. Washington, D.C.: Office of the Assistant Secretary of Defense (Manpower, Installations, and Logistics).

Eitelberg, M. J., with Lathrop, M. E., and Laurence, J. H. Manpower for
1985 Military Occupations. Monterey, CA: Naval Post Graduate School.

# EXPLORING A STATISTICALLY VIABLE ASSIGNMENT BASIS USING ASVAB

DR. M. MARK SCHWARTZ

U.S. ARMY INTELLIGENCE SCHOOL
FT. DEVENS, MA

INTRODUCTION. The utilization of the ASVAB tests for determining likelihood of success in MOS training is premised on establishing minimum scores on 1 or 2 of the tests, above which a person is deemed suitable for training in that MOS. This minimum-score concept, in essence, states that (a) all persons above this score are equally likely to succeed and (b) the other tests in the battery are not essential information for determining success in MOS training.

These two points raise issues of (a) is there a maximum score above which a person may be not qualified, and (b) why not use all the data collected? Although there is a range of personal and training-environment factors influencing the likelihood of training success, this paper addresses only the characteristics captured and described by the ASVAB tests. The test results are aptitude/ability indices, summarized in a profile of 10 scores.

Sample profiles are:

TABLE I

| | CO | FA | EL | OF | GM | MM | CL | ST | GT | SC |
|---|---|---|---|---|---|---|---|---|---|---|
| STUDENT 1 | 135 | 147 | 135 | 124 | 128 | 122 | 141 | 138 | 133 | 133 |
| STUDENT 2 | 89 | 94 | 108 | 88 | 109 | 95 | 92 | 101 | 95 | 91 |

As an introduction to this paper's direction, these are real data, where student 1 did not succeed (DROP) and student 2 did succeed (GRAD). Using these two for discussion purposes, these could be called two good examples of "other factor" influence. Both met the criterion for entrance, although the seemingly better candidate wasn't successful. An alternate reaction to these two cases could be that this is simply an example of predictable and expected" classification error". However, both of these explanations dismiss the possibility that there is a meaningful source of explanation in the context of the available data. Even a response of "oh my, it's time to revalidate", although a professional response, is not the only alternative. All three of these reactions, I contend, divert our attention from the system, which collects a lot of data, summarizes it in a profile of 10 scores, and then proceeds to ignore 80 to 90% of the available information in the decision-making process. This is not a condemnation but rather a description, and it begs the question: does the data contain useful patterns of information?

FOCUSING THE QUESTION. By definition, the word "patterns" states that the entire profile and aggregates of profiles are to be addressed. This is immediately different from the concept of selecting one or two scales from the set and then fixing a point on those scales as criteria. The assessment schema has been shifted from looking at individual scores on an absolute scale to looking at the entire profile collectively, and on a relative scale.

Just as the word "patterns" did, the word "relative" signals a different conception. In doing this, we are allowing that there may be contributory and interacting effects of the characteristics these 10 test scores describe. In essence, let us assume that all the information is relevant and that what needs to be determined is how much of each one, relatively, is present in GRADS and DROPS. For example, referring to the two sample profiles in Table 1, the focus is now on the relative amounts of ability/aptitude each person possesses. Notice that on an absolute scale, case 1 and case 2 are distinctly different if we compare their scores on any given test, and although student 1 has higher scores on every test, both students show different patterns of ability/aptitude within their own profile. One way to easily see this is to rank each individual's score profile--in essence, for each individual what are the relative amounts of CO, FA, EL, ETC,-- and in doing so, a picture emerges which shows discrimination between these students, but in a different context.

Ranked, with ties, the two profiles become:

TABLE II

| TEST | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| DROP | 6.5 | 10 | 6.5 | 2 | 3 | 1 | 9 | 8 | 4.5 | 4.5 |
| GRAD | 2 | 5 | 9 | 1 | 10 | 6.5 | 4 | 8 | 6.5 | 3 |

Notice that tests 4,8,9, and 10 show roughly equivalent ranks (both high or low), while tests 3,5, and 6 are low for the DROP when compared to the GRAD, and conversely for tests 1,2, and 7. Note that test 8, which is used for the school selection criterion, shows an "8" for both individuals, indicating that within each individuals repetoire of skills/knowledge/ability/aptitude/-etc., the characteristics captured in this test rank the same relative to the other characteristics in the profile. This analysis presents a different picture from the one seen when looking at raw ASVAB scores. Table I shows clear discrimination between DROP and GRAD, while Table II indicates three different nodes of information -- one in favor of the DROP, one in favor of the GRAD, and one "even". What is now occuring is a horizontal and vertical scanning of"quantity of test X", and it is this simultaneous within and between comparison that provides a basis for "collective and relative" comparison. The question now is

whether this is a significant and useable pattern within and between groups of DROPs and GRADs, and if so, how can these patterns be assessed?

PSYCHOMETRIC CROSSFERTILIZATION. Ranking ignores absolute ASVAB value, as in the case of the ranks of test 4, where a score of 124 (rank = 2) is roughly equivalent to a score of 88 (rank = 1). Realizing inherent theoretical questions about rank and the use of rank means, in order to step up from talking about individuals to groups the mean is the most convenient metric to use. Then each individual rank can be compared to the group GRAD and DROP mean for each test to see if on that characteristic, the individual "looks more like" A DROP or a GRAD. The expression "looks more like" translates roughly to does the individual look more like a member of their group or the other group, using closer-to-the-mean as the criterion.

The purpose of this comparison is to transform the individual's profile of ranks, which was tranformed from the ASVAB profile, into a profile of yes-or-no, symbolized as 1 = looks more like a GRAD on this test, and 0 = looks more like a DROP on this test.

Below is a summary of all the transformations so far.

TABLE III

| | ASVAB | | Rank | | Group Rank Means | | Status (G=1; D=0) | | Overall Classification | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Drop | Grad | Drop | Grad | Drop | Grad | Drop | Grad | Drop | Grad |
| TEST | | | | | | | | | | |
| 1 | 135 | 89 | 6.5 | 2 | 5.17 | 5.05 | 0 | 1 | Drop | Grad |
| 2 | 147 | 74 | 10 | 5 | 7.17 | 6.35 | 0 | 1 | | |
| 3 | 135 | 108 | 6.5 | 9 | 5.53 | 5.17 | 0 | 1 | | |
| 4 | 124 | 88 | 2 | 1 | 4.83 | 4.63 | 1 | 1 | | |
| 5 | 129 | 109 | 3 | 10 | 4.17 | 4.98 | 0 | 1 | | |
| 6 | 122 | 95 | 1 | 6.5 | 3.23 | 3.60 | 0 | 1 | | |
| 7 | 141 | 92 | 9 | 4 | 6.48 | 6.05 | 0 | 1 | | |
| 8 | 138 | 101 | 8 | 8 | 5.85 | 6.57 | 1 | 1 | | |
| 9 | 133 | 95 | 4.5 | 6.5 | 7.03 | 6.70 | 1 | 1 | | |
| 10 | 133 | 91 | 4.5 | 3 | 5.53 | 5.90 | 0 | 0 | | |

The status profiles can be examined to assess an individual's standing on each test. The question is how to collectively summarize all these individual assessments. It is to be noted that performing these transformations can contribute to some loss of fidelity. However, the ultimate results indicate that a viable system is operating. The answer to the question of how to collectively assess the overall status from the individual status indices lies in the application of a Bayesian technique for calculating the probability of group membership. Item Response Theory relies on the natural dichotomy of an item being right or wrong, plus other important item characteristics. But the

concern here is not right or wrong, but "looking like". This Bayesian procedure takes into account each groups total "likeness" for each test (something like the p-value in item analysis) and uses this information with the individual's "likeness" to compute a probability, using all 10 scores, that the individual is a GRAD or DROP. Despite the apparent outcome of classifying the GRAD and the DROP when looking at the status column in Table III, it is to be noted that these cases were selected for demonstration, and not all cases are as clear.

AN EMPIRICAL VIEW AND DISCUSSION. Two simple tests of this conception were done. The procedure was to take 30 actual DROPS and GRADS ASVAB data (a second case had 16 each), transform the scores, apply the Bayesian procedure, and then compare the actual status with the classification outcomes.

The results were:

TABLE IV

| | | Classification | | | | Classification | |
|---|---|---|---|---|---|---|---|
| | | GRAD | DROP | | | GRAD | DROP |
| | GRAD (N=30) | 21 | 9 | GRAD (N=16) | | 12 | 4 |
| ACTUAL | | | | | | | |
| | DROP (N=30) | 11 | 19 | GRAD (N=16) | | 6 | 10 |
| | TOT= | 32 | 28 | TOT= | | 18 | 14 |
| | | Case 1 | | | | Case 2 | |

Statistically, both cases are significant, with chi-square = 6.70 for case 1 and 8.57 for case 2. Classification error for both cases are:

TABLE V

| | Case 1 | Case 2 |
|---|---|---|
| Grad Error Rate | 39% | 25% |
| Drop Error Rate | 37% | 38% |

The values in Table V are roughly equivalent, which can be interpreted as an indication that the approach used operates across these 2 MOSs, suggesting that there may be a viable basis for expectation that it can be applied for all MOS.

Given the small samples used in this pilot research, the somewhat high error rates are just indicators and neither a minimum nor a maximum expectation. As sample size increases, there is no certainty as to the expected change in mean rank, which is a key

value in the overall process. Theoretically, if the law of large numbers operates, all means would go to 5.5. However, some of the reported means in Table III suggest that the likelihood is that a systematic, non-random principle is operating, and the means would not converge.

CONCLUSION. This is by no means conclusive, but rather exploratory. The goal here was not to have error-free classification but rather to demonstrate that models other than the current ones used for assessing profiles of data do show promise and merit further research. This model could be extended to the point of realizing that all the requisite data could be compiled and updated at each school, transmitted to every recruiting office, and utilized to evaluate the likelihood of successful training in a range of MOSs by simply entering an individual's profile of scores into a computer.

# The Validity of ASVAB for Predicting Training and SQT Performance

Paul G. Rossmeissl
U.S. Army Research Institute[1]

Donald H. McLaughlin, Laurese L. Wise and David A. Brandt
American Institutes for Research

This paper is a condensation of a larger report (McLaughlin, Rossmeissl, Wise, Brandt, & Wang; 1984) which investigated the validity of the Armed Services Vocational Aptitude Battery (ASVAB) for predicting success in Army jobs or Military Occupational Specialties (MOS). The ASVAB is a cognitive test battery used by the military services as their primary instrument for selecting and classifying enlisted personnel. This particular research was based upon ASVAB forms 8/9/10 which was composed of ten subtests: General Sciences (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto/Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehensive (MC), and Electronics Information (EI). The two verbal subtests, WK and PC, are most often combined into a single measure of verbal ability called VE. The current version of ASVAB (forms 11/12/13) uses parallel forms of these same subtests.

Scores on the ten ASVAB subtests are typically combined into aptitude area (AA) composites. Examples of these composites are given in Table 1. The Army composites serve as the basis for assignment of personnel to Army MOS in that a minimum qualifying score on one of the aptitude area composites is required for admission to Army initial level training courses. For example, the CO composite is used to classify recruits into the infantry and armor specialties. Similarly, the MAGE composites are used by the Air Force to select and classify prospective personnel into Air Force specialties. The final set of composites routinely in use are the High School Composites which have been developed for use when ASVAB is administered to high school students as a career guidance tool. Maier and Truss (1983) have also recommended that the first four of these composites be used to select and classify enlisted personnel within the Marine Corps.

The goal of the McLaughlin et al. (1984) research was twofold. First, the validities of the composites then in use by the Army and other DoD agencies were evaluated with regard to predicting success within the Army. Second, an additional set of composites were derived empirically in hopes of obtaining a composite system with maximal predictive validity.

In all cases the validation criterion were MOS specific end-of-course training scores or skill qualification tests (SQTs). All of the criterion measures were trimmed of outliers and then standardized before any

---

[1]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.

validation analyses. The separate training and SQT data were combined for validation analyses at the MOS level. All validities were corrected for restriction of range using the multivariate adjustment due to Lawley (1943) and described by Lord and Novick (1968).

Table 1
Typical ASVAB Composites

| | | |
|---|---|---|
| **Army Composites (1983)** | | |
| Clerical/Administrative | CL | VE + NO + CS |
| Combat | CO | AR + CS + AS + MC |
| Electronics Repair | EL | GS + AR + MK + EI |
| Field Artillery | FA | AR + CS + MK + MC |
| General Maintenance | GM | GS + AS + MK + EI |
| Mechanical Maintenance | MM | NO + AS + MC + EI |
| Operators/Food | OF | VE + NO + AS + MC |
| Surveillance/Communications | SC | VE + NO + AS + CS |
| Skilled Technical | ST | VE + GS + MK + MC |
| **MAGE Composites** | | |
| Mechanical | M | MC + AS + GS |
| Administrative | A | VE + NO + CS |
| General | G | AR + VE |
| Electronic | E | AR + MK + GS + EI |
| **High School Composites** | | |
| Mechanical Trades | HSMT | AR + MC + AS + EI |
| Office and Supply | HSOS | VE + CS + MK |
| Electronics/Electrical | HSEE | AR + EI + MK + GS |
| Skilled Services | HSSS | AR + VE + MC |
| Academic Ability | HSAA | AR + VE |

Composite System Validities

Table 2 gives the adjusted validities for each of the composite systems displayed in Table 1. Validities and sample sizes are given for each of the nine clusters of MOS now in use by the Army. The validities were obtained by averaging the validities for the individual MOS within each cluster and weighting by the number of soldiers within each MOS.

Table 2
Validities of Established Composite Systems

Army Composites (1983)

| Cluster of MOS | (N) | Composite CL | CO | EL | FA | GM | MM | OF | SC | ST |
|---|---|---|---|---|---|---|---|---|---|---|
| CL | 10368 | 48 | 51 | 53 | 54 | 49 | 46 | 50 | 50 | 53 |
| CO | 14266 | 36 | 44 | 43 | 43 | 43 | 42 | 44 | 40 | 44 |
| EL | 5533 | 38 | 47 | 46 | 47 | 46 | 47 | 44 | 44 | 47 |
| FA | 5602 | 39 | 49 | 48 | 48 | 49 | 49 | 49 | 45 | 44 |
| GM | 2571 | 39 | 48 | 46 | 46 | 47 | 48 | 48 | 45 | 47 |
| MM | 7073 | 36 | 48 | 46 | 45 | 48 | 48 | 48 | 43 | 46 |
| OF | 8704 | 38 | 48 | 47 | 45 | 48 | 47 | 48 | 44 | 48 |
| SC | 3729 | 39 | 49 | 48 | 47 | 48 | 47 | 48 | 45 | 49 |
| ST | 7061 | 51 | 56 | 57 | 57 | 55 | 54 | 56 | 54 | 58 |

MAGE Composites

| Cluster of MOS | (N) | Composite M | A | G | E |
|---|---|---|---|---|---|
| CL | 10368 | 45 | 48 | 54 | 53 |
| CO | 14266 | 42 | 36 | 42 | 43 |
| EL | 5533 | 45 | 38 | 46 | 47 |
| FA | 5602 | 48 | 39 | 46 | 48 |
| GM | 2571 | 46 | 39 | 44 | 46 |
| MM | 7073 | 48 | 36 | 44 | 46 |
| OF | 8704 | 47 | 38 | 47 | 47 |
| SC | 3729 | 47 | 39 | 47 | 48 |
| ST | 7061 | 52 | 51 | 57 | 57 |

High School Composites

| Cluster of MOS | (N) | Composite HSAA | HSMT | HSOS | HSSS | HSEE |
|---|---|---|---|---|---|---|
| CL | 10368 | 54 | 47 | 54 | 53 | 53 |
| CO | 14266 | 42 | 43 | 40 | 44 | 43 |
| EL | 5533 | 46 | 47 | 43 | 47 | 47 |
| FA | 5602 | 46 | 49 | 44 | 49 | 48 |
| GM | 2571 | 44 | 47 | 43 | 47 | 46 |
| MM | 7073 | 44 | 49 | 41 | 47 | 46 |
| OF | 8704 | 47 | 48 | 43 | 48 | 47 |
| SC | 3729 | 47 | 48 | 44 | 49 | 48 |
| ST | 7061 | 57 | 54 | 56 | 58 | 57 |

The main diagonal of the upper portion of Table 2 gives the validities of the composites that were associated with each of the nine clusters in 1983. The most interesting feature of the data in Table 2 is the uniformity of the validities. All of the entries are between .36 and .58, with the mean validity of each system being about .45. One MOS cluster, ST, appears to be slightly more predictable than the others; and another cluster, CO, appears to be slightly less predictable. The remaining clusters show very little variance.

Identification and Validation of Alternative Composites

In order to develop alternative composites the MOS were partitioned into clusters, based on similarity of ASVAB profiles of successful criterion performance. The similarity between each pair of cells was defined as correlation of the predicted criterion performance in the two cells for the applicant sample. The performance predictions were based on ridge regressions, using the ASVAB subtests as predictors. The cells were clustered by adapting standard "leaf to stem" procedures. Upon finding that the results of the clustering were unstable, due to the high intercorrelations of the predicted criterion scores, the clustering procedure was modified to use as a starting point the Army's current grouping of MOS into aptitude area clusters.

Once a cluster had been defined the unit-weight composite with maximal predictive validity for that cluster was identified. It was found that optimal unit-weight composites for four clusters possessed a root mean square (RMS) predictive validity within 97% of the RMS validity of the ridge regression vectors computed separately for each of the 98 MOS included in the sample. The composition of these four alternative composites are given in Table 3, and their predictive validities are given in Table 4.

Table 3
Optimal Four Composite Solution

| Composite | | Subtests |
|---|---|---|
| Clerical/Administrative | (ACL) | VE + AR + MK |
| Skilled Technical | (AST) | VE + AR + MK + AS |
| Operations | (AOP) | VE + AR + MC + AS |
| Combat | (ACO) | VE + MK + MC + AS |

Inspection of Table 4 shows that by focusing on the most valid portion of the ASVAB, the primary aim of this aspect of the research was achieved: the validities went up. The aggregate RMS predictive validity for the four alternative composites for their assigned MOS is .486, in comparison with RMS validity for the 1983 Army composites of .454. Certain members of the 1983 Army composite set account for a large part of the difference in validity between the two composite sets. When compared

Table 4
Predictive Validities of the Alternative Composites

| Cluster of MOS | (N) | Composite ACL | AST | ACO | AOP |
|---|---|---|---|---|---|
| CL/ACL | 10368 | 56 | 54 | 52 | 51 |
| CO/ACO | 14266 | 42 | 44 | 44 | 44 |
| EL/ACO | 5533 | 46 | 48 | 48 | 48 |
| FA/ACO | 5602 | 47 | 49 | 50 | 50 |
| GM/ACO | 2571 | 45 | 48 | 48 | 48 |
| MM/AOP | 7073 | 44 | 48 | 49 | 49 |
| OF/AOP | 8704 | 46 | 49 | 49 | 49 |
| SC/AOP | 3729 | 47 | 49 | 50 | 50 |
| ST/AST | 7061 | 58 | 58 | 57 | 57 |

to validities of the optimal composites for the same cluster of MOS, the 1983 Clerical composite (CL) appeared to be weak, with a validity of .48 versus a potential of .56. Another composite Surveillance and Communications (SC), was mildly weak, with a validity of .45 versus a potential .50.

Recommendations

A major purpose behind the McLaughlin et al. (1984) report was to present recommendations to the Army as to how the composite system then in use to select and classify enlisted personnel could be improved. The average validity of the set of four empirically derived alternative composites was .48 versus .45 for the existing composite systems. Thus, from a purely statistical point of view the results in terms of predictive validity tended to favor the alternative four composite solution over the nine composite system the being used or any of the alternatives being used by other armed services.

However, considering the costs of implementing a whole new composite system, it was decided that a more favorable proposal would be to maintain a nine composite system but to replace the the two composites which were the major source of the deficiency of the 1983 composites. The new CL composite would be comprised of the VE, AR, and MK subtests and would have a predictive validity of .56. The new SC composite would have a predictive validity of .50 and be made up of the VE, AR, MC, and AS subtests. The average validity of the revised nine composite system would be .47. The Army officially adopted this composite system on October 1, 1984.

The gain in expected performance resulting from the change in the CL and SC composites can only be approximated, because of the constrained nature of the selection and classification process. If, however, the choice were purely between assignment to an individual MOS and rejection, application of Cronbach's formula yields an expected gain of .05 standard deviations of criterion performance per person in the two clusters of MOS from the introduction of the two revised composites.

## References


Lawley, D. (1943). A note on Karl Pearson's selection formulae. Royal Society of Edinurgh, Proceedings, Section A. 62, 28-30.

Lord, P., & Novick, M. (1968). Statistical theory of mental test scores. Reading MA: Addison-Wesley Publishing Company, Inc.

Maier, M. H., & Truss, A. R. (1983). Validity of ASVAB Forms 8, 9, and 10 for Marine Corps training courses: Subtests and current composites. (Center for Naval Analyses Memorandum No. 83-3107). Alexandria, VA: Marine Corps Operations Analysis Group.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A. & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY 1981 and FY 1982 enlisted accessions. Technical Report No. 651, U. S. Army Research Institute for the Behavioral and Social Sciences, Alexandria VA.

# GENDER, ETHNIC GROUP, APTITUDE AND PERSONALITY DETERMINANTS OF U. S. COAST GUARD ATTRITION

ROBERT L. FREY, JR.

U. S. COAST GUARD HEADQUARTERS

All five Armed Services have been using aptitude measures for enlistment screening and technical school qualification for decades. Research consistently has shown that premature first term attrition (i.e., before the end of one's enlistment) is related to aptitude.

However, as we all know, aptitude is only a small part of the potential predictor space. In the Coast Guard we are interested in the possible use of personality assessment to improve our screening. One of the major responsibilities of the Coast Guard is to board vessels for fishing law inspections and drug interdiction investigations. These boardings are conducted by armed personnel; obviously, management is quite concerned that assessment of personnel to determine suitability for armed boarding party duty be as accurate as possible.

The nature of these boardings is similar in many respects to law enforcement activities. For example, it is critical that the people involved exercise utmost discipline and be able to deal with sudden stress without resorting to the use of firearms unless absolutely necessary. Naturally, extensive special training is given to personnel before they are allowed to take part in armed boarding parties in order to meet such performance requirements. In addition, however, this is a performance domain where we expect people's behavioral predispositions to be quite predictive. Accordingly, we are investigating the potential of using personality profiles as part of the assessment process. A major objective in our use of personality profiles is to improve assessments of new recruits for their general suitability to serve and remain in the U. S. Coast Guard.

The personality assessment instruments which we used for this study were the Sixteen Personality Factor Questionnaire (16PF) and the Clinical Analysis Questionnaire (CAQ). There were three major reasons for our choice: (1) The 16PF/CAQ covers the full realm of both normal traits and clinical syndromes, 2) the 16PF/CAQ seems to be the best validated instrument for law enforcement and other occupations involving great stress, and 3) normative data show that the 16PF/CAQ profiles are virtually the same for minorities as for whites--that is, the 16PF/CAQ is fair for minorities.

## RESULTS AND DISCUSSION

The subjects for this study were the cohort of FY83 recruits. The total N was 4,288. There were 3,544 whites and 744 minorities. Also, there were 3,732 males and 556 females. On the predictor side, aptitude test scores and 16PF/CAQ scores were obtained for each recruit. The aptitude measure used was a composite of verbal and arithmetic subtests from the ASVAB or a similar test battery (During part of FY83, the USCG was still using an old Navy battery). The aptitude composite has a normative mean of 100 and a standard deviation of 20. 16PF/CAQ scores are scaled to have a normative mean of 5.5, SD of 2 and a range of 1 to 10.

The analysis design was a crossed factorial multivariate analysis of variance. The three factors were: 1) Gender (male vs female), 2) Ethnic Group (Minority vs Majority), and 3) Status (In the CG vs out of the CG). Since the Coast Guard only has four year enlistments, anyone already out is a case of premature first term attrition. This resulted in a 2 X 2 X 2 design with eight cells. The criteria for this analysis were the aptitude score and the 16PF/CAQ scores.

Since assignment to the cells of the design is by self-selection, the cell N's were quite disparate. Accordingly, all the tests for interactions and main effects were accomplished using a hierarchical, least-squares model. The attrition rates by ethnic group were: 1) majority group - 16.4%, 2) minority group - 28.9%. The attrition rates by gender were: 1) males - 17.5%, 2) females - 25.7%. The overall attrition rate was 18.6%.

The Gender (G) X Ethnic Group (E) X Status (S) multivariate interaction was not significant ($F = 0.928$, $p < .575$). However, the E X S multivariate interaction was significant ($F = 1.692$, $p < .013$, $R = .105$). In order to interpret this multivariate interaction, the discriminant function structure coefficients (or loadings) were evaluated. The structure coefficients are used to interpret the underlying dimensionality of the E X S discriminant function. Four of the criteria explained the discriminant function:

| CRITERION | LOADING |
|---|---|
| WARMTH | .501 |
| APTITUDE | .499 |
| LOW ENERGY DEPRESSION | - .318 |
| SUICIDAL DEPRESSION | - .311 |

As defined in Krug, 1980

Warmth: High-scoring individuals are "...personable and easy to get along with. ...prefer to adapt to other people's schedules. ...share their feelings with others.

Low Energy Depression: "High scoring individuals report frequent feelings of sadness and gloom. ...almost never sleep soundly or wake up full of energy...have little zest for life and are worn out and low."

Suicidal Depression: "Centers around thoughts of self-destruction. High scoring individuals report that they are disgusted with life...entertain thoughts of death as a viable alternative to their present hopeless situation."

The univariate interaction means for these four criteria are:

| | GROUP | | | |
|---|---|---|---|---|
| CRITERION | MAJ-IN | MAJ-OUT | MIN-IN | MIN-OUT |
| WARMTH | 4.36 | 4.34 | 5.02 | 5.54 |
| APTITUDE | 108.2 | 104.3 | 99.2 | 98.8 |
| LOW ENERGY DEPRESSION | 5.01 | 5.77 | 5.05 | 5.48 |
| SUICIDAL DEPRESSION | 4.61 | 5.27 | 4.84 | 5.13 |

NOTE:

MAJ-IN: Majority group personnel still in the Coast Guard

MAJ-OUT: Majority group personnel already out of the Coast Guard
MIN-IN: Minority group personnel still in the Coast Guard
MIN-OUT: Minority group personnel already out of the Coast Guard

By definition of an interaction, we would expect the criteria to operate differently within each ethnic group. For example, within the majority group, there is no difference in Warmth between those still in the Coast Guard vs those already out of the Coast Guard. In contrast, within the minority group, those who are already out of the Coast Guard scored considerably higher on Warmth than those who are still in the Coast Guard. Looking at the Aptitude score, within the majority group, those who are still in the Coast Guard scored considerably higher than those already out of the Coast Guard. Within the minority group, there is no difference on the Aptitude score between those still in the Coast Guard vs those already out of the Coast Guard. For both the Low Energy Depression score and the Suicidal Depression score, the interaction pattern is the same. That is, within the majority group, those who are already out of the Coast Guard scored considerably higher (i.e., "worse") than those who are still in the Coast Guard. Within the minority group, however, the difference between those still in the Coast Guard vs those already out of the Coast Guard is very small.

So far, we have examined those criteria which "explain" the E X S multivariate interaction. However, follow-on analyses are required to clarify the differential status profiles within each ethnic group. Separate analyses of the status factor (using the common multivariate error matrix) were done within each Ethnic Group. That is, an internal analysis of the E X S multivariate interaction was accomplished. The primary interest in the internal analysis was to see whether the discriminant function composite of aptitude and personality variables for predicting status was different for the majority and minority ethnic groups.

The results of the internal analysis indeed showed that the underlying dimensionality of the status prediction composite was quite different for the two ethnic groups.

The multivariate test of the status factor was significant within the majority group ($F = 8.4$, $P < .001$, $R = .229$). Five of the criteria explained the discriminant function:

| CRITERION | LOADING |
|---|---|
| HYPOCHONDRIASIS | - .646 |
| LOW ENERGY DEPRESSION | - .601 |
| APTITUDE | .582 |
| SUICIDAL DEPRESSION | - .559 |
| PSYCHOLOGICAL INADEQUACY | - .534 |

As defined in Krug, 1980:

Hypochondriasis: "High-scoring individuals are depressed and preoccupied with bodily disfunctions. ...[they] feel that their health is worse than others ... feel sluggish ... and generally run down".

Low Energy Depression: (previously defined)

Suicidal Depression: (previously defined)

Psychological Inadequacy: "High-scoring individuals describe themselves as no good for anything ... think of themselves as doomed or condemned ... learned helplessness pattern".

The majority group univariate means for these criteria were:

| CRITERION | GROUP IN CG | GROUP OUT CG |
|---|---|---|
| HYPOCHONDRIASIS | 4.69 | 5.47 |
| LOW ENERGY DEPRESSION | 5.02 | 5.72 |
| APTITUDE | 108.33 | 103.89 |
| SUICIDAL DEPRESSION | 4.60 | 5.31 |
| PSYCHOLOGICAL INADEQUACY | 4.88 | 5.51 |

The four personality scores noted above are clinical syndromes from the CAQ portion of the 16PF/CAQ. Thus, a high score is "worse" than a low score. As expected, those majority group members already out of the Coast Guard scored higher than those still in the Coast Guard. Depression and feelings of inadequacy seem to be a common theme for the majority group members already out of the Coast Guard. Also, those already out of the Coast Guard had an aptitude score lower than those still in the Coast Guard.

The multivariate test of the status factor was significant within the minority group ($F = 2.034$, $P < .001$, $R = .115$). Five of the criteria explained the discriminant function:

| CRITERION | LOADING |
|---|---|
| WARMTH | - .492 |
| HYPOCHONDRIASIS | - .490 |
| SENSITIVITY | - .462 |
| ANXIOUS DEPRESSION | .395 |
| PSYCHOLOGICAL INADEQUACY | - .390 |

As defined in Krug, 1980:

Warmth: (previously defined)

Hypochondriasis: (previously defined)

Sensitivity: High scoring individuals are "... tender-minded, dependent, overprotected, fidgity, clinging, and insecure. ... They prefer to use reason rather than force in getting things done".

Anxious Depression: "High-scoring individuals describe themselves as clumsy and shakey ... they lack self-confidence, and seldom speak out and say what they think. They are confused and unable to cope with sudden demands..."

Psychological Inadequacy: (previously defined)

The minority group univariate means for these criteria were:

| CRITERION | GROUP IN CG | GROUP OUT CG |
|---|---|---|
| WARMTH | 5.02 | 5.53 |
| HYPOCHONDRIASIS | 4.85 | 5.37 |
| SENSITIVITY | 5.19 | 5.71 |
| ANXIOUS DEPRESSION | 5.32 | 5.76 |
| PSYCHOLOGICAL INADEQUACY | 4.88 | 5.51 |

It is immediatley apparent that the dimensionality of the discriminant function composite for the minority group is different from that of the majority group. First of all, aptitude is not involved. Also, two normal traits from the 16PF portion of the test are part of the composite (Warmth and Sensitivity). Interestingly, those minority group members already out of the Coast Guard score higher on these 2 criteria than do those still in the Coast Guard. As expected, those minority group members already out of the Coast Guard score higher on Hypochondriasis, Anxious Depression and Psychological Inadequacy than do those still in the Coast Guard.

Regarding the Warmth and Sensitivity criteria, minorities scored higher than the majority. Since the USCG enlisted force is composed of 83% majority members, one could posit that the majority group culture dominates the enlisted force. It may be that there is a "cultural clash" in the USCG for a significant number of the minority recruits. Thus, the higher Warmth and Sensitivity scores of the minority group members already out of the Coast Guard may indicate those who were most discrepant from the dominant culture.

There is also a practical significance to the separate discriminant function composites for majority and minority group members. The status discriminant function composite was computed using two different methods: 1) based on the total sample, ignoring the E X S multivariate interaction (i.e., from the status main effect test), and 2) separately for each group, based on the internal analysis of the E X S multivariate interaction. Point biserial correlations between the discriminant function composites and actual status were then computed for each ethnic group. For the majority group members, there is no real change--.24 under method 1 and .25 under method 2. This is not surprising since the total sample is 82.6% majority members. However, for the minority group members the correlation under method 1 is .17 while the correlation under method 2 is .26. This is a meaningful improvement for screening purposes.

The results of this study indicate that when one expands the predictor domain beyond aptitude, differential prediction for majority ad minority ethnic groups may be called for. Naturally, in the wider context, the results of one study are always tentative. Meta-analyses (such as are now done with aptitude scores) will be necessary before any conclusions can be reached. With all the work being done on new and supplementary predictors, the possibility of different prediction dimensions for majority and minority group members should not be overlooked.

## REFERENCE

Krug, S E Clincal Analysis Questionnaire Manual. Champaign, ILL.:IPAT, 1980

# STUDY OF WASTAGE FROM THE TERRITORIAL ARMY (UK ARMY RESERVE)

D M Blyth

Personnel Psychology Division
Army Personnel Research Establishment
Ministry of Defence
Farnborough
Hampshire
United Kingdom

## INTRODUCTION

1. At last year's MTA Conference the results of the first phase of a study of the wastage problem in the UK Army Reserve Forces were reported in a paper entitled "Retention in the UK Territorial Army". (Dennison and Blyth, 1984). At that stage the work consisted of a literature review which focussed on international comparisons of reserve force attrition; an analysis of the available wastage and turnover statistics; a programme of interviews and group discussions with serving reservists at a number of Territorial Army (TA) units. The interviews and group discussions were carried out by staff from the Army Personnel Research Establishment (APRE).

2. The purpose of this paper is to summarise the findings of the study as a whole including the second phase which consisted of a nationwide interview survey of people who had been discharged from the TA in the preceeding twelve months. This work was carried out by a commercial Market Research agency using a structured interview schedule designed in collaboration with APRE. The aims of the survey were to quantify the relative importance of the factors associated with wastage, to isolate those factors on which the TA could realistically be expected to take action, either locally or centrally, and to make practicable recommendations for improvements.

## BACKGROUND

3. The British Army's reserves fall into two categories. The first consists of officers and soldiers who have completed their service with the Regular Army. These "Regular Reservists" have a compulsory liability for reserve service on mobilisation and moves are currently underway to encourage them to undertake a limited amount of peacetime training. They are not, however obliged to do so.

4. The second category consists of civilian volunteers (including some ex-regular soldiers and officers) who undertake military training in their spare time and comprise the Territorial Army.

5. The role of the TA is to provide units, some highly specialised, to reinforce the Regular Army at home and abroad in the event of mobilisation. The TA provides almost 30% of the Army's mobilised strength.

6. In 1981 the Ministry of Defence published plans to expand the TA in two phases from 70,000 to 86,000 by 1990 to meet NATO and other commitments. Actual increases in strengths, however, have not kept pace with increases in target strengths. For example, the average total strength of officers and soldiers in September 1984 was 72,041 (having fallen from a high point of 72,703 in April 1983) compared to the target figure of 75,051.

7. The urgent need to improve retention of TA personnel led to the commissioning by the then Director, Territorial Army and Cadets of a study of the reasons for high wastage to be carried out by APRE. The study consisted of an element that was carried out by APRE staff and a separate element for which a research agency with a field force of interviewers was used. The latter consisted of interviews with 426 ex-TA soldiers drawn from a cross section of regions and types of TA unit. The APRE element led to a number of conclusions and recommendations some of which were reported in the MTA Paper referred to earlier. However, rather than report separately the findings of the national survey of leavers, all of the main findings and conclusions from the study as a whole will be reported briefly.

## RESULTS AND CONCLUSIONS

8. The question of whether the current 30% annual turnover is an unreasonably high figure or is capable of being improved upon significantly was examined. The element of turnover attributable to voluntary wastage cannot be measured accurately at present. This is due to inconsistencies among units in categorising leavers and in some cases, ignorance of the real reasons for leaving. It would be more difficult still to estimate how much voluntary wastage could be prevented by remedial action by the TA. However, in view of the similar turnover rates experienced by other countries' reserve forces and the consistency of our own turnover over the years, despite major improvements to terms and conditions, we conclude that no more than a modest improvements can be expected from action likely to be at the disposal of the TA.

9. The available statistics on strengths, wastage and turnover can be misleading. Strength figures include a high proportion of soldiers who have done no training nor attended camp for at least a year but have not been discharged. Discharge categories used to describe types of leaver are not used consistently across units nor within units. It was recommended that more informative monthly, quarterly and annual statistical breakdowns should be produced and that firm guidance should be given to units on the use of discharge categories.

10. The results of the Market Research survey were in broad agreement with those of the APRE interviews and group discussions. In the survey of leavers, when asked for the single most important reason for leaving the TA, 33% gave job pressures, including a change to shift work and related financial problems; 23% gave the harmful effect that the demands on time had on family and social life; 27% referred to inadequacies within the TA such as lack of action and boredom; 7% gave physical demands or medical reasons; the rest gave an assortment of other reasons.

Table 1 gives a breakdown of responses to the questions "What is the single most important reason for not carrying on with the TA?" and "What other things helped you to make up your mind?"

TABLE 1. REASONS GIVEN FOR LEAVING

(What is) the single most important reason why you didn't carry on with the TA?
What other things helped to make up your mind..?

| REASONS FOR LEAVING | TOTAL* % | MOST IMPORTANT REASON % | OTHER FACTORS* % |
|---|---|---|---|
| JOB/MONEY | 48 | 33 | 15 |
| Demands of job | 18 | 15 | 3 |
| Change to shift work | 16 | 12 | 4 |
| Employer pressure | 6 | 3 | 3 |
| Not enough money | 8 | 3 | 5 |
| SOCIAL/TIME | 46 | 23 | 23 |
| Took too much time | 16 | 7 | 9 |
| Travelling time | 5 | 2 | 3 |
| Too many weekends | 2 | 1 | 1 |
| Affected family/social life | 23 | 13 | 10 |
| 'INTERNAL' REASONS | 47 | 27 | 20 |
| Boring/lack of action | 22 | 13 | 9 |
| Discipline too strict | 1 | 1 | * |
| Discipline not strict enough | 5 | 3 | 2 |
| Didn't get skill training | 6 | 3 | 3 |
| Objections to higher ranks | 5 | 2 | 3 |
| Didn't get promotion | 4 | 3 | 1 |
| Equipment inadequate | 1 | 1 | * |
| Didn't like unit | 3 | 1 | 2 |
| PHYSICAL | 10 | 7 | 3 |
| Physical demands | 5 | 3 | 2 |
| Medical | 5 | 4 | 1 |
| OTHER | 17 | 12 | 5 |

* Reasons for leaving were provided by respondents, therefore, these columns do not add up to 100%.

11. Pay and bounty (ie. annual payments contingent on meeting a minimum training commitment) were not perceived to be as important a contributory factor in wastage as might have been expected. Very few rated an increase in bounty as an improvement likely to improve retention, though its effect on recruitment and morale is likely to be more beneficial. A substantial increase in bounty or an alternative major financial inducement would be needed to improve retention significantly.

12. The quality of the training in the TA did not appear to be a major cause of wastage or dissatisfaction; the organisation of training events, however, particularly the frustration and inactivity caused by cancellations and delays was mentioned frequently.

13. The supply of modern equipment for training appears to be patchy but, on the whole, was not thought to be a major problem. Personal clothing and equipment allowances were criticised. Although it is thought unlikely to be an important factor in wastage, the cumulative effect of such minor irritants should not be underestimated.

14. The quality of both TA and Regular Officers and non-commissioned officers was considered to be reasonably high. The crucial importance of energy and imagination in the contribution of Regular training staff was stressed.

15. For those receiving supplementary benefit (ie. a UK welfare payment), and declaring it to the Department of Health and Social Security (DHSS), TA pay can result in benefit (and linked payments such as rent allowance) being reduced or eliminated, causing considerable resentment. Ignorance and rumour can make matters worse. More effective education and counselling for the unemployed TA recruit is recommended. A recent increase in the concession or "disregard" allowed by the DHSS from £4 to £8 per week may ease the sense or unfairness felt by those affected. The effects of this increase can be monitored and, if a problem continues to exist, the MOD will no doubt resume its efforts to persuade the DHSS to take a more sympathetic view.

16. Many leavers do not make a positive decision to leave. A temporary absence may be extended indefinitely through embarrassment, inertia etc. A more personal sympathetic follow-up of non-attenders (visits, telephone calls etc) by TA staff may reduce unnecessary wastage of this sort or, where this is unsuccessful, would improve the unit's knowledge of the reasons for wastage.

17. In the survey of "leavers" those who had stayed the longest tended to be older and more likely to be employed, married and to have children. They are slightly less likely than short stayers to have left school with academic qualifications. They are also more likely to have been members of uniformed youth organisations such as scouts, cadets etc. and to have had relatives in the Regular Army or the TA. This profile is of interest but of limited value in selection since the TA is unlikely to be in a position to reject suitable applicants on the basis of such factors.

18. Respondents were presented with five possible changes to terms and conditions currently under consideration (see Table 2). Nearly half said that one or more of the six would have made them less likely to leave but only the proposal for flexible terms of service (ie. the opportunity to earn some bounty by fulfilling less than current minimum commitment to obligatory training and annual camp) were regarded as particularly persuasive. While recognising the drawbacks, on the evidence of this study, we supported moves to increase flexibility in appropriate cases.

TABLE 2

POSSIBLE CHANGES TO TA TERMS AND CONDITIONS CURRENTLY UNDER CONSIDERATION

1. Splitting the bounty to allow payment of at least £100 as a bonus for attending camp, payable at camp.

2. Raising the third year bounty from £400 to, say £500.

3. Raising the first and second year bounties much closer to the third year bounty, say £300.

4. Issuing Rail Cards (ie. allowing cheap rail travel) to members of the TA similar to Student Rail Cards ie. for a small annual payment.

5. More flexible terms of service for the trained soldier to allow him to stay in the TA and qualify for some bounty for a lower training commitment - perhaps, an 8-day camp and a total of 15 instead of 27 days.

19. This study examined the wastage problem at a single point in time only. We concluded that there is a strong case for instituting a regular survey of currently serving TA soldiers and/or leavers in order to provide an up-to-date assessment of prevailing attitudes (and trends in attitudes) within the TA. This can in turn be compared to wastage rates, changes in terms and conditions and to external social and economic factors to provide more of the information needed for effective policy making.

20. Generally, the efforts made by central and local TA management to maintain the high quality of its internal operations and its relationships with the outside world were evident. Two suggestions for bringing the less effective units up to the quality of the best were:

a. Examine communications within the TA with a view to improving the spread of ideas and best practice.

b. In the overall evaluation of each unit's activities, which currently focusses on training effort and adminstration, an assessment of factors such as liaison with local employers, induction processes, and local efforts to reduce wastage should be included.

21. Local efforts to liaise with employers should be reinforced by a central initiative to educate, involve and organise employers.

## SUMMARY

The study led to the following conclusions:

a. No dramatic improvement in the turnover and wastage rates from the present levels can be expected.

b. There is no single cause of wastage; job and social pressures, perceived characteristics of the TA, and a number of other factors all play a part.

c. No single measure will significantly affect wastage but a number of actions will improve the situation. These include greater flexibility in training commitments, the follow-up of non attenders, avoidance of inactivity during training events, counselling for the unemployed on DHSS benefit entitlements, and further attempts to improve the DHSS treatment of unemployed TA soldiers.

22. In addition to the particular measures covered in the study a number of general recommendations were made in the areas of communications within the TA, compilaton of statistical data, evaluation of units' activities, inplementation of a regular survey of attitudes, recruitment and selection, and the involvement and education of employers.

## REFERENCES


Blyth D M. Study of Soldier Wastage in the Territorial Army. Report of the Army Personnel Research Establishment. November 1984.

Blyth D M and Dennison D. Retention in the UK Territorial Army. MTA Conference Paper, 1984.

Doering Z D and Grissmer D W. What we know and how we know it: A selection research and methods for studying active and reserve attrition/retention in the US armed forces. Second Symposium on motivation and morale in NATO Forces - contributions to the enhancement of the quantity and quality of military personnel, 1984.

Payne G D. Retention in the Army Reserve - A Review. Publication of the Australian Army Psychological Corps, 1981.

Shapland P C. Report of the Committee on the Study of Wastage. Internal UK Ministry of Defence Report, 1978.

Sub-Committee on Armed Forces of the Standing Committee on External Affairs and National Defence, House of Commons. (1981, December 18). Action for reserves. Hull. Quebec KIA 059: Canadian Government Publishing Centre. (Known as the Harquail Report.)

# DIMINISHING RETURNS FROM AN ASSESSMENT CENTRE

by GAIL R HARDY

Department of Senior Psychologist(Naval), (SP(N)), Ministry of Defence UK.[1]

## HISTORY

Royal Marine training has for many years had a high wastage rate, around 40%, but by the late Nineteen Seventies wastage was around 50%, and the Corps was consistently under-manned. In-house investigations had indicated ways in which the training regime and environment could be improved, but the climate of opinion amongst the trainers was that selection was the main problem. They were not convinced that the relatively brief selection procedure conducted in Careers Offices throughout the country was adequate to establish that the applicant had the physical fitness and motivation required to start training. As is so often the case, he trainers were inclined to feel that they could do a better job of selection than the Careers Staff.

In view of the urgency of the problem it was agreed to introduce a Potential Recruits Course (PRC) in October 1981, which combined some of the qualities of an Assessment Centre and a Realistic Job Preview. Each course lasts two and a half days and includes timed gymnasium tests, a timed run, an interview with a Personnel Selection Officer (PSO), psychometric tests, and other informative and recreational activities. Only applicants who have already satisfied the Careers Staff as to their suitability are forwarded to the PRC.

The aims of the PRC may be summarised as follows:

a. to determine whether the applicant is sufficiently physically fit to commence Royal Marine training, and has the necessary determination to overcome difficulties.

and b. to enable applicants to assess the type of life and training offered so that those who are not suitable may self-select out.

Additionally it was hoped that the experience would reduce the 'culture-shock' likely to assail new recruits, and that the trainers' morale would be improved by their participation in the selection process.

The final assessment and selection decision is based on marks awarded by three elements of the PRC team viz., a) the Physical Training

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of Senior Psychologist (Naval), or the Ministry of Defence.

Instructor (PTI) b) the PSO and c) the corporals looking after the course, who assess such factors as humour, cooperation etc.

It is worth noting that 64 recruit training places were lost in order to introduce the PRC, but the sacrifice was thought justified in the hope of reducing the excessive wastage rate. 64 places represents over 10% of recent annual recruit intakes.

EFFECTS OF THE PRC ON WASTAGE

PRCs were held twice weekly during training terms from 1st October 1981, and the first successful course members started training in late 1981/early 1982. Recruit training officially lasts six months, but back-trooping for various reasons can extend it to twelve months. Weather extremes during the year can also influence training results, so it was not until 1983 that a fair assessment of the PRC was possible. Statistical returns for recruits entering during calendar year 1982 showed a training wastage of about 25% - half the figure that had caused so much concern - and it would be fair to say that joy was unconfined.

The folly of unconfined joy became apparent when this author began to prepare a paper on the PRC (Hardy, 1984) to present to the 26th Annual Conference of the Military Testing Association. The first doubt related to the observation that training wastage had been declining steadily for two years before PRC-selected recruits entered training. The second was that the wastage graph for more recent recruits seemed to show a worrying upward trend.

Figure 1 demonstrates all too clearly that the second worry was justified.

As the PRC was introduced in October 1981, approximately one quarter of those entered in the year ending March 1982 had been through the procedure, as against approximately three quarters of those entered in the year ending September 1982. Wastage was down to 36% for the last year shown prior to PRC entrants, and 'bottomed out' for the year erding September 1982 at 24%. Therefore the maximum saving that could conceivably be attributed to the PRC is 12%, and that only if one chose to ignore the existing downward trend in wastage. As there were only 374 entrants in the year ending September 1982, a 12% saving represents approximately 45 recruits; rather less than the 64 training places that were sacrificed to accommodate the PRC.

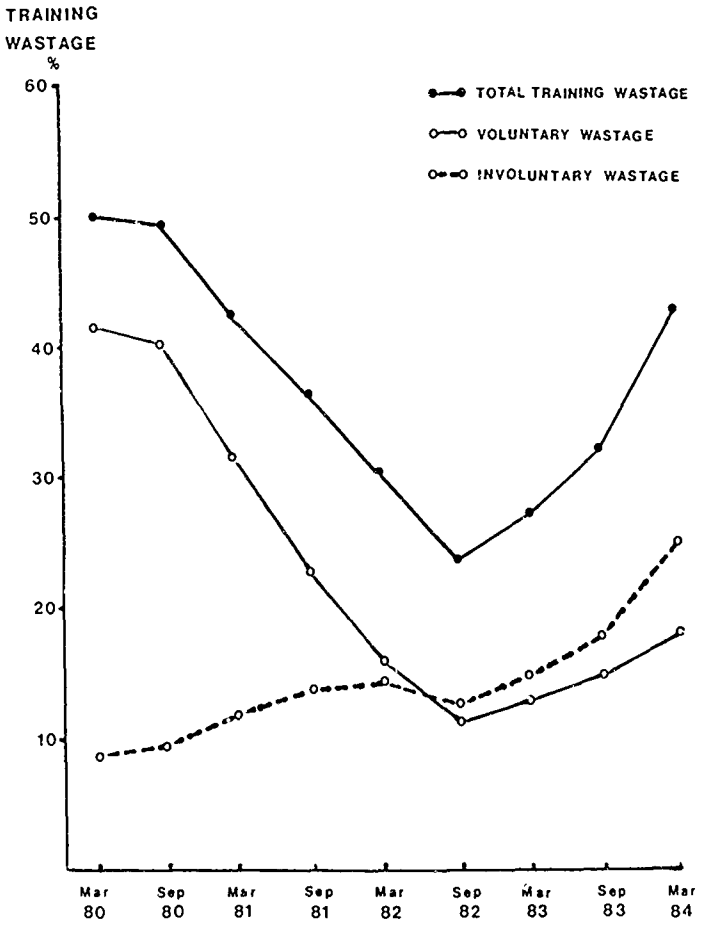
Of greater concern is the clear upward trend in wastage since the low for 1982 Entrants. For the year ending March 1984 wastage was around 42%, 6% higher than it was before the PRC was introduced.

A conspicuous feature of Figure 1 is the change in proportions between voluntary and involuntary wastage. In 1980 voluntary wastage was four

Figure 1

Training Wastage Rate Plotted Against Date of Entry

(Moving Annual Totals)



times as frequent as involuntary wastage, but recent figures show that involuntary wastage is now in the majority.

## THE SEARCH FOR AN EXPLANATION

It has been argued that fear of unemployment was responsible for the original decline in voluntary wastage. As the unemployment of yourg people became more widespread, so the social stigma attached to that condition decreased, with the effect that unhappy recruits were again willing to leave. This is very probably true, but it does not account for the fact that the main increase in wastage is involuntary wastage, that is the trainers rejecting the recuit rather than the recruit opting out.

Early discussions of the problem with responsible authorities produced the comment that the quality of the recruits had declined, and this led to an examination of all the measurable and testable indicators of 'quality' held on the SP(N) data bank. These included psychometric tests scores, educational qualifications, age on entry, and even an assessment of 'Personal Qualities' made by the Careers staff. All these measures remained remarkably stable throughout the period in question, if anything showing a small improvement in some cases. Of course this does not disprove the 'quality' argument, but it makes it more difficult to sustain.

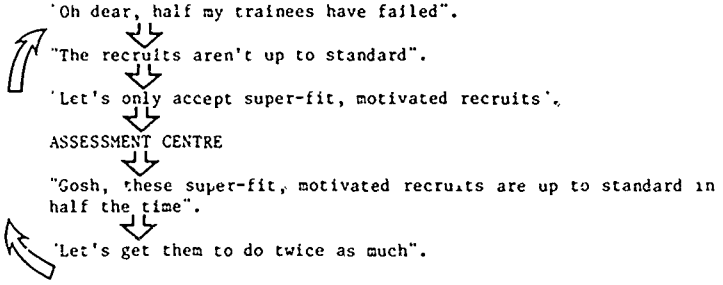
The early decline in voluntary wastage coincided with a marked reduction in the size of the recruit intake, and this led to some promising ideas that strain on training facilities and resources might contribute to voluntary wastage. Fleeting consideration was even given to a perverse type of equity theory, that is, whatever the size of the intake, whatever the selection method and whatever the quality of the recruit, the trainers would always produce the same number of trained Marines. Happily for those of us still clinging to the idea of an ordered world, the 'size of intake' theory has collapsed with more recent figures. Whereas an intake of 374 in year ending September 1982 had a wastage rate of 24%, an intake of 372 for the year ending March 1984 had a wastage rate of 42%.

When wastage trends were discussed with the Training Staff, PSOs and PRC personnel at the training establishment, there was a widely-held view that although training objectives had remained the same, trainers were informally raising their own standards to 'match' the improved capabilities of PRC-selected recruits. It is not in a keen young PTI's nature to allow a troop to lounge around if they have completed the assault course in half the allotted time. However, any unplanned raising of standards will mean that recruits who should have coped moderately well will have difficulty with training, and those who might have enjoyed training may become demoralised and opt to leave - or at least to stop trying.

The syndrome of diminishing returns - even from a successful assessment centre - is illustrated at Figure 2.

Figure 2

Diminishing Returns from an Assesment Centre




A further complication is the impact of the wastage rate on the PRC team. Members of the team are in informal daily contact with the trainers, and it is improbable that the trainers are restrained in their comments on 'poor' selections. Thus the PRC team are under continuous pressure to maintain or raise their selection standards. The PRC pass rate is relatively stable, although it does show a slight downward trend. However, the current pass rate of 55% is low in view of the fact that all course members have been tested and interviewed and found suitable at a Careers Office before being forwarded to the PRC. It seems likely that the standards being set by the PRC, particularly for physical fitness, are more relevant for Marines in training than for civilians wishing to start training. Anecdotal evidence suggests that an excess of zeal during some courses is responsible for several potentially good recruits withdrawing their applications. Both these factors reduce the potential recruit intake, and add to the problems of reaching complement.

HAS THE PRC FAILED?

It is impossible to know what the wastage rates would now be without the PRC - possibly even higher. However, on present evidence the PRC has not achieved what it set out to do. Wastage was falling before the PRC came into operation, and continued to fall as PRC selections started training. By 1983 wastage was climbing again and it is now higher than it was prior to the PRC.

It would be rash to attribute any of these effects solely to the influence of the PRC (or the consequent effect on the trainers). Unemployment, and changing social attitudes to unemployment, almost certainly have some, unquantifiable, effect. Similiarly, it is undeniable that when intake numbers were very high the relative

shortage of irons, drying facilities, telephones, etc for the numbers of users placed additional strain on the recruits. Implementation of some of the recommendations of earlier investigations very probably contributed to the initial fall in wastage. Nonetheless, this author does not believe that the PRC is justifying its existence on present performance.

It should not be thought that the PRC team, the training establishment, or their directing authorities are complacent observers of wastage trends. A detailed report on the problem has been prepared, and a thorough investigation and over-haul of the PRC is under way.

## SOME LESSONS DRAWN FROM THE PRC EXPERIENCE

It is important to remember that although the PRC has many of the characteristics of an Assessment Centre, it is very much an in-house affair, only the PSO being professionally trained in selection. This is not to suggest that the other team members are amateurs, but rather that their professionalism lies in other spheres. However, this is probably true of many of the new crop of Assessment Centres, and the following observations may apply to them equally well.

a. An assessment team that is not working towards a definite quota of passes will tend to raise its' standards rather than accept 'training risks', particularly if it consists mainly of trainers. It is more socially acceptable to be seen to have high standards than low ones, and feed-back tends to concentrate on disasters.

b. An unsupervised assessment team may not set up any scheme to monitor and validate its' own performance, nor even be aware of the need to do so.

c. An over-rigorous assessment centre program may discourage potentially good recruits and cause them to withdraw their applications.

d. An Assessment Centre needs to know the quality of recruit it is required to produce. If the 'customer' informally raises his requirements to match the improved calibre of the recruits, the Assessment Centre is doomed to apparent failure.

## REFERENCE

Hardy, G.R. An Assessment Centre and Realistic Job Preview for Selecting Men for the Royal Marines. **Proceedings of the 26th Annual Conference of the Military Testing Association** 1984. 1049-1054.

# INCREASING BASIC SKILL LEVELS OF MODERATE RISK OFFICER CANDIDATES PRIOR TO COMMISSIONING IN HISTORICALLY BLACK COLLEGES

By
William D. Sprenger
Office of the Deputy Chief of Staff Army ROTC

It has been a source of some concern by authorities within the Army ROTC Headquarters that the basic skills of reading comprehension, writing, and computational ability were not as high as desired. Officer failures may take place at the initial training in an officer's branch which is called the officer Basic Course (OBC). It is here that young second lieutenants from the 16 branch schools across the United States are brought in order to learn the academics of their branch. From a total of 8,500 officers produced each year, the failure rate is about 1.4 percent. This may or may not be considered high, but to some it represents an unacceptable loss in manpower and resources. In 1984 there were 104 failures at the OBC and it was determined that 73 percent were caused by academic failures attributed to low ability to process information and/or poor writing skills.

Due to the complexity and range of quality of institutions of higher education in the United States, the guarantee of appropriate developed abilities cannot be assured--at least for ROTC requirements. Therefore, in 1983, Army ROTC felt the need to institute a qualifying examination titled the Officer Selection Battery (OSB), but this examination was subsequently disallowed primarily because of the adverse impact that it would have on minorities. The expected number of false negatives was deemed unacceptable. Partly as a result of this decision not to permit the OSB to be used as the sole screening vehicle into the advanced phase, an achievement testing program was initiated during the school year (SY) 1983-1984 that was to test the entire population of cadets (which at that time was about 72,000) in reading comprehension, English expression and mathematics. It is hoped that all of this will be completed and the new cognitive screens in place for the school year 1986-1987.

A concomitant benefit is that the testing program immediately places ROTC in the "value-added" business which is probably the most discussed program in higher education, although the techniques and its usage have been around for a considerable time.

Besides their appeal to the American psyche, value-added programs seem to be very worthwhile for certain institutions with certain types of students. In effect, ROTC will have approximately 1,000 "soft value-added" programs and 21 "hard value-added" programs. The qualifier "soft" is added since few contingencies are placed on the Professor of Military Science (PMS) to ensure completion of the program at this time, and since no criterion score is required for contracting or

commissioning (either in HBC or non-HBC) other than the PMS recommendation and a grade point average of 2.0.

Yoked to the achievement testing program is the main topic of this paper, which is the Enhanced Skill Training ROTC Cadet Skills Development Program (EST). From 1916 to 1976, 21 Historically Black Colleges (HBC) were established as host institutions with an additional 53 formed as non-host schools. Only those enrolled at host colleges are eligible to participate in the skills program. These 21 schools account for 59 percent of the black officer production with the extension centers and cross-enrolled producing the remaining 41 percent (HBC only). Only those at HBC and enrolled in the EST courses have what could be a contingency in order to obtain a commission, that is, they must complete the courses in order to be commissioned, although in some cases doing so extends the college career by a semester, and in some cases, two semesters. The EST program is curious in that the federal government is involved in cooperating with 19 schools (two institutions elected not to participate during the SY 84-85) to produce higher basic skill levels. So far, the cooperation of the HBC has been nothing short of admirable even though both the schools and the Office of the Deputy Chief of Staff for ROTC had to work through the complexities of government contracting.

Only those cadets enrolled at the 21 host HBC colleges are eligible to participate in the skills program. These 21 schools account for 59 percent of the black officer production with the extension centers and cross-enrolled producing the remaining 41 percent (HBC only).

Over the years the HBC have produced exemplary officers, but due to several rather recent social-political factors, the percent of black OBC failures has increased and is considered unacceptably high. Of the total number of OBC failures, black officers (HBC and non-HBC) failures account for 55 percent with HBC contributing at the rate of 30 percent. The situation is becoming intensified because the Army intends to meet objective force requirements in 1990 and beyond, and to do so Army ROTC will be required to produce a steady state of 12,110 officers each year. Of this number, approximately 2,500 black officers will be required to mirror in leadership positions the enlisted force composition. To meet the challenge of producing black officers in the numbers needed and to avoid the adverse impact that a single mental screen would provide (such as the OSB) the EST program was launched in September 1985.

The essential features of the program are:

1. It provides additional classroom instruction funded by the Army that is designed to correct weaknesses in communicative skills, mathematics, reading, and cognitive skills.

2. Instruction complements rather than duplicates institutional subjects.

3. Instruction meets the professional education requirements for Army writing.

4. It requires individuals to be tested prior to entry in order to determine deficiencies and again at the end to measure progress made (in other words, a value-added program).

5. Instruction will be provided the enrolled cadet from the beginning of the freshman year up to the time of commissioning.

6. Instruction augments, but is not conducted in lieu of military science training.

7. A finite amount of funding is available through the contracting procedure which is based on ROTC enrollment figures. Cadets with the highest scores are to be selected for the enhanced skill training.

Each HBC will enroll contracted and non-contracted cadets if the cadet scores below the 6th stanine on the Stanford Mathematics Test, below the 6th stanine on the Missouri College English Test and below the 7th stanine on the Nelson-Denny Reading Test. A student may test out or exit the courses if he/she scores at or above the stanine mentioned above. For contracting purposes the school is considered to have met the terms of the contract if a student scores at the 5th stanine in mathematics, the 4th stanine in English, and the 5th stanine in reading.

Table 1 shows the mean percentile and raw scores on the achievement tests for the four ROTC regions and for males and females. As can be seen from the table, the Region means are ordered consistently across all tests, with Region IV being the highest followed by Regions II, I, and III. Region I includes those states up and down the East Coast, Region II is comprised of states in the Upper Midwest, Region III is located in the Central South and Southeast, and Region IV holds states in the Upper Great Plains and West.

These regional differences appear to reflect the distribution of HBC but one cannot assume that the HBC are the sole causal factor in mean score ranking. Female cadets scored higher than males on the English test, whereas males scored higher than females on the mathematics test. There were no overall differences between men and women.

Table 2 shows the means of HBC and non-HBC in each Region. Among the HBC, Region II means are the highest, followed by Regions I and III. Region II has the fewest HBC; Region IV has none. Among the non-HBC, Region IV scored the highest, but the means of Region I and II are not far below the Region IV means. Region II means are the lowest among the non-HBC schools.

The results of all tests indicate a sizable difference in the mean scores among racial groups. For example, on the English test, 93 percent of the white cadets attained the percentile score of 10 (raw score of 33) compared to 59 percent of the

black cadets and 67 percent of the Hispanic cadets. Of the total number, 15 percent tested would fall below this score (see the Total Column), but blacks and Hispanics would constitute a disproportionate number of those failing to achieve a raw score of 33. It is estimated that a raw score of between 45 and 50 is an indicator of at least average writing ability.

The HBC cadets have talent and brain power and one must not be led astray by average scores. It is the intention of Army ROTC to obtain from the HBC those cadets that have the highest test scores and those that show "reaction potential" toward EST. In any case, those cadets that are recommended for commissioning by the PMS will have scores in the range that will predict ability to compete in the Army system as beginning lieutenants as well as through the field grade ranks.

Table 1. Means of Percentile and Raw Scores by Region and Gender
(MS IV, SY 83-84)

| | | Mean Percentile | Mean Raw Score | No. of Cases | % of Total |
|---|---|---|---|---|---|
| English: | Total | 40.2 | 52 | 7358 | |
| | Region I | 39.6 | 52 | 3097 | 42% |
| | Region II | 42.1 | 53 | 1614 | 22 |
| | Region III | 35.5 | 50 | 1456 | 20 |
| | Region IV | 44.7 | 55 | 1191 | 16 |
| | Males | 38.9 | 52 | 6178 | 84 |
| | Females | 46.7 | 55 | 1180 | 16 |
| Reading: | Total | 43.0 | 126 | 7515 | |
| | Region I | 42.5 | 126 | 3262 | 43% |
| | Region II | 45.5 | 128 | 1598 | 21 |
| | Region III | 36.7 | 119 | 1424 | 19 |
| | Region IV | 49.0 | 131 | 1231 | 16 |
| | Males | 43.3 | 126 | 6267 | 84 |
| | Females | 42.4 | 125 | 1196 | 16 |
| Math: | Total | 43.2 | 34 | 8215 | |
| | Region I | 41.9 | 34 | 3189 | 39% |
| | Region II | 45.5 | 34.5 | 2032 | 25 |
| | Region III | 37.1 | 33 | 1664 | 20 |
| | Region IV | 50.5 | 35.5 | 1330 | 16 |
| | Males | 44.9 | 34.5 | 7001 | 85 |
| | Females | 33.2 | 32 | 1214 | 15 |

The percentile scores are based on the following comparison groups:
English: Four-year university and college freshmen (1964).
Reading: Four-year university and college freshmen through seniors (1981).
Math: College-preparatory high school seniors (1965).

Table 2: Means of Percentile and Raw Scores by Region and School Type (MS IV, SY 83-84)

| | | HBC | | | NON - HBC | | |
|---|---|---|---|---|---|---|---|
| | | Mean Percentile | Mean Raw Score | No. of Cases | Mean Percentile | Mean Raw Score | No. of Cases |
| English: | Region I | 19.1 | (40) | 358 | 42.2 | (53) | 2739 |
| | Region II | 25.7 | (45) | 42 | 42.5 | (54) | 1572 |
| | Region III | 14.8 | (37) | 241 | 39.6 | (52) | 1215 |
| | Region IV | | NO HBC | | 44.7 | (55) | 1191 |
| Reading: | Region I | 13.1 | (91) | 328 | 45.8 | (128) | 2934 |
| | Region II | 25.5 | (106) | 42 | 46.0 | (128) | 1556 |
| | Region III | 10.5 | (86) | 237 | 41.9 | (125) | 1189 |
| | Region IV | | NO HBC | | 49.0 | (131) | 1221 |
| Math: | Region I | 18.1 | (27) | 382 | 45.1 | (34.5) | 2807 |
| | Region II | 19.1 | (28) | 34 | 45.9 | (34.5) | 1998 |
| | Region III | 16.1 | (26) | 252 | 40.8 | (33.8) | 1412 |
| | Region IV | | NO HBC | | 50.5 | (35.5) | 1330 |

The percentile scores are based on the following comparison groups:
English: Four-year university and college freshmen (1964).
Reading: Four-year university and college freshmen through seniors (1981).
Math: College-preparatory high school seniors (1965).

# DIRECT MEASUREMENT OF ROTC CADETS' WRITING SKILLS

by

James P. Hanlon
Shippensburg University

We all agree that literacy is important. We believe it is especially important that our organizational leaders and managers are indisputably and demonstrably literate. But do we really know why literacy is important? Conventional explanations of the importance of literacy simply expand on the ideal of "an officer and a gentleman." That is, we rightly acknowledge how literacy enables a person to function, to perform his duty, to be an officer: That person must be able to transmit written messages clearly, concisely, and crisply, so also should he be able to receive them accurately and act upon them properly, with dispatch. No one, then, disputes the practical, functional role of literacy. On the other hand, we at least defer to, pay lip service to the idealistic role of literacy in officership. So literacy is likewise supposed to be an attribute of the gentleman--the person of sensibility, of humane tradition and values. Thus Breaker Morant, officer-hero of the fine Australian film bearing his name and depicting a tragic incident of the Boer War, was all the more a compelling protagonist because, in addition to being a fine soldier, he was as well an accomplished singer and poet: an officer and a gentleman.

But when the crunch comes, when we draw bottom lines, when we make hard budgetary and personnel decisions, we recognize the inadequacy of lip service and conventional reasoning. The functional rationale for literacy is all too self-evident: People who cannot deal competently with written texts sooner or later simply do not make it in the professional world we share. The idealistic rationale for literacy, on the other hand, is pretty much archaic cavalier and Victorian baggage--good enough for movies, I suppose.

Why then, is literacy important? To get at this question, we first should break literacy down into its two components: reading and writing. For writing is what I really want to talk about. Instead of viewing these activities as communications theorists do from a sending/receiving perspective, let's view them on a passive/active scale. Reading is essentially passive, even though more complex texts require higher-order interpretative, cognitive, and analytic skills. So reading may carry you from passive to active on my hypothetical scale. But writing begins with activity. And it may lead to true productivity, or pro-activity. A writer leads the reader through the text, whereas a reader follows a text. A fully literate person is a writer as well as a reader.

Only very recently has our educational establishment begun to seek out and deeply acknowledge the learning potential of writing. Put simply, writing is probably the best vehicle for learning and knowing in our educational arsenal. A good writer exhibits very high-level critical skills, reasoning ability, organizational skills, and --through awareness of audience--interpersonal skills. But, more importantly, through practicing the craft of writing, the student writer develops these skills. Perhaps writing is our only widely available means for developing these skills fully and collectively. Because we value these skills and abilities, because these skills and abilities contribute greatly to organizational effectiveness and dynamics, we seek and need potential career leaders who can and will write. A writer can reveal that he knows and what he really knows. In fact, through writing he discovers what he knows. So I believe this epistemological justification (writing as a way of knowing)

for writing provides the real, as opposed to the conventional, explanation for the high value we place on literacy.

Having then briefly advanced the fundamental argument for the importance of writing, allow me to share a few of my assumptions regarding the nature of writing, assumptions that are more or less germane to my proposal that the writing of officer candidates should be directly examined.

1) Writing is a technology: It is a very old technology, and a critical event in the European Renaissance was the invention of the printing press, which provided our first mass-produced consumer item, the book. As a technology, it can be learned on a skills-acquisition model, as opposed to a biological developmental model. Whereas we may be justifiably concerned if a young person does not develop listening and speaking capabilities by a certain early age, we should not make adverse judgments regarding a person's intelligence or learning abilities if that person has not acquired sophisticated writing skills by, say, the age of twenty. Nor should we blithely assume that, simply because that person has graduated from high school and perhaps done some college work, he is a poor learner of writing. He may well not have been instructed in writing, nor may he have been required to write in school, or anywhere else.

2) Writing is a discipline: Writing systems are wholly arbitrary and thoroughly conventional, nothing less than an elaborate social contract. Mastering a writing system requires incredible discipline--to say nothing of time. Further, the task of writing--of employing the system--is likewise exacting, frustrating, and time-consuming.

3) Writing individuates: No two persons will respond alike to a given writing situation. Whereas it is highly unlikely that a person would seek to copyright his response to the SATs (and, given the number of takers, he may have trouble proving his was original, for there is some statistical probability that other sheets are identical to his), that person could well seek sometime to copyright a love lyric, even though such lyrics abound in our culture.

4) Writing is not a rule-governed activity: We do not write, or learn to write, formulaically, by strict sequential prescriptions. To be sure, through automatization, we must acquire sufficient control of the surface features of the arbitrary writing system to focus on the content of our texts, rather than on the mechanics of producing them. A California throughway driver who has to think about where his brake pedal is and about which foot he wants to use on it is in BIG TROUBLE when, in a split-second of awareness, he knows he must stop his car instantly. So also is a writer in a different kind of trouble if he thinks predominantly about spelling and mechanics of punctuation as he writes. Unfortunately, this troubled writer may have been taught (should we say mistaught?) that spelling and mechanics are what writing is. In fact, writers develop incredibly varied, but seldom disclosed, strategies for composing. They learn these strategies through practice as well as through their experience as readers in learning how texts work.

5) Writing in the real world is a corporate activity: Almost all published, finished texts are produced through collaboration. Early drafts are shared and often benefit from early trusted-reader input. Many drafts are worked up and worked through in committee. Editors and advisors intrude and sometimes revise. Software programs mechanistically edit and silently correct predictable surface errors. Proofreaders check the final draft in scrupulous detail. Developing writers should

learn early how to buy into the corporate writing network, to use it to enhance and promote their authorship.

Since writing is, then, a corporate activity and since effective writing is essential to organizational welfare, I believe that there should be a larger measure of organizational concern for written performances within the organization. For example, many colleges and universities are developing programs of writing-across-the-curriculum. Such programs involve many faculty, rather than only English faculty, in teaching writing. Likewise, many colleges have developed writing centers and tutorial programs in writing to provide extra instruction in writing for those who want, need, or seek it. Graduate and professional schools are offering more and more writing courses for students seeking degrees in law, business, medicine, and even theology. So also are writing samples being used increasingly as part of selection criteria for admissions into both advanced undergraduate and graduate programs.

Officer training programs could as well use writing samples both as an instructional resource and as part of the selection process. I have recently proposed that the Army ROTC collect three writing samples from all cadets during their first two years in the program, prior to their contracting for the Advanced Course. Consider these justifications:

1) Writing samples provide direct evidence of writing ability: Other measurements provide only indirect evidence, even though such evidence may indicate very reliably how well the test-taker will perform in a writing course. In effect, measures of reading ability, of proofreading skills, and of common rhetorical ploys provide predictable measures of writing ability. But a series of writing samples provides a fully valid direct indication of writing ability.

2) Writing samples teach: Well-designed assignments show test-takers how to approach the task. Also, preliminary announcement of topics allows students time for the pre-writing tasks of talk and reflection. Likewise, people learn to write by writing, so each writing experience is a learning experience. Finally, the content of the writing could, through design of the assignment, relate directly to the ROTC Program of Instruction.

3) Collection of writing samples sends out the proper message: Consider this definition of motivation: People do what they have to do in order to do what they want to do. It follows, then, that a person who wants to be an officer will learn to write competently . . . if that person is certain he has to write competently. If that person is taking an indirect measure of writing ability, he will learn vocabulary words, proofreading rules for punctuation and capitalization, recognition of common grammatical errors, because these are things he has to know. But if that person is submitting a series of writing samples, then he will have to learn to write. In fact, you may see him skulking off to a writing center or seeking trusted-reader advice on a draft from a cadet who is recognized as a good writer. Likewise, through the simple expedient of allowing the cadet to submit as many samples, through time, as he wants to and of using only the final sequence of three as efforts of record, the cadet can then use the evaluation as a true learning experience and vehicle for improvement. He can, in effect, write his way toward commissioning.

Time and space do not allow me to elaborate upon the administration and evaluation of writing samples. Nonetheless, I will itemize some suggestions regarding procedures for handling and evaluating writing samples through the ROTC programs:

1) Each semester, PMSs should be provided with instructional packets for administering a writing sample. All first- and second-year cadets should submit samples each semester under carefully defined controlled circumstances.

2) PMSs should forward the sample to a contracted agency for holistic or primary trait evaluation on a 1-6 scale by two readers, and a third reader in the event of widely divergent first and second readings. The nationwide sample should be subjected to a common evaluation.

3) The contractor should report evaluations in a timely fashion to ROTC Headquarters, which would break out and distribute sets of evaluations region by region, detachment by detachment, cadet by cadet.

4) The contractor should store samples for a reasonable period of time, maintain accurate records of individual reader evaluations of each sample, and provide accurate records of reader reliability.

Without doubt, ROTC commitment to collecting and evaluating writing samples would be a considerable administrative and financial undertaking. But such an undertaking would provide indisputable evidence of the Army's commitment to quality in written communications at the entry level of commissioning. In the final analysis, this would be nothing less than a commitment to officer and organizational quality themselves.

# PROCEDURES FOR REFINING WRITTEN MEASUREMENTS AT USAF OTS

SYDNEY SAKO
AND
LT COL WILLIAM J. SLAUGHTER

Officer Training School, USAF
OTS/MTC, Lackland AFB, Texas 78236

## INTRODUCTION

Personnel at the Officer Training School, USAF, are always looking for ways to improve the product of the Officer Basic Military Precommissioning Course. During the past 25 years of school operation, with the production of over 80,000 graduates, one of the key factors of training has been the refinement of testing instruments. Among the several different types of measurements, the use of the consolidated written tests stands out as one of the major criteria in the graduation of students and selection of distinguished graduates. This paper presents some of the ways of refining written tests. It discusses test item writing, Test Review Board decision-making, item analysis, curriculum area achievement analysis, and statistical analysis of test scores. Information from these validational procedures was used in making decisions for conducting changes on specific items of the written tests. A total of over 200 test item revisions were made during 1983-1985. These revisions have resulted in improving significantly the test item bank by increasing the curricular validity of the items, lowering the standard error of measurement, and producing more reliable test scores.

## POPULATION SAMPLE

The population sample used in this study consisted of over 5,000 students enrolled at OTS during the period of 1983-1985. Approximately 70 percent of the population have had no previous military service, and the remainder had served in enlisted rank, or active duty under the Airman Education and Commissioning Program. Their median age came close to 25, a drop from 27 observed during the years of 1978-1980. Almost 40 percent of them are married. All of them possess bachelor degrees, with approximately 5 percent having master or doctorate degrees.

## MEASUREMENTS

At OTS we use a total of eight different types of measurements which include Consolidated Written Tests (CWTs), a Military Briefing, Graded Letters, a Communication Security Test, an Airman Performance Rating, a Flight Drill Performance Rating, Physical Fitness Tests, and two Officer Trainee Effectiveness Ratings. We have limited our study to the CWTs, since they are one of the most important criteria for graduation and for selection of distinguished graduates. The CWTs are designed to measure student achievement in the curriculum areas of Communicative Skills (CS), Leadership and Management (LM), Professional Knowledge (PK), and Defense Studies (DS). There are five different levels of CWTs used at OTS, with three alternate forms for each level. CWTs are administered at specific intervals in training; namely training days 15, 25, 34, 42, and 54. They are designed as criterion-referenced

tests, using a total of 235 test items that cover 67 criterion objectives of the course. (See Atch 1 for samples of criterion objective topics.) The minimum achievement standard for meeting the criterion objectives is set at 80 percent.

## PROCEDURES AND RESULTS

The procedures used at OTS for test refinement include test item writing, Test Review Board decision-making, item analysis, curriculum area achievement analysis, statistical analysis, and pretest/posttest analysis. Pretest/posttest analysis is seldom used at OTS because of its high cost; therefore, it will not be discussed in this paper. The blueprint for test construction is the table of test specifications which is accomplished by a joint effort of testing specialists and subject matter experts. In order to maintain a high curricular validity, all test items are written by subject matter experts, also called Curriculum Area Managers (CAMs). Testing specialists make further refinement by using various statistical and other validation studies. The CAMs develop all course material and present lectures in their area of responsibility. They also write test items using well-established procedures which they have learned from attending the Test and Measurement Course on base and the 5-week Academic Instructor Course at the Air University. Once a CAM has developed a test item, it is submitted to a Test Item Review Committee for evaluation. This committee of other CAMs and division supervisory personnel critically review each test item, suggests improvements, and eventually approve every question before it can be included in a CWT.

During item analysis, we can tell how each item is functioning by looking at ease indexes and tabulation of alternate responses. As .80 is the minimum cut-off, any item with an ease index of .79 or below is checked for causes of low index reading. The cause may be due to ambiguity in item construction, inadequate emphasis of material, or insufficient assimilation or retention by students. An item analysis of over 5200 item administrations based on testing of 20 classes (Class 84-06 through 85-11) produced 271 items which fell below the ease index of .80. Flight commander (FC) comments and items below .80 were reviewed at the Test Review Board meetings attended by the chiefs of Curriculum Instruction Branch and Measurement Branch, CAMs, and FCs. The Test Review Board revised a total of 201 items.

The purpose of the curriculum area (CA) achievement analysis is to determine the amount of attainment of curriculum material in CS, LM, PK, and DS. This is accomplished by tallying the number of individual student failures in each flight, computing the percent failure rate by squadron, and then averaging the failure rate by class (see Atch 2). The goal is to achieve at least 90 percent in each area.

In CA achievement analysis categorized by class, the failure rate should not exceed 15 percent. If it is greater than 15 percent we must again determine whether the high failure rate is due to instruction, the test itself, or the curriculum material. In CA analysis by class, the failure rate was analyzed on 240 different CA administrations. Studies showed that the failure rates of 214 of them were less than 16 percent. The remaining 26 had failure rates of 16 percent or more distributed throughout the 21 classes (see Atch 3). The areas which showed the largest number of failures were DS with a total of 12, CS with 9, and LM with 5. There were none in PK which showed that all

students had passed the minimum requirements. This information is particularly useful to testing specialists, CAMs, and FCs. Testing specialists can pinpoint specific items with low ease indexes. CAMs can check their training material for any discrepancies, and FCs can emphasize those particular areas of difficulty during instructional hours.

In statistical analysis of test scores we obtain the differences of mean score (M), standard deviation (SD), reliability index (R), and standard error of measurement (SE) between the standardized and new tests. The norms of the standardized tests were obtained by averaging the data of operational tests administered to 20 classes. The new tests used in this study were administered to Class 85-11. Statistical results of the new tests were compared with those of the standardized tests, as shown below:

| CWT 1 | M | SD | R | SE |
|---|---|---|---|---|
| Standardized | 90.1 | 5.6 | .91 | 1.7 |
| New | 90.3 | 6.3 | .90 | 2.0 |
| Differences | .2 | .7 | .01 | .3 |

| CWT 2 | M | SD | R | SE |
|---|---|---|---|---|
| Standardized | 90.8 | 5.2 | .90 | 1.7 |
| New | 91.7 | 5.6 | .91 | 1.7 |
| Differences | .9 | .4 | .01 | 0 |

| CWT 3 | M | SD | R | SE |
|---|---|---|---|---|
| Standardized | 90.9 | 5.1 | .90 | 1.6 |
| New | 89.9 | 5.8 | .89 | 1.9 |
| Differences | 1.0 | .7 | .01 | .3 |

| CWT 4 | M | SD | R | SE |
|---|---|---|---|---|
| Standardized | 89.7 | 5.8 | .89 | 1.9 |
| New | 87.5 | 6.5 | .87 | 2.4 |
| Differences | 2.2 | .7 | .02 | .5 |

| CWT 5 | M | SD | R | SE |
|---|---|---|---|---|
| Standardized | 91.3 | 4.9 | .91 | 1.5 |
| New | 90.5 | 4.3 | .89 | 1.4 |
| Differences | .8 | .6 | .02 | .1 |

The differences between the standardized and the new tests were not significant; therefore, the new tests were put into operational status with only minor changes. By having the CAMs write the test items on their own material; by using the Test Review Board meetings to identify and take immediate corrective action on weak areas in testing, instruction, and curriculum materials; and by taking other curricular and test improvement actions when identified through item analysis, CA achievement analysis, and statistical analysis of test scores; we are continuously refining the testing and instructional programs at OTS in order to produce the finest young officers for the United States Air Force.

## SAMPLES OF CRITERION OBJECTIVE TOPICS AT OTS

| | |
|---|---|
| CS-2 | EFFECTIVE WRITING |
| CS-3 | COMMUNICATIVE PROCESS |
| CS-4 | AF PUBLICATIONS |
| LM-1A | TI' MANANGEMENT & GOAL SETTING |
| LM-1E | STRESS MANAGEMENT |
| LM-2B | LEADERSHIP SURVEY |
| LM-2C | LEADERSHIP AUTHORITY |
| LM-2E | LEADERSHIP RESPONSIBILITY |
| LM-2F | LEADERSHIP CONCEPTS |
| LM-3C | MORAL OBLICATIONS AND STANDARDS OF CONDUCT |
| LM-4E | GROUP DYNAMICS |
| LM-5A | MANAGEMENT IN THE AIR FORCE |
| LM-5B | MANAGEMENT FUNCTIONS & DUTIES |
| LM-5C | SUPERVISION |
| LM-5D | MOTIVATION THEORY |
| LM-5F | DELEGATION PROCESS |
| LM-6 | COUNSELING |
| PK-1 | AIR FORCE CUSTOMS AND COURTESIES |
| PK-2 | AIR FORCE UNIFORM |
| PK-5 | PROFESSIONAL RELATIONS |
| PK-6A | PAY, ALLOWANCES & LEAVE |
| PK-6B | MILITARY JUSTICE & DISCIPLINE |
| PK-8 | PERFORMANCE RATINGS |
| PK-9 | AIR FORCE DRUG/ALCOHOL PROGRAMS |
| DS-1 | AIR FORCE HERITAGE |
| DS-2 | MAN AND CONFLICT |
| DS-3 | POWER IN CONFLICT |
| DS-4 | DEPARTMENT OF DEFENSE |
| DS-5A | EMPLOYMENT OF AEROSPACE FORCES |
| DS-5B | STRATEGIC ISSUES OF THE 80s |

Atch 1

## SAMPLE

## CURRICULUM AREA (CA) ACHIEVEMENT ANALYSIS

CLASS 85-06 CWT 4 TOTAL STUDENTS. 202 AVERAGE SCORES: 90.9

Part I: Number of Students Failed

| SQ | FLT | STUDENTS | CS | LM | PK | DS | TOTAL CA FAILURES |
|---|---|---|---|---|---|---|---|
| 1 | 10 | 19 | 4 | 0 | 0 | 1 | 5 |
| 1 | 12 | 18 | 5 | 0 | 3 | 2 | 10 |
| 1 | 14 | 20 | 6 | 1 | 1 | 2 | 10 |
| TOTAL: | | 57 | 15 | 1 | 4 | 5 | 25 |
| 2 | 10 | 16 | 4 | 0 | 2 | 1 | 7 |
| 2 | 12 | 17 | 6 | 2 | 1 | 0 | 9 |
| 2 | 14 | 15 | 5 | 1 | 0 | 1 | 7 |
| TOTAL: | | 48 | 15 | 3 | 3 | 2 | 23 |
| 3 | 10 | 18 | 4 | 0 | 0 | 1 | 5 |
| 3 | 12 | 17 | 4 | 1 | 0 | 0 | 5 |
| 3 | 14 | 15 | 2 | 0 | 2 | 2 | 6 |
| TOTAL: | | 50 | 10 | 1 | 2 | 3 | 16 |
| 4 | 10 | 18 | 4 | 0 | 1 | 2 | 7 |
| 4 | 12 | 14 | 0 | 0 | 0 | 1 | 1 |
| 4 | 14 | 15 | 1 | 0 | 1 | 1 | 3 |
| TOTAL: | | 47 | 5 | 0 | 2 | 4 | 11 |

Part II: Percent of CA Failures by Squadrons

| | SQ 1 | SQ 2 | SQ 3 | SQ 4 | CLASS AVG |
|---|---|---|---|---|---|
| Communicative Skills | 26.3% | 31.2% | 20.0% | 10.6% | 22.2% |
| Leadership and Management | 1.7 | 6.2 | 2.0 | .0 | 2.4 |
| Professional Knowledge | 7.0 | 6.2 | 4 0 | 4.2 | 5.4 |
| Defense Studies | 8.7 | 4.1 | 6.0 | 8.5 | 6.9 |

Atch 2

CURRICULUM AREA ACHIEVEMENT ANALYSIS
BY CLASS AND CUTs (PERCENT FAILURE RATE)

| CWT | CURRICULUM AREA | 84-06 | 84-07 | 84-08 | 84-09 | 84-10 | 84-11 | 84-12 | 84-13 | 84-14 | 84-15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CS 2,3 | 6 | 5 | 7 | 8 | 8 | 11 | 7 | 2 | 5 | 7 |
| | LM 1A,2B,2R,4A,4E | 5 | (18) | 6 | 5 | 11 | 7 | 6 | 2 | 8 | 13 |
| | PK 1,2 | 4 | 9 | 9 | 8 | 4 | 3 | 5 | 11 | 6 | 10 |
| | DS 1 | 13 | (18) | 10 | 8 | 5 | 13 | 8 | 9 | 4 | 12 |
| 2 | CS 2,4 | 7 | 5 | 7 | 6 | 2 | 5 | 6 | 5 | 4 | 2 |
| | LM 2C-H,5A,B | 2 | 3 | 6 | 3 | 9 | 8 | 2 | 8 | 12 | 8 |
| | PK 6 | 8 | 2 | 4 | 2 | 14 | 6 | 6 | 6 | 2 | 4 |
| | DS 1,2 | (18) | 14 | (29) | (27) | 3 | 6 | 6 | 3 | 3 | (18) |
| 3 | CS 2,5A | 5 | 6 | 6 | 11 | 6 | 6 | 7 | 5 | 5 | 5 |
| | LM 5C,5D,5F,5B | 13 | 11 | 4 | 5 | 2 | 1 | 15 | (17) | 2 | 7 |
| | PK 5A,8A1,9A | 3 | 0 | 1 | 2 | 4 | 1 | 7 | 4 | 3 | 0 |
| | DS 1C,3 | 10 | 14 | 8 | 6 | (22) | 8 | 5 | (23) | 9 | 15 |
| 4 | CS 2 | (20) | 11 | 6 | 11 | 6 | 7 | 14 | (19) | 14 | 14 |
| | LM 1E,3C | 8 | 8 | 9 | 17 | 14 | 9 | 6 | 5 | 13 | 12 |
| | PK 7,8 | 10 | 5 | 5 | 13 | 6 | 7 | 7 | 2 | 3 | 5 |
| | DS 1D,4 | 10 | 5 | 6 | 8 | 12 | 11 | 7 | 9 | 14 | 7 |
| 5 | CS 2 | 1 | 2 | 2 | 4 | 14 | (18) | 13 | 7 | 5 | 7 |
| | LM 6 | 1 | 5 | 7 | 5 | 6 | 3 | 1 | 2 | 3 | 6 |
| | PK 7C,5F,11,13,10 | 6 | 4 | 10 | 6 | 2 | 1 | 1 | 4 | 3 | 7 |
| | DS 1E,5,6 | 1 | 4 | 4 | 4 | 10 | 15 | 4 | 6 | 9 | 5 |

| CWT | CURRICULUM AREA | 85-01 | 85-02 | 85-03 | 85-04 | 85-05 | 85-06 | 85-07 | 85-08 | 85-09 | 85-10 | 85-11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CS 2,3 | 8 | 9 | 9 | 4 | 0 | 4 | 13 | 6 | 2 | 12 | 11 |
| | LM 1A,2B,4A,4E | (16) | 2 | 11 | 8 | 3 | 6 | 2 | 14 | 12 | 14 | (19) |
| | PK 1,2 | 7 | 3 | 5 | 6 | 2 | 4 | 7 | 2 | 6 | 5 | 9 |
| | DS 1 | 5 | 5 | 6 | 1 | 0 | 1 | 1 | 1 | 0 | 4 | 3 |
| 2 | CS 2,4 | 3 | 5 | 5 | 5 | 7 | 4 | 7 | 7 | 13 | 1 | 3 |
| | LM 2C-H,5A,B | 4 | 9 | 8 | 1 | 1 | 8 | 2 | 7 | 9 | 4 | 13 |
| | PK 6 | 1 | 5 | 8 | 5 | 4 | 1 | 1 | 5 | 5 | 6 | 5 |
| | DS 2 | 14 | 3 | 2 | 2 | 3 | 6 | 7 | 4 | 9 | 4 | 13 |
| 3 | CS 2,5A | 4 | 5 | 3 | 5 | 6 | 2 | 1 | 11 | 10 | (20) | 11 |
| | LM 5C,D,F,H | 11 | 2 | 2 | 1 | 4 | 4 | 9 | 4 | 5 | 3 | 0 |
| | PK 5A,9A | 1 | 2 | 3 | 2 | 2 | 1 | 4 | 2 | 1 | 3 | 11 |
| | DS 3 | (20) | 7 | 8 | 5 | 4 | 5 | 5 | 7 | 11 | 10 | 13 |
| 4 | CS 2 | 11 | 6 | 9 | (23) | 8 | (22) | (18) | 6 | 15 | 11 | 14 |
| | LM 1E,3C | (16) | 14 | 8 | 10 | 8 | 2 | 3 | 13 | 12 | 6 | (16) |
| | PK 7,8 | 7 | 2 | 5 | 2 | 7 | 5 | 6 | 2 | 2 | 2 | 5 |
| | DS 4 | 9 | 15 | 11 | 13 | 10 | 7 | (17) | 2 | 5 | (16) | (22) |
| 5 | CS 2 | 13 | 7 | 8 | 9 | 8 | 11 | 15 | 11 | (20) | 7 | 10 |
| | LM 6 | 1 | 2 | 5 | 1 | 2 | 2 | 1 | 4 | 1 | 1 | 4 |
| | PK 5E,10,11,13 | 8 | 1 | 5 | 2 | 2 | 2 | 4 | 2 | 4 | 6 | 6 |
| | DS 5 | 5 | 5 | 4 | 2 | 0 | 5 | 4 | 3 | 1 | 1 | 1 |

Atch 3

## REFERENCES


1. Anastasi, Anne. 1982. Psychological Testing. 5th Ed. New York: MacMillan.

2. Borsuth, John R. 1970. On the Theory of Achievement Test Items. Chicago: Univ of Chicago Press.

3. Buros, O.K. ed. 1978. The Eighth Mental Measurements Yearbook, Vol 1 and 2. Highland Park, NJ. The Gryphon Press

4. _____ ed. 1974. Tests in Print II. Highland Park, NJ. The Gryphon Press.

5. Cronbach, L.J. 1984. Essentials of Psychological Testing. 4th ed. New York: Harper and Row.

6. Edwards, Allen L. 1973. Statistical Methods. New York: Holt, Rinehart & Winston.

7. Green, J.A. 1975. Teacher-Made Tests. New York: Harper & Row.

8. Gronlund, Norman E. 1982. Constructing Achievement Tests. 3rd Ed. Englewood Cliffs, NJ, Prentice-Hall.

9. _____ 1985. Measurement and Evaluation in Teaching. New York: MacMillan.

10. Guilford, J.P. and Fruchter, B. 1977. Fundamental Statistics in Psychology and Education. 6th Ed. New York: McGraw-Hill.

11. Lyman, H.B. 1978. Test Scores and What They Mean. Englewood Cliffs,NJ: Prentice-Hall.

12. Mehrens, W.A. and Lehmann, I.J. 1984. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart & Winston.

13. Objectives & Tests. Vol III. AFP 50-58. Handbook for Designers of Instructional Systems, 15 Jul 1978. Department of the Air Force, Washington DC, USAF.

14. Randall, R.A. 1972. "Contrasting Norm-Referenced and Criterion-Referenced Measures." Paper presented at American Educational Research Organization,1972. Washington, DC.

15. Roid, G.H. and Haladyna, T.M. 1982. A Technology for Test-Item Writing. New York: Academic Press.

16. Sako, S. 1980. "Validation Studies of Written Achievement Tests used at USAF OTS." Proceedings of Military Testing Association 22nd Annual Conference.

17. Sako, S. 1983. OTS Measurements Development Handbook for Instructors. Lackland AFB, Texas.

18. Standards for Educational and Psychological Tests. 1985. Washington DC: American Psychological Association.

19. Student Measurements Administration Control and Special Individualized Assistance. 1984. Officer Training School Regulation 50-11. Lackland AFB,Texas.

20. Principles and Techniques of Instruction. 1984. Air Force Manual 50-62. Department of the Air Force.

# Utility Estimation in Five Enlisted Occupations

Newell K. Eaton, Hilda Wing and Alan Lau

U.S. Army Research Institute for the Behavioral and Social Sciences[1]

In most organizations the decision to develop and implement selection and/or classification tests rests on the assumption that their costs will be outweighed by their benefits in terms of increased employee performance and tenure. The initial costs of testing programs have been increasing due to more stringent requirements for documentation of validities, test administration using computers, and the potential for legal challenges to test fairness. With the increasing costs of starting and maintaining testing programs, more attention is being paid to assessing their benefits. The purpose of this paper is to expand on methods used by several researchers in this area (Eaton, Wing, & Mitchell, 1985, Hunter & Schmidt, 1982).

Brogden (1949) and Cronbach and Gleser (1965) provided the first systematic descriptions of the utility of testing programs indexed in dollars. They linked performance levels to the dollar values estimated for those performance levels. Their formula for the gain in productivity, or utility (U$), obtained by using valid selection procedures includes (a) Ns, the number of individuals selected, (b) SD$, the standard deviation of performance, scaled in a utility metric such as dollars, and (c) the average performance expected on the criterion by the selected group as estimated from a valid predictor, given by Rxy Zx:

$$U\$ = Ns\ SD\$\ Rxy\ Zx$$

The formula was subsequently modified to account for testing costs. A more complete description of such formulations can be found in Cascio (1982), Cronbach and Gleser (1965), and Hunter and Schmidt (1982).

While the values of most of the variables on the right hand side of the Brogden-Cronbach-Gleser formulas are known, the estimation of SD$, the standard deviation of performance scaled in dollars, is problematic. One 'SD$ Estimation Technique' is based on estimates of the dollar value to the organization of performance at the 50th percentile level, the 85th percentile level (one standard deviation above the mean), and, sometimes, the 15th percentile level (one standard deviation below the mean). The dollar difference between the 15% and 50% estimates, and the 50% and 85% estimates, provides an estimate of SD$ (Cascio & Silbey, 1979; Hunter & Schmidt, 1982, and Schmidt, Hunter, McKenzie, & Muldrow, 1979).

---

[1]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.

A second method is the 'Superior Equivalents Technique' proposed by Eaton et al.(1985). It is somewhat like the SD$ Estimation Technique. Instead of using estimates of the dollar value of 85th percentile performance, however, the technique uses estimates of the number (N85) of superior (85th percentile) performers who would be needed to produce the output of a fixed number (N50) of average (50th percentile) performers. This estimate, combined with an estimate of the dollar value (V50) of average performance, provides an estimate of SD$:

$$SD\$ = V50\ [(N50/N85) - 1].$$

Eaton et al. speculated that this method would be more appropriate in situations where the nature of the work is such that managers are more accustomed to considering the relative productivity of employees or crews than the relative costs of producing given levels of output.

A third estimation strategy has been proposed by Hunter and Schmidt (1982). In reviewing the results of a variety of studies, they note that SD$ typically falls between 40% and 70% of annual salary. This might be termed the 'Salary Percentage Technique.'

In their recent paper, Eaton et al. showed that the Superior Equivalents Technique provided more stable estimates of U.S. Army tank commanders' SD$ than did the SD$ Estimation Technique. They also noted that both these techniques provided substantially larger estimates of SD$ than did the Salary Percentage Technique. The purpose of this paper was to compare the results of the Superior Equivalents Technique with the SD$ Estimation Technique across five different U.S. Army enlisted military occupational specialties (MOS). This was intended to assess both the variability of SD$ values across the five MOS as well as the results with the two techniques. The paper was also intended to determine whether a 'short hand' estimation procedure could be developed for military occupations, such as the Salary Percentage Technique. Last, because the research was conducted with supervisors who were both noncommissioned officers (NCOs) and commissioned officers, it was possible to assess the impact of level of management on SD$ estimates.

## METHOD

### Instrument

A questionnaire based on earlier research (Bobko et al. 1983, Burke & Frederick, 1984; Eaton et al. 1985, Schmidt et al. 1979) was developed to measure the comparative worth to the Army of first-term soldiers operating at different performance levels. Separate forms were administered to supervisors in each of the five MOS studied. The first method asked supervisors to think about how much ten average soldiers (50th percentile) contributed to the Army. Supervisors then estimated how many superior (85th percentile) soldiers would be needed to do the same amount of work. The second method asked supervisors to first consider the worth of average and superior first tour soldiers to the Army. They were then asked to estimate how much an average (50th percentile) first-term soldier and a superior (85th percentile) soldier

are worth by considering such factors as salary, output, responsibility, and equipment. Dollar estimates of the yearly value to the Army of average and superior soldiers were then requested.

Subjects

Supervisory estimates were obtained from 270 NCOs and officers in five MOS across three different posts. The five MOS were infantrymen (11B), armor crewmen (19B), light wheel vehicle/power mechanics (63B), medical specialists (91B), and radio teletype operators (05C). Of the 270 supervisors, 226 (83 percent) were NCO and 29 (11 percent) were officers. The remainder did not provide rank information. Four supervisors (one percent) did not respond to the methods of estimating utility and their responses are not included in the analyses. Of the remaining 266, 13 did not provide useable estimates for the first method and (a different) eight did not provide useable estimates for the second method.

Other Data

To obtain the value of average performance for the Superior Equivalents Technique, as well as the data required for the Salary Percentage Technique, we used published pay and allowance tables. In 1985 the base pay for Army enlisted personnel with two years of service ranged from $9,000 to $10,000. Non-taxable allowances for such items as housing, post exchange, vacation, and travel benefits could amount to more than $6,000 for the typical married soldier with dependents. Our estimate of an equivalent civilian salary would be about $16,000 per year. This is consistent with Henderson's (1985) estimates for the compensation of a Private First Class living off post with dependents.

## RESULTS

The results from the Superior Equivalents Techniques indicated that, across MOS, 5.20 superior first-tour soldiers performed as well as 10 average soldiers. Using $16,000 as the value of average performance (V50), 5.20 as the number of superior equivalents (N85), and 10 as the number of average soldiers (N50), the Superior Equivalents Technique yielded a SD$ estimate of $14,769. Of the 253 supervisors responding, 7% indicated 1 or 2 superior first-tour enlisted soldiers were equivalent to 10 average soldiers, 23% indicated 3 or 4, 51% indicated 5 or 6, 17% indicated 7 or 8, and 3% responded with 9 or 10. There was only a modest difference between estimates for the five MOS. the number of superior equivalents ranged from 4.90 to 5.58 with SD$ estimates from $12,681 to $16,720. The results by MOS are shown in Table 1. Full ANOVA results were computed, including as factors MOS and RANK of the supervisor providing the estimates. The differences by MOS did not reach statistical significance, nor did RANK, nor the MOS x RANK interaction.

The results from the SD$ Estimation Technique indicated that, across MOS, average soldiers were worth about $16,725 per year while superior soldiers were worth about $25,969. This yields an SD$ estimation of $9,244. Of the 258 supervisors responding, 11% provided SD$

Table 1· Estimated Number of Superior First Tour Soldiers Equaling 10 Average Soldiers and Computed SD$ by MOS

| MOS | N | Number Superior | SD$ |
|---|---|---|---|
| 11B | 48 | 5.54 | $12,881 |
| 19E | 60 | 5.40 | $13,630 |
| 91B | 36 | 4.89 | $16,720 |
| 63B | 67 | 5.15 | $15,068 |
| 05C | 42 | 4.90 | $16,653 |
| Totals | 253 | 5.20 | $14,769 |

estimates of less than $2,000, 14% between $2,001 and $4,000, 19% between $4,001 and $6,000, 12% between $6,001 and $8,000, 16% between $8,001 and $10,000 15% between $10,001 and $16,000, and 12% over $16,000. These appear to be more variable than Superior Equivalents estimates. Larger, between MOS differences also were found with the SD$ estimation technique, ranging from about $6,254 to $11,150. The average values assigned average and superior soldiers, as well as SD$ estimates for the five MOS, are shown in Table 2.

Table 2: Dollar Estimates of Value to the Army of Average and Superior First Tour Soldiers by MOS

| MOS | N | Average | Superior | SD$ |
|---|---|---|---|---|
| 11B | 53 | $19,226 | $29,000 | $ 9,774 |
| 19E | 63 | 13,736 | 20,190 | 6,254 |
| 91B | 38 | 18,000 | 27,132 | 9,132 |
| 63B | 64 | 15,719 | 26,344 | 10,625 |
| 05C | 40 | 18,200 | 29,350 | 11,150 |
| Totals | 258 | $16,725 | $25,969 | $ 9,244 |

Full ANOVA results were computed on the SD$ estimates following the procedures outlined for the Superior Equivalents estimates. With the SD$ estimates, however, the effect of MOS was significant ($F = 4.23$, $df = 4,225$, $p < .01$). Duncan's multiple range tests indicated the SD$ estimates for first tour armor crewmen (19E) were lower than those for medics (91B) mechanics (63B) and radio telephone operators (05C). Neither the RANK nor MOS x RANK effects were significant.

Last, SD$ values obtained using both the Superior Equivalents and SD$ Estimation Techniques were compared to the estimated civilian equivalent salary and to base pay. Using $16,000 as the best estimate of estimated civilian equivalent salary and $9,500 as base pay, estimates of SD$ would be 58-92% of estimated civilian equivalent salary based on superior equivalents and SD$ estimates, respectively. Using only base pay as a salary basis, SD$ would be estimated at 97%-156%. The value of 125% of base pay may be chosen as an estimate of SD$. Assuming a value of $10,000 per year as base pay (for simplicity, rather than the $9,500 figure used in previous analyses), then SD$ = $12,500 and U$ can be estimated (Cascio, 1982, pp 220-226). Table 3 displays the estimated U$, per first tour soldier selected, as a function of the validity of the test and the proportion of applicants selected.

Table 3: Estimated US Per Selection as a Function of Test Validity and Proportion of Applicants Selected

| | | Test Validity | | | | | |
|---|---|---|---|---|---|---|---|
| | | .1 | .2 | .3 | .4 | .5 | .6 |
| Proportion | .2 | $1,750 | $3,500 | $5,250 | $7,000 | $8,750 | $10,500 |
| of | .4 | 1,200 | 2,400 | 3,600 | 4,800 | 6,000 | 7,200 |
| Applicants | .6 | 813 | 1,625 | 2,438 | 3,250 | 4,063 | 4,875 |
| Selected | .8 | 413 | 825 | 1,238 | 1,650 | 2,063 | 2,475 |

If 100 soldiers were selected from among 125 applicants, using a test with a validity of .3, the estimated utility would be 100 x $1,238 = $123,800 per year.

## DISCUSSION

The first purpose of this research was to assess the SD$ of performance in five Army enlisted military occupational specialties using two methods. For both methods there were numerical differences in SD$ across the MOS, and they were ordered logically. The lowest SD$ values were obtained for team/crew occupations - infantryman and tank crewman - while the highest SD$ values were obtained for those who perform many duties as individuals - medics, mechanics, and radio/telephone operators. However, between MOS differences were statistically significant for SD$ values obtained for only one method, the SD$ Estimation Technique, and these differences were not clear cut.

The Superior Equivalents Technique, designed for use in military settings, did not provide reliable between-MOS differences. It did, however, yield estimates with considerably smaller levels of between-subjects dispersion. This is consistent with the results of the earlier Eaton et al. research. On balance, it would seem that both techniques provide SD$ estimates which yield a useful range in which the 'real' SD$ probably falls. But neither is sufficiently precise at this time to provide between-MOS comparisons in which one can be confident.

Obtaining a ball-park estimate of SD$ may well be sufficient for most purposes. Seldom does one face a decision where the utilization of a selection or classification test rests on cost tradeoffs of plus or minus 10% or 20% of testing and start up costs. Rather, such programs are more typically initiated only if the potential payoff is several times the costs. As a consequence, estimating a reasonable range of SD$ values can be quite useful.

Fortunately, this and prior research (Eaton et al. 1985) show that such an estimate may be obtained using a variant of the Hunter & Schmidt (1982) Salary Percentage Technique. In the Eaton et al. work, SD$ was 89% of estimated civilian equivalent salary, and 178% of base pay. For the two methods compared in this research, results ranged from 58%-92% of civilian equivalent salary, and 97%-156% of base pay. Given this consistency it would seem that a rough estimate of SD$ for first-tour enlisted personnel is about 125% of base pay.

Such an estimate is likely to be quite conservative. Eaton et al. found SD$ values obtained with the two methods used in this research to be about half those obtained with yet a fourth method, the System Effectiveness Technique, designed to incorporate equipment, maintenance, and other support costs. Burke and Frederick (1984) and Schmidt et al. (1982) also obtained results suggesting the conservative nature of SD$ values obtained with the SD$ Estimation Technique. The use of such a rough estimate may well make a useful contribution to front end analyses designed to assess the potential utility of initiating research on, or implementation of, a selection and classification testing program. Table 3 provides figures which make such estimates relatively simple.

## REFERENCES


Bobko, P., Karren R., & Parkington, J.J. (1983). The estimation of standard deviations in utility analyses: An empirical test. Journal of Applied Psychology, 68, 170-176.

Brogden, H.E. (1949). When testing pays off. Personnel Psychology. 2, 171-183.

Burke, M.J., & Frederick, J.T. (1984). Two modified procedures for estimating standard deviations in utility analyses. Journal of Applied Psychology, 69, 482-489.

Cascio, W.F. (1982). Costing human resources The financial impact of behavior in organizations. Boston: Kent Publishing Co.

Cascio, W.F., & Silbey, V. (1979). Utility of the assessment center as a selection device. Journal of Applied Psychology, 64, 107-118.

Cronbach, L.J., & Gleser, G.C. (1965). Psychological tests and personnel decisions (2nd ed.), Urbana: University of Illinois Press.

Eaton, N.K., Wing, H., & Mitchell K.J. (1985). Alternate methods of estimating the dollar value of performance. Personnel Psychology. 38, 27-40.

Henderson, W.D. (1985). Cohesion. The human element in combat. Washington, D.C., National Defense University Press.

Hunter, J.E., & Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In Fleishman E.A., Dunnette M.D. (Eds.), Human performance and productivity· Volume I. Human capability assessment. Hillsdale, NJ,: Erlbaum.

Schmidt, F.L., Hunter, J.E., Mckenzie, R., & Muldrow, T. (1979). The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 64, 609-626.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work force productivity. Personnel Psychology. 35, 333-347.

# RELATION OF MENTAL AND EDUCATION LEVELS TO NAVY ENLISTED PERFORMANCE

Charles H Cory[1]

Navy Personnel Research and Development Center

As part of a joint services effort, the Navy Personnel Research and Development Center has been investigating the validation of enlisted personnel selection standards against job performance as well as the training school achievement criteria presently used Personnel-record based job outcome measures are one type of criterion being investigated. This paper summarizes the findings from two studies which measured the association of mental and education levels of Navy enlisted personnel to their performance on job outcome criteria

The first study, conducted on the CY76 male enlisted cohort, investigated the relationship of education (ED) and mental level (ML) to 6 job outcome criteria--2 advancement, 2 attrition/retention, and 2 maladaptive behavior criteria. Relationships were examined within the three entry pay grades in the cohort E-1, E-2, and E-3.

The approximately 71,000 male, non-prior service personnel accessioned during CY76 who were 4-year enlistees and had complete records on the variables of interest constituted the sample The predictors were (1) Education Category coded with three values. Non-high school Graduate=1, GED Certificate=2, and High School Diploma Graduate=3 and (2) mental level category(ML), coded with five values 93-99th centile(ML 1)=1, 65-92nd centile(ML 2)=2, 49-64th centile(ML High-3)=3, 31-48th centile(ML Low-3)=4, and 10-30th centile (ML 4)=5

The two advancement criteria were. (1) percent of time spent at E-4 or higher(PTE4+) and (2) achievement/non-achievement of E-4 during the 4-year enlistment(E4/Non-E4) The attrition/retention criteria were (1) completion/non-completion of enlistment and (2) attrition/non-attrition within one year of accession The maladaptive behavior criteria were (1) number of demotions and (2) overall behavior record, a composite score formed from the number of unauthorized absences + twice the sum of the demotions and desertions

## Validities of ED and ML across Ratings

Figure 1 shows the validity coefficients of ED and ML for the two advancement criteria. The four columns for each predictor represent the validity coefficients for the subsamples accessioned as E-1s, E-2s, E-3s, and the total sample, respectively. Because the scaling of ED and ML are arbitrary and the focus of the research is to compare their magnitudes, the absolute values of the validity coefficients are shown. In other words, for this figure, the negative coefficients for ML are presented as positive values.

---

[1]The opinions expressed in this paper are those of the author, are not official, and do not necessarily reflect the views of the Navy Department

The top part of the figure shows that both education and mental level had substantial effects on percent of time spent at E-4 or higher during the first enlistment. Validity coefficients for ED range from .21 to 11 and the validity of ED overall, without subgrouping by entry pay grade, was 24. Validity coefficients for mental level ranged from 10 to 25 and the overall coefficient was 27. The predictiveness of ED and ML for E4/Non-E4 was similar to but slightly lower than for percent of time at E4+.

For attrition/retention criteria (Figure 2) the predictiveness of ED was from two to three times that of ML, depending on the criterion and the entry pay grade group. For Completed the 4-year Enlistment, validity coefficients for ED ranged from .20 to .06 compared with a range for the ML coefficients of .01 to .03 For Attrited during the First Year, validity coefficients for ED ranged from .14 to .07, compared with a range of .03 to .06 for ML.

ED was also more predictive for Maladaptive Behavior criteria (Figure 3) than was ML For Overall Behavior, validity coefficients of ED ranged from .18 to 08. Those for ML ranged from 01 to 02 Coefficients of ED for Number of Demotions ranged from .09 to 03, those for ML ranged from 01 to .02

Validities of ED and Mental Ability Within Ratings

The predictiveness of education and mental ability variables for job outcome criteria also were studied within ratings The sample selected for this research consisted of personnel in three representative ratings Sonar Technician (Surface), Hospital Corpsman, and Machinist Mate. The samples were taken from a data base composed of longitudinal records which covered the 4-year enlistment outcomes of all first-term non-prior service male personnel who were accessioned during FYs 77, 78, and first quarter 79

The data base contained complete rating and pay grade records of personnel during the 4-year period, or that part of it during which they served in their first enlistment, and a calculation of total basic pay, based on time in each pay grade. Samples were selected to include all 4-year non-prior service male enlistees who at some time in their four years had been either (1) assigned to the rating, (2) assigned to formal school training for the rating or (3) promoted to E-4 or higher in the rating. Sample sizes were 956 Sonar Technicians, 2,278 Hospital Corpsmen, and 3,177 Machinists Mates

For this study the only criterion used was time at E4 or higher, the criterion which had been found to be the most promising in previous research This was measured as Months at the Full Performance Level (MFPL), i.e., time spent at E-4 or higher

Figures 4 and 5 show the predictiveness for MFPL of the selector composite which is used for classification into the rating and the three variables which define Navy requirements for enlistment These are AFQT score, Education level categorization (ED), and SCREEN score. SCREEN, an acronym standing for Success Chances of Recruits Entering the Navy, is a composite score, scaled from 48 to 96, which is based on AFQT, education, age and marital status Personnel with scores of 70 or higher are eligible for Navy enlistment SCREEN was developed to exclude from Navy enlistment personnel who are high attrition risks.

Figure 4 shows, for personnel accessioned as E-1s, the predictive validity of AFQT, ED, SCREEN, and the selector composites for the three ratings. Sonar Technician has the highest coefficients overall, and Hospital Corpsman and Machinist Mate follow in that order Validity coefficients for STGs range from .10 to .27, with similar ranges for the other two ratings All 12 validity coefficients are statistically significant. For the three ratings SCREEN was the best predictor of MFPL and ED was a better predictor than AFQT.

Figure 5 shows for personnel accessioned as E-3s, the same type of predictive information shown previously for personnel accessioned as E-1s. These coefficients are much lower than those for E-1s Validity coefficients for Hospital Corpsman, the most predictable rating, range from .11 to .17 and those for Machinists Mate, the least predictable rating, range from .01 to .06. Only 7 of the 12 coefficients in the slide are statistically significant.

Comparisons of this type are not being presented for E-2s because E-2 accessions are much less common than E-1 and E-3 accessions. E-2s constitute only about 7% of enlisted accessions, compared with 74% accessioned as E-1s and 19% accessioned as E-3s Therefore, the sample sizes for E-2s are much smaller than those for the other two levels, and the predictive relationships are more irregular

Relative Utility of Personnel Accessioned as E-1s, E-2s, and E-3s

The same data base was also used for calculations to develop a dollar criterion of utility Because the utility relationships were similar for the three ratings, only those for Machinist Mate are presented Figure 6 shows the relative performance of personnel accessioned as E-1s, E-2s, and E-3s in terms of mean months at full performance level, mean total basic pay cost per month at FPL, and first term attrition rate All four variables have been scaled in multiples of the lowest value Thus, for the leftmost variable, the mean MFPL of E-1s is scaled as 1 0 that for E-2 is roughly 25% greater, etc.

Mean months at full performance level for E-3 accessions were more than twice those for E-1s and about 80% greater than those for E-2s. In contrast, the mean cost of E-3s per month at FPL was about 35% that of E-1s. The reasons for this large discrepancy in mean basic pay cost per month are shown in the next two categories. Mean total basic pay costs for E-3 accessions for the entire enlistment were about 4/5ths those of E-1s and E-2s Since E-3s are paid more than E-1s and E-2s and they spend more time at E-4+, one reason for their surprisingly small mean total basic pay is shown in the next category to the right E-3 accessions attrited at nearly twice the rates of E-1s and E-2s. Attrition rates of 28 0%, 24.9%, and 47 3% were recorded for E-1, E-2, and E-3 accessions, respectively

Statistical Corrections to Improve Estimates of Validity

Although these analyses are preliminary, one gratifying result from them is that they show that selection and classification variables in the Navy are associated with important aspects of enlisted performance Moreover, it is clear that the predictivenesses of ED and mental ability for job outcome criteria are even greater than is shown in the preceding figures for instance, inspection of

the AFQT means and standard deviations for the samples used for the research showed that restrictions in range had occurred. Correction of the validity coefficients for these restrictions in range will produce a more accurate measure of the predictive relationships; it will also show them to be larger than the coefficients presented in Figures 1 through 5.

Also, because the shapes of the distributions of both ED and ML differ substantially from normality, the actual maximum validity coefficient is substantially less than the theoretical maximum, 1 The departure from normality was the most pronounced for the ED distribution. 19% Non-HS Grad, 7% GED, and 74% HS Grad--an asymetrical U-shaped distribution.

Estimates of the effects of these departures from normality on the effective maximum validity coefficients for the predictors in the CY76 conort were made. For this step a method recommended by Carroll was used to calculate the effective maximum correlation coefficient, given the non-normal characteristics of the ED and ML distributions for two criteria Percent of time spent at E4+ and Completion/Non-completion of Enlistment. The maximum coefficients for ML for PTE4+ and Completion/Noncompletion were .91 and .57, respectively, and for ED, .42 and 64.

This substantial shortfall of the maximum rs from the theoretical maximum, 1 00, indicates that both ED and ML are more strongly related to the performance criteria than the absolute magnitudes of their validity coefficients indicate This is particularly true for ED-PTE4+ ED accounts for 33% of the explainable variance of PTE4+, rather than 6% of the variance as is indicated by the absolute magnitude of the computed coefficient

Summary and Conclusions

These research findings support the following conclusions. (1) The education level, mental ability and SCREEN variables which are used for enlisted selection in the Navy are significantly predictive for important job outcome performance criteria, both within and across ratings (2) ED and mental ability are about equally predictive for advancement criteria, but ED is much more predictive than mental ability for attrition/retention and maladaptive behavior criteria. (3) Selection and classification variables are more predictive of job outcome variables for E-1 accessions than for E-3 accessions. (4) Adjustment of the data to compensate for the distortions caused by departures from normality of the distributions of the predictors and restriction in range can be expected to show that the true predictiveness of these variables for job outcome criteria is even greater than the substantial coefficients found in the present research (5) The practice of accessioning qualified personnel at E-2 and E-3 pay grades is a very cost effective way of providing to the Navy enlisted personnel who can perform effectively at the full performance level.

This report describes preliminary findings from research designed to develop a methodology for calculating a composite job outcome measure which expresses an individual's overall value to the Navy The research will continue during FY86 and beyond

## VALIDITY COEFFICIENTS OF ED AND ML FOR ADVANCEMENT CRITERIA

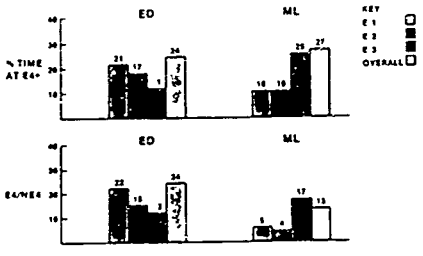



FIGURE 1

## VALIDITY COEFFICIENTS OF ED AND ML FOR ATTRITION/RETENTION CRITERIA

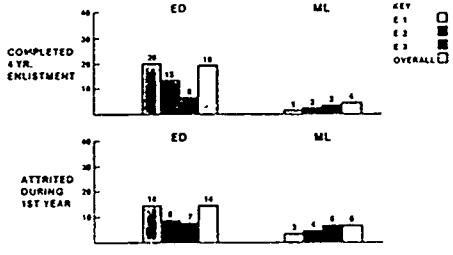



FIGURE 2

## VALIDITY COEFFICIENTS OF ED AND ML FOR MALADAPTIVE BEHAVIOR CRITERIA

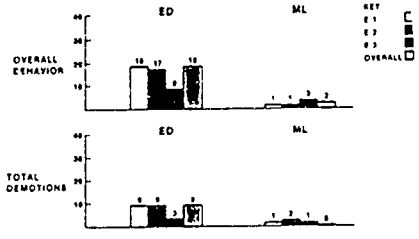



FIGURE 3

## PREDICTIVE VALIDITY OF AFQT, LEVEL OF EDUCATION, SCREEN, AND SELECTOR COMPOSITE FOR MFPL (FOR E-1 ACCESSIONS)

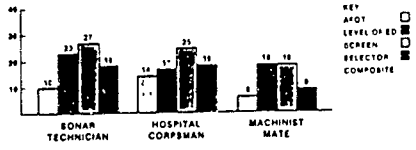



FIGURE 4

## PREDICTIVE VALIDITY OF AFQT, LEVEL OF EDUCATION, SCREEN, AND SELECTOR COMPOSITE FOR MFPL (FOR E-3 ACCESSIONS)

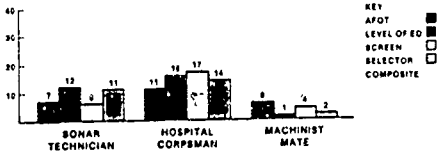



FIGURE 5

## ATTRITION AND ADVANCEMENT CHARACTERISTICS of E-1, E-2, and E-3 MACHINIST MATE ACCESSIONS

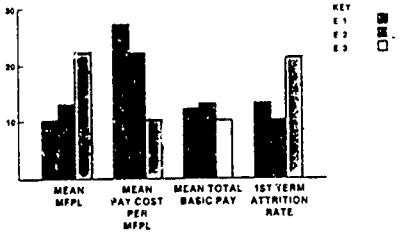



FIGURE 6

# SELECTING CRITICAL TASKS FOR A RADIOMAN HANDS-ON PERFORMANCE TEST

Herbert George Baker, PhD
and
Gerald. J. Laabs, PhD

Navy Personnel Research and Development Center
San Diego, California 92152-6800

Selecting an array of critical tasks is one of the most important steps in developing a hands-on job sample test. This paper describes the procedures used in critical task selection for the Radioman (RM) performance test and for surrogate instruments.

## Background

In cooperation with the Joint Service Job Performance Measurement/Enlistment Standards Project, the Navy is exploring relationships between predictors and various measures of job performance. The main objective is to investigate measurement approaches that might be used to make personnel classification more performance-based. The outcomes of the research may allow a recruit's potential for successful on-the-job performance to be considered more fully in the personnel accessioning process, perhaps through the addition of an additional algorithm component in the automated classification and assignment system.

The observation of job performance would seem to be the ultimate basis upon which to validate tests used to screen military applicants. Thus, the strategy of the Joint-Service Project is to construct hands-on performance measures and investigate the use of these measures as criteria for predictor validation, with a major research focus on the development of job sample tests of technical proficiency.

The institution of a service-wide program to obtain performance criteria via job sample testing would be prohibitively expensive. Therefore, a concurrent research focus of the Joint-Service Project is the development of economical substitutes, or surrogate tests. Costs will be reduced by using hands-on performance tests as benchmark performance measures and developing substitute measures that are less expensive and easier to administer, such as paper-and-pencil or computer simulations of job samples.

## Radioman Performance Test Development

The Radioman rating is one of the jobs selected by the Navy for performance test development because it is critical to mission success, has a large population -- including substantial numbers of women and ethnic minorities -- and is similar to radioman jobs in the other armed services. In accordance with the Joint-Service research strategy a job sample test is being designed and developed, along with a job sample simulation and a set of rating scales. This work is being performed under contract by the Human Resources Research Organization (HumRRO) and the Personnel Decisions Research Institute (PDRI).

## Critical Task Selection

For the job sample test under construction to be valid, it must adequately represent the important tasks done by first-term radiomen. In fact, the validity of the entire test package -- the hands-on test, simulation, and rating scales -- depends on the method used to identify and select tasks to be tested. Critical task identification and selection is the first and perhaps most important step in the test development process.

The Navy and its contractors have completed this first step, gathering information on critical tasks from subject matter experts (SME) who are senior enlisted personnel in the RM rating. As an additional check on the data gathered from RMs working in the fleet, a Quality Control Review Panel (QCRP) composed of additional SMEs was formed. This was done by soliciting nominations for panel membership from headquarters, school, and operational Navy commands that have cognizance over the RM rating. This panel will provide expert review of the work at each step in the process, assuring the highest quality end-products. In addition, it will establish an audit trail, facilitating acceptance of the final test package by the commands whose representatives are on the panel.

Out of the plethora of tasks done by RMs, a manageable subset must be selected as candidates for the development of hands-on test items. These items will be the critical job tasks that can feasibly be measured in the hands-on mode, and in turn will guide the development of substitute, or surrogate, test items. To guide critical task identification and selection, we have adopted Guion's (1979) paradigm for reducing the job to a job sample (Figure 1). His four major steps include determining the: (1) job content universe; (2) job content domain; (3) test content universe; and (4) test content domain.

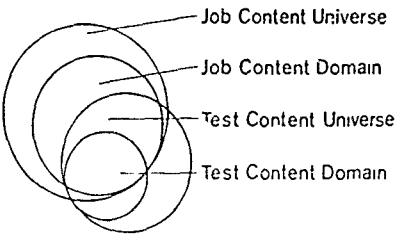



Figure 1. Venn diagrams relating job content to test content (adapted from Guion, 1979)

### Job Content Universe

The RM job content universe was defined by a comprehensive job task analysis. Two sources of information were used: Data from the Navy Occupational Task Analysis Program (NOTAP), which is used to develop the RM occupational standards; and the Job Task Inventory (JTI), which is used in developing the entry level training (i.e., A-School) RM curriculum. The two task analyses overlap considerably and contain 500-600 different job task statements.

## Job Content Domain

Next, the job content domain for first-term RMs was specified. The original list of task statements was refined and reduced through two workshops that included SMEs who were supervisors of first-term RMs. Both ship and shore installations were represented. Seven senior Navy enlisted personnel and one civilian attended the first workshop which was held in San Diego. Then, the shortened list was further refined at a second workshop held at Norfolk, in which seven senior Navy enlisted personnel participated.

To ensure that all task statements were written at the same level of generality, subtasks were accumulated into more comprehensive tasks, while broader, more general tasks were separated into their constituent parts. The Joint-Service research strategy addresses technical job proficiency only; that is, it specifically excludes purely military or routine administrative tasks, as well as other job performance considerations such as team tasks, and motivational, situational, or stress factors. The workshop panels deleted tasks falling into those categories, as well as tasks not within the normal purview of first-term RMs.

A list containing 125 task statements resulted. These statements were evaluated by the review panel, who gave approval to a final list that included 127 task statements. These tasks comprise the technical proficiency domain of first-term RMs.

## Test Content Universe

The next step in the procedure was to determine the test content universe. Put simply, this universe potentially embraces all those tasks that could possibly be included in a hands-on test, plus those elements necessitated by the testing situation. The testing situation includes such things as those conditions that are imposed to achieve relatively standardized testing and the procedures used to observe and record responses. For the RM rating, the test content universe is essentially the same as the job content domain, except for the addition of the required testing conditions, and a way must be found to reduce the size of the test content universe to manageable proportions. That is, some priorities must be established to guide task selection.

One of the typical methods to set task selection priorities is to gather information on the criticality of performing each task correctly. Consequently, the Navy launched an extensive survey effort to gather judgments from personnel in the rating, for use in setting task selection priorities. The 127 task statements were incorporated in a questionnaire, the Radioman (RM) Survey Form, designed to isolate the critical first-term RM job tasks.

The review panel addressed four proposed judgment scales:

1. frequency of task performance by first-termers,
2. difficulty or complicatedness of the task,
3. importance of the task to mission accomplishment, and
4. frequency of errors on performing this task by first-termers.

As a result of the review, the scales were changed to reflect absolute rather than relative judgments, with the review panel providing the scale values or anchors. The questionnaire was developed in two forms: one for first-term RM personnel (defined as RMs with four or less years of active duty), and one for supervisors (those RMs with between four and ten years of active duty). First-term RMs judged the frequency and difficulty of the tasks; supervisors judged the task importance and error frequency. The RM Survey was pilot tested on board a ship, using a small sample of RMs and supervisors.

The review panel also assisted in developing the sampling plan for distribution of the RM Survey Forms. Panel consensus was that the jobs of first-term RMs probably differ depending primarily on whether the job is situated in a large or a small communications facility. Other differences might accrue dependent on the job site: that is, ship or shore installation. Therefore, hull types and shore installation types were both dichotomized by the panel into large and small categories.

Using the Enlisted Master Tape, the total number of RMs (both first-term and supervisor) assigned to each Navy facility categorized by the review panel was determined. Population totals in each of the four cells made by crossing facility size by sea/shore location was used to determine the number of questionnaires to be mailed.

Because the data to be collected that had the most variability consist of the proportions of personnel performing tasks, the simple random sampling formula for binomial data (with correction for finite populations) was applied to each cell. This ensured an adequate sample for each of the four potential job subtypes, should separate tests later prove necessary; that is, if a common core test can not be achieved.

In view of the fact that each cell includes a number of different types of ship or shore installations, the sample was drawn proportionally within each cell according to installation type. The PQ split used in the sampling formula was worst case (i.e., .5), the level of tolerance was .10, and the probability level set at .95. Thus:

$$N = \frac{N'}{1 + N'/\text{Population}}$$

The final questionnaire was sent to approximately 500 first-term RMs and 500 supervisors of RM personnel throughout the world. To facilitate timely return of the questionnaires, personnel assigned to ships that were deployed, on overseas cruises, or in overhaul status were eliminated. Excepting this restriction, sampling was random. The review panel asserted that there is no difference whatever in the RM job between overseas and CONUS locations. Therefore, we expect this exception to have no bias on sampling and consequently no effect on task selection.

The questionnaires were addressed to individuals by name, and delivered via the commanding officer. Completed questionnaires have been received and their information entered into a computertized data base. Figure 2 reflects total sent, returned, and usable questionnaires for each sampling matrix cell.

FIRST - TERM

| | Large | Small |
|---|---|---|
| Shore | 140<br>117<br>90 | 120<br>91<br>59 |
| Ship | 137<br>109<br>85 | 142<br>105<br>81 |

SUPERVISOR

| | Large | Small |
|---|---|---|
| Shore | 138<br>112<br>90 | 119<br>92<br>65 |
| Ship | 114<br>85<br>6 | 132<br>92<br>82 |

Figure 2. Questionnaires sent and returned[1]

Test Content Domain

While final analyses are not yet complete, preliminary data indicate that we obtained reliable judgments. PDRI researchers calculated intraclass correlation coefficients for the two major scales (frequency and criticality). These reliabilities were corrected using the Spearman-Brown formula. All were in the .90s. Similar calculations could not be done on the percentage performing or complicatedness scales because the absence of a rating for a task not performed or supervised could not be given a score. Thus, the frequency and criticality scales will be given emphasis in identifying and selecting the sample of tasks for item development, with less weight given to the scales pertaining to errors and complicatedness (which also showed more restricted variability across tasks).

To ensure that the job of the first-term RM is adequately represented in the final sample of tasks, functional job categories were superimposed on the set of 127 tasks. Final selection will be proportional to the number of tasks included in a category rather than from the entire list.

In order to derive the job categories a taxonomy of tasks was developed at a workshop comprised of 12 senior staff personnel drawn from the Navy's entry level RM training course in San Diego. Cards containing the 127 task statements were sorted into categories of tasks that are of the same type or that deal with the same type of equipment. The sort yielded the following categories: (1) Preparing and processing messages/Establishing communications; (2) Setting up equipment; (3) Maintaining equipment; and (4) Handling secure materials.

---

[1]Personal communication. Mr. S. Lammlein, PDRI, 23 October 1985.

The judgments made in the questionnaires will guide task selection. A successive hurdles technique will be employed to obtain the final job sample. First, the most critical tasks in each category will be selected. Then, tasks with low frequencies or low percent performing or supervising both overall and within each of the sampling cells will be replaced. Finally, tasks that have low values in percentage of performance errors or complicatedness will be replaced. Figure 3 summarizes the task selection procedure. Tasks will be subsumed within the four functional categories, with each category reflected in the final test composition by at least one task. Ultimately, the test content domain for the RM rating will be a sample of the 127 tasks, organized into functional categories.

1. Select tasks with overall most critical ratings

2. Replace any tasks with:
   a. low frequency of being performed
   b. low percentage of sampling populations performing
   c. low percentage of performance errors
   d. low complicatedness

3. Ensure proportional representation by category

Figure 3. Successive hurdles for final task selection

It is important to note that, at this point, none of the 127 tasks has been eliminated because it is unsuitable for hands-on testing. The complete list of tasks selected as a result of the survey will be presented to the review panel for its judgment as to suitability for testing in the hands-on mode. Unsuitable tasks, if any, will be replaced using the procedures used in developing the original sample. Subsequently, a hands-on test, computer-based simulation test, and set of rating scales will be developed.

## Summary

Because no hands-on tests exist for measuring on-the-job performance, it was necessary to design and develop them. In turn, this required selecting critical tasks for which hands-on test items will be developed. Critical task selection was accomplished by an approach that combines supervisor and job incumbent questionnaire response through survey administration with quality control checks exercised through experience-based expert judgment. The result is a meaningful and manageable subset of tasks that can be converted into performance test items. The procedures used to obtain the job final sample ensure that the tests will have the highest possible content validity. We expect the results of critical task selection effort to materially contribute to the success of the Navy job performance measurement program, and ultimately to enhanced selection, classification, and assignment methodology.

## Reference

Guion, R. M. (April 1979) Principles of work sample testing: III. Construction and evaluation of work sample tests (ARI TR-79-A10). Alexandria, VA: Army Research Institute for the Behavioral & Social Sciences.

# Cognitive Predictors of M1 Tank Gunner Performance

Barbara A. Black
U.S. Army Research Institute-Fort Knox

Potential weapon system capability and achieved capability differ in large measure as a function of crew performance. To maximize Armor system effectiveness, the U.S. Army is committed to optimally selecting and training tank crewmembers. Within the four-man tank crew, specific emphasis has been placed on identifying soldiers who possess the requisite aptitudes and abilities to become proficient Armor tank commanders (TCs) and gunners. Early identification of these high-ability soldiers can lead to improvement in overall tank crew performance and in the cost effectiveness of training programs.

The realization of this goal is contingent on possessing the capability to measure abilities that have been identified as critical to job success and then being able to demonstrate the relationship between those abilities and actual job performance. The first step in this process is the analysis of job requirements for each position under consideration; the second step is the development of ability measures or tests. The U.S. Army Research Institute has for several years conducted research in job analysis and the development of tests for the prediction of Armor crewmember performance. This research has involved the development of tests for predicting success in basic and advanced individual Armor training, as well as in operational units.

Both job sample and paper-and-pencil tank gunnery predictor tests have been evaluated by ARI over a period of about ten years. The results of a meta-analysis completed on 15 of these data sets indicate that job sample tests were, across studies, better predictors of performance by job incumbents than were paper-and-pencil tests (Black & Campbell, 1982). Drawbacks to job sample testing do exist; they are similar to those identified from the psychomotor testing programs of the 1940s and 1950s: cost, increased administration time, and equipment unreliability (Melton, 1947). However, the advent of microprocessors and the increasing availability of high fidelity simulators may remove or reduce several of the major concerns in the use of job sample tests, specifically, the requirement for special equipment, the need for continuous calibration, and the difficulties involved in unit-collocated testing facilities. Job samples tests developed for incorporation into on-line or forthcoming unit-located simulators may improve the cost effectiveness of testing, reduce testing time requirements, and eliminate the need for special equipment apart from the simulator itself.

In addition, certain demographic variables have been found to correlate with tank gunnery performance across numerous studies for the past few years. These findings characterize the successful tank crew as being commanded by (a) a noncommissioned officer (NCO) with more time in the TC position than other TCs, (b) a TC who has trained longer with the gunner with whom he fired (Eaton & Neff, 1978), and (c) a TC

who has a history of having qualified crews (Evans & Sauer, 1982). None of these findings is particularly unexpected, but unfortunately, none is useful in the early identification of high-performing TCs. Yet this information is valuable in terms of providing data on variables whose covariance with the predictor measure may obscure the relationship of interest.

Initial efforts to evaluate predictors of crewmember performance in tank firing, driving, and loading used paper-and-pencil tests because they are the most cost effective and least time-consuming approach to performance prediction. Greenstein and Hughes (1977) used Armor trainees and limited their effort to the use of paper-and-pencil tests in the psychological literature or in use by the Army at that time. For example, they included Lauer's (1952) tests of Visual Memory and Attention-to-Detail, as well as the Armed Forces Qualification Test (AFQT) and three composites from the Army Classification Battery (ACB): Combat Operations (CO), Field Artillery (FA), and Motor Maintenance (MM). Significant correlations were obtained between the paper-and-pencil tests and loading errors and driving performance. None of the 11 paper-and-pencil tests in the study predicted tank firing scores. However, in a study commonly referred to as the Gideon report, Wallace (1982) presented the results of the 1982 European Canadian Cup trophy competition. He correlated the AFQT scores of tank commanders on the American team with their crew's live-fire gunnery scores and obtained a coefficient of .739 ($p < .01$, $N = 13$). The correlation between AFQT for the gunner and firing score was not significant. This research has prompted considerable interest in the existence and strength of relationships between the mental abilities of crewmembers and successful tank crew performance.

Scribner (1984) cited by Phillips (1985) in his report on Career Management Field (CMF) 19, found significant relationships between gunner/TC AFQT and tank Table VIII firing scores for 1131 tank crews training in Europe. This effort represents the largest database from which AFQT relationships have been obtained for live-fire performance.

While previous research indicates that certain testing techniques hold promise for Armor crewmember performance and more information is now available concerning important intervening variables, the availability of appropriate and useful criteria against which to validate predictor tests has remained a problem. Criterion measures used in past research include scores form live-fire gunnery exercises, Multiple Integrated Laser Engagement System (MILES) exercises, supervisory ratings, peer ratings, Skill Qualification Tests (SQTs), tests administered during the course of normal Armor training. Efforts to explain the inconsistencies found in past research have brought to light many disadvantages associated with the current job performance criteria available in Armor, especially those associated with gunnery.

Scores obtained from live-fire gunnery exercises often provide data that are not comparable between units or even between tanks. It is conceivable that with a company of tanks firing over a period of

several days, the condition of the weather, tank equipment, and range equipment could change to such a degree that no tanks fire the same engagements. In addition, for any specific tank, changes in ammunition characteristics, equipment performance, and firing conditions may reduce the reliability or increase the error variance for within-tank performance measures. Thus, low reliability of the criterion measure may have been a large contributing factor to the relatively inconsistent findings of past research.

In addition, it should be pointed out that tank gunnery tables are collective exercises. Engaging targets and measuring the results of those behaviors in such values as "time to engage" or "proportion of hits" produces a crew-level evaluation or, in the case of a Table IX exercise, a platoon-level evaluation. The relative contributions of individual crewmembers are difficult to ferret out. In fact, it is not uncommon for unit commanders who are short on high-quality personnel to pair mature, experienced TCs with novice or ineffective gunners to ensure that the tank crew will be rated "qualified." On the other hand, very effective gunners may find themselves in crews with ineffective TCs and fail to qualify their tanks during annual gunnery, thus making it virtually impossible to use the results of tank table exercises to make statements about individual performance.

The present effort involves the evaluation of a battery of M1 gunner performance prediction tests. Specifically it is an evaluation of the relationship between AFQT and the individual performance of gunners, using ratings from their supervisors, hands-on tests, and the crew/collective criterion, Table VIII gunnery. By virtue of the human factors engineering of the M1 tank system, the M1 gunner has more responsibility for fire control than the gunner on any other U.S. tank system. This increased responsibility, combined with the knowledge that from the soldiers selected and assigned as gunners come the future tank commanders (TCs) of the Armor force, makes it important to optimize the selection and assignment of M1 gunners.

## METHOD

### Subjects

The subjects, 123 M1 tank gunners representing four battalions, were selected for participation based on supervisors' ratings. Ratings were completed by company commanders and a senior NCO chosen by the commander. The two raters from each company were instructed to reach a consensus rank order for gunners, based on each gunners' demonstrated ability in performing gunnery-related tasks and on their availability for testing. Raters were asked to disregard gunners' performance in such nongunnery areas as military courtesy. In addition, raters were asked to consider the gunners' performance apart from that of their respective tank commanders or crews; that is, to rate gunners high if they were proficient, even though their crews may not have qualified on the most recent gunnery. Eight gunners were selected for testing from

each company. The four rated most proficient in each company and the four rated least proficient were tested. One company was exempted from the testing because of prior commitments. Thus two groups were formed, highly rated gunners and gunners receiving low ratings.

ASVAB. Four subtests from a research version of the Armed Services Vocational Aptitude Battery (ASVAB) were used to obtain estimates of soldier scores on the AFQT. Due to time constraints imposed on the overall testing process, a scaled-down version of each subtest was used. Subtests were shortened by randomly selecting 50% of the questions for administration. The AFQT consists of a combined score obtained from the following ASVAB subtests: Numerical Operations (NO), Paragraph Comprehension (PC), Arithmetic Reasoning (AR), and Word Knowledge (WK).

Tracking Test. This hands-on test used a snakeboard, an M1 tank, and an M55 laser boresighted with the main gun. A specially built device was used to pulse the laser automatically once per second for periods of 60 seconds. Soldiers were instructed to use the gunner station power control handles to track the snakeboard 12 times, 6 times from left to right and 6 times form right to left. To determine the gunner's accuracy on each trial the test administrator counted the number of laser pulses that fell on the snake. Speed was determined by recording the location of the final pulse, thus indicating the distance tracked during the 60-second trial.

M1 Computer Panel Test. Soldiers were tested on three operations of the M1 ballistic computer by means of a microcomputer-controlled simulation of the M1's computer panel. The simulator used a screen digitizer or touch panel placed over the face of a 12-in. color monitor. The operations consisted of enter - check data (ECD), and run computer self-test (CST). The software for the M1 computer test was developed to provide the soldiers with 3 instructional trials for each type of operation followed by 10 scored or test trials. The number correct on each operation and the time required to complete each test trial were recorded.

Criterion measures. The criterion measures included the supervisor's ratings and Table VIII scoresheets from the soldiers' most recent gunnery. Each gunner's overall Table VIII percentage was recorded as were the totals for day and night. Subtotals were also computed within both day and night scores for the moving main gun engagements.

## Procedures

The validation effort was conducted double-blind; that is, neither the soldiers tested nor the test administrators were aware of the supervisors' ratings. The testing schedule allowed four soldiers to be tested in the morning and four in the afternoon. Sixteen days were required to test 123 gunners.

## RESULTS AND DISCUSSION

Analysis of the data was completed in three steps: a) descriptive analyses by battalion, b) correlation between job sample measures and AFQT, and c) a break-out of performance by mental category.

Means and standard deviations for AFQT were computed for each battalion. No significant differences were found. However, significant differences were found among battalions for performance on the job sample tests: tracking and M1 computer panel. For this reason raw scores were converted to z-scores for all subsequent analyses.

Significant correlations between job sample measures and AFQT are presented in Table 1. It is interesting to note that AFQT correlated with number of hits for the hands-on tracking task ($r = .360$, $p < .0001$), indicating that gunners with higher AFQTs had more hits; that is, they were more accurate. Because tracking hits (accuracy) and tracking distance (speed) were negatively correlated ($r = -.444$, $p < .0001$), it is conceivable that gunners with a higher AFQT approached the tracking test with a greater emphasis on accuracy than did low AFQT gunners. Tracking distance did not correlate with AFQT. It is not possible to test the interaction between AFQT and the speed/accuracy tradeoff with these data.

Table 1
Job Sample Measures That Correlated with AFQT

| Tests and Measures | r | (p <) |
|---|---|---|
| **Tracking** | | |
| Number of Hits | .360 | (.0001) |
| Composite Hits and Time | .316 | (.0004) |
| **M1 Computer Panel** | | |
| Enter-Check Data (ECD) Correct | .214 | (.0172) |
| Enter-Check Data (ECD) Time | -.513 | (.0001) |
| Enter-Check Data (ECD) Composite | .444 | (.0001) |
| Computer Self-Test (CST) Time | -.339 | (.0001) |
| Computer Self-Test (CST) Composite | .237 | (.0082) |

AFQT is more commonly reported in its grouped or categorized form, that is, by AFQT category rather than by percentile. When gunner performance on three job sample measures is presented by mental (AFQT) category the contrasts are striking. Note in Table 2 the drop in performance for category 3B and 4 soldiers. The measures presented include time and accuracy composites for the two M1 Computer Panel tests

and a General Ability Composite (GAC) formed by totaling the two Computer Panel composites and a time/accuracy composite from the Tracking Test. The GAC was developed using stepwise regression techniques; it was found to yield the most accurate prediction of AFQT. Average gunner scores were obtained by taking the average standardized score, adding a constant, then multiplying by 100. The correlation between GAC and AFQT was $r = .49$ ($p < .0001$).

Table 2
Average Gunner Score by AFQT Category

| | AFQT Category | | | | |
|---|---|---|---|---|---|
| Job Sample | 1 (N = 13) | 2 (N = 41) | 3A (N = 27) | 3B (N = 17) | 4 (N = 25) |
| ECD Composite | 198 | 141 | 139 | 47 | 19 |
| CST Composite | 170 | 151 | 91 | 55 | 10 |
| General Ability Composite | 346 | 211 | 105 | 26 | 5 |

The AFQT measures correlated with the high cognitive weighted tasks for gunner--ECD and CST on the M1 Computer Panel but not with the Table VIII scores. Because job sample tests were constructed with a combat criterion in mind, the snakeboard used in the hands-on tracking task involved tracing an extremely circuitous route, more similar to a threat target employing evasive maneuvers than to the slow-moving flank silhouette targets normally found on a Table VIII range. Thus one might conclude that if Table VIIIs were designed to more accurately mirror threat scenarios, the relationships between job sample tests or AFQT and live-fire gunnery might be easier to establish.

## Conclusion

The present effort has involved the development and evaluation of skills tests for use as combat performance predictors. No relationships were observed for the supervisors' ratings and Table VIII day scores. These findings suggest that a UCOFT-implemented job sample testing approach may be more appropriate. UCOFT will allow high fidelity simulation of threat scenarios without the concomitant costs and safety hazards associated with a live-fire exercise.

Pending further research on job sample ability testing, a multiple hurdle approach to M1 gunner and tank commander selection may be suggested. Job sample ability testing for position-specific requirements may be combined with the on-site commander's evaluation of crewmember performance. This testing may offer a feasible approach to crewmember selection. Further research is required to assess how job sample testing can apply to tank commander selection.

# A Retrospective Analysis of Instructional Technology Innovation Using General Systems Theory

Ana G. Ekstrom
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

The U.S. Military today operates in a worldwide, multi-lingual, multi-cultural context. The ramifications of this fact on linguistic training requirements presents a formidable challenge. There is a consistent need for a vast and constantly replenished supply of linguists trained in a wide range of languages with specialized skills to perform many and diverse tasks. Organizationally, the training requirement includes the capacity to teach even obscure infrequently needed languages as well as the ability to respond to the periodic, sudden and often unpredictable surges in demand for a particular language as military requirements develop.

The Defense Language Institute Foreign Language Center (DLIFLC), the foreign language training arm of the military, has found, despite claims made for various methods -- from Berlitz to Super Learning -- that learning a foreign language is a time-intensive process. Yet linguist personnel must be rapidly trained to functional levels of proficiency if the military is to benefit from short term enlistments, as well as meet surge requirements.

The use of interactive video in foreign language instruction has high face validity. Beyond the learning enhancement provided by multi-sensory presentation of instructional materials, interactive video offers the capacity to present a language in realistic situational and cultural contexts wherein the ability to understand and make computer responses, visibly results in simulated real-world consequences. Organizationally, a machine-based instructional system affords a versatile and flexible means of augmenting existing teaching structure, thereby potentially increasing institutional flexibility.

To take advantage of potential technological contributions to training capacity, DLIFLC established a specialized office to examine innovative instructional technologies. The New Systems Training Division (NST), identified the potential of the interactive videodisc in foreign language education and began preliminary exploration of developmental feasibility. This preliminary exploration led to what would ultimately become a complex "system" of individuals and organizations with the objective of field testing a Video Enhanced German Gateway Program (VEGG). *

* The term VELVET (Video Enhanced Learning Video Enhanced Testing) has been adopted by the developer for the video materials. The evaluation, concerned with assessing the use and impact of the materials on the German Gateway Program, rather than evaluating the materials themselves, uses the VEGG acronym.

This paper uses the evaluator's experience and knowledge of the VEGG system to illustrate the principles and utility of a systems perspective. This is not a report of findings and conclusions of a system evaluation. Rather it is intended as an example of how a systems perspective can assist in the organization and interpretation of events, actions, and reactions experienced in evaluation efforts even where the evaluation itself is not explicitly system oriented. We will begin with a description of the VEGG system in terms of its components and then illustrate the system principles described by Atwood (1985) using examples drawn from the VEGG system.

Initial development processes entailed feasibility and needs assessments related to materials development and the establishment of support within the military system to provide the requisite authorization, funding, and sponsorship base. The components of this particular instructional technology development system evolved incrementally with organizational units added in process. First requirements were for a "test bed", materials development personnel, equipment, and initial funding. The test bed requirement brought in the German Gateway Program (GG); materials development added two contractors; start-up equipment and materials development funding was provided by the Army Communicative Technology Office (ACTO) which became the technical sponsor; and test/evaluation funding was provided by the Training Development Institute (TDI) which became test sponsor. TDI subsequently tasked the Army Research Institute (ARI) with evaluation responsibility. TRADOC reorganization turned TDI into a new office of Training Technology Agency (TTA) and ACTO into Communicative Education and Training Systems Management Office (CETSMO) with consequent changes in personnel and uncertain sponsorship.

As the project moved toward the implementation phase, contractor and equipment funding were assumed by the National Security Agency (NSA) and only one of the two original contractors being retained. Within DLIFLC, quality assurance and research components were identified to review instructional materials and evaluation procedures respectively. Arrangements for video production were made with specialists from Ft. Dix and a full time Subject Matter Expert (SME) was provided to the project by the DLIFLC German school.

With the major VEGG system components now identified, we can examine the VEGG system applying General Systems Theory. The illustrations are necessarily drawn from the view available from the evaluator's window. Looking out the window of a different system component would certainly give a very different picture of the system itself as well as supply a whole different range of experiences from which to draw.

## Synergism

Synergism is an outgrowth of regularized interaction, communication, and coordinated activity in the shared pursuit of system objectives. Time, distance, structural arrangements, expertise, and policies all impact on system synergy. Momentum in project development can be seriously affected by interruptions. The VEGG project experienced more than a one year delay between contractor selection and contract award for start-up activities and an additional one year postponement of field testing while awaiting completion of instructional materials. Such delays may divert evaluator and contractor focus and energies away from system objectives and into other projects, and reduce levels of interest, commitment, and involvement of sponsors and test units.

Conversely, delay may also serve positive functions by allowing participating units to accommodate to system requirements -- especially in the early stages of development. For example, the SME has a critical function in this system as the intermediary between developers and GG, translating language and instructional requirements into programmable lessons. The previously mentioned contracting delay provided the time needed for the SME to arrange to be released from teaching activities and to complete test development responsibilities. In this instance, the delay facilitated project development by enabling the SME to devote full time to instructional development when the project got underway. Likewise, the same period provided the evaluator (ARI) with the opportunity to become more familiar with the specialized areas of foreign language education and interactive video systems.

Ideally from a synergy point of view, core development staff (and perhaps evaluator staff) would share a common central facility. Where this is not possible, means of communication and feedback which simulate this type of working environment need to be established. The VEGG project began development efforts relying on the mails, phone calls, and periodic site visits back and forth to coordinate DLIFLC and contractor efforts. The contractor sent completed lessons to DLIFLC for review. Feedback, corrections, and suggested changes were phoned or mailed back. Prompted by something of a crisis created by the delay in materials development, direct telephonic computer links were established between the contractor and NST enabling instant feedback, correction, and literally simultaneous developmental effort to occur.

System Structures

Lines of authority and division of responsibility define vertical and horizontal structure respectively. When control of decision-making and resource allocation resides within the development system, the system is able to organize to achieve its objectives. When major control does not reside in the development system, it can only be understood as a sub-system within a larger system(s). Under the latter condition, the development system's objectives may become secondary to primary objectives of the larger system(s). Resources then must be negotiated and participants may be subject to conflicting authority.

ARI's original relationship to the VEGG project was on a Technical Assistance Service (TAS) basis which is by definition a short-term, limited resource effort. This TAS-level relationship specifically defined ARI as a limited stakeholder in the VEGG system. The long-term intensive commitment required for a systematic systems evaluation is precluded - - the VEGG system's evaluation resources could be diverted at any time into ARI's primary program objectives. In addition, ARI evaluation staff is subject to both the authority of the ARI structure in determining evaluation design features, procedures, and time commitments and to the German Gateway management in defining permissible data collection intrusions into the program.

By virture of his position as intermediary for a number of components of this system, the SME is even more enmeshed in multiple lines of authority. As translator of foreign language pedagogy into technology based instruction, the SME may be caught between the conflicting policies or needs of the various

components of the system. For example, GG instructional policies demand exclusive use of German in educational materials. This GG policy came into direct conflict with an NST/contractor decision to use limited English to support students' initial use of the IVD, resulting in directly opposed demands on the SME. This illustration exemplifies a type of technical issue that emerges in the course of converting conventional instructional methods to media-based technologies.

As evident from the above, lines of authority determine where critical decisions are made, by whom, and in terms of what objectives. Depending on their impact on system relations, decisions also promote, build on, inhibit, or destroy system synergy.

Stability

The ability to achieve a degree of internal stability is required to function in terms of system goals. At the same time, instructional technology development systems must overcome stability in the existing educational structures in order for technological innovations to be adopted. The VEGG project faces both challenges.

The process by which the VEGG system was constituted was essentially additive. That is, units were added as required by either the system's purpose or as imposed by external authority. The result of this process is something akin to a perpetual motion system forever adjusting to the requirements of new components. To illustrate from the evaluator's perspective, ARI, in its evaluation effort, identified the inadequacy for evaluation purposes of the existing student performance measures used in the GG program. This necessitated the development and administration of a new test, a requirement which, when combined with data collection needs, significantly added to the responsibilities of teaching and testing staff of the GG program. When the contractor added evaluation staff approximately half way into the baseline data collection (contractor funding actually occurred after data collection had begun), weaknesses in evaluation measures were identified by their foreign language specialist evaluator which imposed additional requirements on ARI and DLIFLC system components. Still later, a DLIFLC researcher entered the system (newly hired to fill a newly created position) and identified additional contractor evaluation responsibilities. The disruptive influence of these events might have been averted if an evaluation component, operating from a systems framework in the early stages of development, had facilitated a broader more realistic perception of the nature and requirements of the overall program.

A systems oriented evaluation also would have broadened the scope of outcome concerns to include examination of processes and relations which facilitate adoption of technology once developed. On the VEGG project two contractors were involved in preliminary feasibility and needs assessment activities. One contractor, with both technical and linguist capability, developed a product - - an interactive videodisc demonstration lesson. The other contractor, lacking linguist capability, operated as a consultant, training members of the DLIFLC staff in design concepts and assisting them in developing sample lesson designs. From the larger system perspective, the product is the desired outcome. Products can be more efficiently acquired from a contractor than by expending limited and essential specialized German

instructional staff resources in product development.

The selection of the first contractor established a "product model" (Rutt, 1984) relationship between the contractor and NST/GG in the development system. This selection limited the demands made of GG staff to that of insuring product quality/acceptablity for their program. This decision reduced both the potential strain on DLIFLC German staff resources and the dependence of the developmental system on those resources. At the same time, it reduced the GG stake in the system. The second contractor, requiring DLIFLC German language and instructional expertise, would have necessarily involved the German staff to a far greater extent. This involvement might have fostered greater appreciation for developmental issues and strengthened investment in the final product and its use in the program.

Optimization

"If sub-systems operate to optimize their own individual performance the net result will almost never be overall system optimization" (Atwood, 1985). Several examples from the VEGG system illustrate this point.

In the case of the previously described successive efforts to modify the evalution design, one component proposed modifications in the evaluation design which would have strengthened the evaluation but would have expanded data collection efforts. The outcome was that the system collectively operated to limit expanded data collection due to the excessive intrusion and demands that would have been placed on GG student and staff resources, ARI evaluation resources, and SME time.

As developers discover the advantages offered by various technological possibilities, additional time is required to integrate new features into product development. For example, VEGG developers became aware of the highly desirable possibility of incorporating audio commentary or questions into their materials through use of the second audio track of the video disc. Understanding the full potential of a technology options is an evolutionary process. That evolutionary process, combined with the constant marketing of new hardware, software, and courseware that occurs in the electronic media field may entice developers into exploring new or newly recognized possibilities at the expense of timely product completion.

Another example deals with quality assurance review of instructional materials. This must necessarily be a constant process, deciding on almost a frame-by-frame basis what is excellent, what can be lived with, and what is totally unacceptable. Grammar and syntax errors are easily recognizable as unacceptable. Reviewers tend to readily identify excellence. When video instructional materials fall into the no-man's-land between the extremes, constant and difficult decisions must be made in terms of what can be lived with. Insistence on excellence only, may result in no product at all.

The fine art of negotiating between "must have/do" and "would like to have/do" characterizes the process of achieving system optimization over optimization of individual component performance. Systematic documentation of this negotiating process would provide information relating to the balance that exists among the array of sub-system objectives and between sub-system and development system objectives.

Equifinality

The principle of equifinality has two important implications for system evaluation. The first relates to the need to identify options and alternatives for "as the number of alternatives increases, the system is more able to adjust to unforseen changes in the environment" (Atwood, 1985). As has been implied throughout this paper, there are many ways to achieve system objectives. Different choices in terms of equipment, contractor, evaluation approach and design, target programs and language would have substantially changed the nature and operation of the VEGG system and its final products. Decisions made in the early stages of development limit what can follow and determine to a great extent what must follow. For example, the choice of a Gateway program limited the test population to officers, dictated a baseline/field test comparison rather than an experimental design, and limited outcome performance measures to short term learning at low level, limited range proficiencies. These limitations may well be less significant to overall system goals than the advantages gained for example from the greater liklihood of product adoption by an already equipment-based (audio cassette) program and the limited but more manageable scope of evaluation. The important point is that the involvement of all parties early in the planning process and in decision-making throughout development increases the system's ability to anticipate its overall needs and to develop contingency alternatives.

The principle of equifinality also maintains that the objectives of a system may be achieved "from a number of different starting conditions and through a variety of means" (Atwood, 1985). Fortunately, a system's ability to achieve its purposes does not depend on a precisely defined set of components, processes, and relationships. However, within the framework of equifinality, the cumulation of a body of information - - through systems evaluations of a variety of instructional development systems - - would begin to refine our understanding of what works better, when, and how - - and not at all.

To the extent that a systems approach to evalution does not result either in information overload or in surfacing such an overwhelming number of issues that instructional development initiatives are stymied, formative systems evaluation enhances the development process. Properly used and understood, the systems approach enables evaluators to perform a comprehensive job, collecting information that can be used both internally and externally to achieve system objectives.

References

Atwood, N.K. (1985). A Systems Approach to Evaluating High Technology Training. Paper presented at Military Testing Association (27th Annual Meeting).

Churchman, C.W. (1968). The Systems Approach. New York: Delacorate Press.

Rutt, D.P. (1984). Consultation in Instructional Development: A First Look. In R. Bass and C. Dills (eds) Instructional Development: The State of the Art II. Dubuque, Iowa: Kendall/Hunt. 294-309.

von Bertalanffy, L. (1968). General Systems Theory. New York: George Braziller.

Wolf, W.C. (1984). Linking Knowledge Production and Needs of Knowledge Users. In R. Bass and C. Dills (eds.) Instructional Development: The State of the Art II. Dubuque, Iowa: Kendall/Hunt. 259-379.

# A Systems Approach to Evaluating High Technology Training

Nancy K. Atwood

U.S. Army Research Institute Field Unit
Presidio of Monterey, California

Military training demands have steadily increased with the advent of the computer age and the development of complex electronics and weapons systems. At the same time, the volunteer force has been faced with a constrained supply of skilled recruits. To address these dual problems, instructional applications of new technologies have received considerable attention in the military training community as a strategy for leveraging training resources (O'Neil, 1981). For example, the U. S. Army Research Institute maintains a "technology watch" to identify potential techniques for increasing the cost effectiveness of training, for raising levels of learning at the same cost as conventional methods, or for achieving the same results as conventional methods at lower cost.

One promising new technology is "Intelligent" Computer Assisted Instruction or ICAI. ICAI, a field within the larger discipline of Artificial Intelligence, draws upon emerging computer technologies and cognitive psychology in an effort to provide a new sort of computer-based learning environment. The intent is to provide an intelligent tutoring system that simulates a human coach or tutor. The enthusiasm of ICAI developers is reflected in the comments of Anderson and Reiser (1985) who argue: ". . . The prospect is great of providing every student with the educational benefits of a private human tutor. When this happens, the consequences for American education will be nothing short of revolutionary."

While ICAI is a rapidly evolving field, functional programs are appearing, funded in large part through the investment of DoD resources. It is vitally important at this time for the military training community to critically examine instructional applications of the new, AI-based computer technologies. This examination should be directed toward understanding the state of the art of ICAI technology and the process for developing ICAI systems with a view toward shaping the development process and the emerging product to military training needs and requirements.

This paper presents a strategy for designing and conducting early or "formative" evaluation of instructional applications of new technologies. The approach, based on General Systems Theory, is first described. This conceptualization was in fact stimulated by problems experienced while attempting to apply a conventional evaluation methodology to three ICAI projects (Baker & Atwood, 1985). Then insights into the utility of the systems approach are presented based on a post-hoc analysis of experience with these projects.

## A Systems-based Formative Evaluation Strategy

Virtually all prescriptive models of the instructional development process specify the need for "formative evaluation" to identify program strengths and weaknesses for program improvement by collecting information on instructional processes and learner performance. A case in point is the Instructional System Development (ISD) model of military training development which

prescribes evaluation as an integral component of the development cycle (O'Neil, 1979). However, evaluation is generally conceptualized as one step in a linear sequence of development procedures which is initiated when a prototype product is ready for field-testing.

In contrast, the systems-based formative evaluation strategy to be described here conceives of evaluation as an ongoing process that begins with project start-up and continues throughout full project implementation. The approach views the organizations involved in instructional technology development as sub-systems within a larger system. The principles of General Systems Theory are applied to derive features of an evaluation strategy that will facilitate successful development and implementation of the technological innovation.

Systems theory encompasses a vast range of issues and approaches ranging from systems philosophy to development of mathematic system theories, to empirical research on system behavior, to systems engineering (Mattesich, 1982). However, at the heart of systems thinking is an emphasis on input-output features, a purposeful orientation with concerns for means-ends relationships, and feedback loops within the system to adjust the behavior of system components to yield the desired end-behavior of the entire system. (See, for example, von Bertalanffy, 1968.) The essence of systems thinking lies in a wholistic, non-linear view that constantly relates components of the system back and forth to the overall system and tries to reconcile their often conflicting goals.

Figure 1 presents a prescriptive systems model of instructional technology development. Generally, there are at least three interacting organizations or system components in the military training setting. These include the contracting agency that funds the project, the research and development organization (usually a private business or university) that develops the product, and the school that is the targeted user. (As shown in the figure, there are also sub-systems operating within each of these components.)

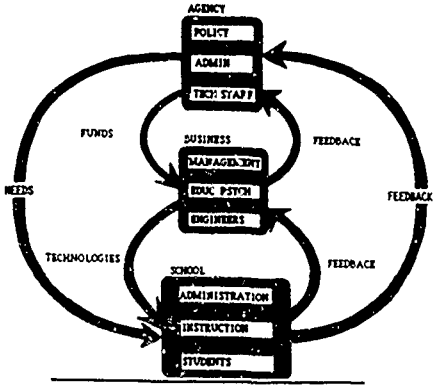



Figure 1 Systems model for instructional technology development

Traditionally, the development process has relied on minimal interaction between the above organizations. Given the natural bureaucratic barriers to

communication, organizations generally take responsibility for their portion of the process, with little communication and coordination with the others. Furthermore, organizations often have competing or conflicting goals (e.g., funding agencies looking for specific training applications, R & D organizations concerned with producing scientific knowledge, and user organizations looking for affordable and feasible enhancements to current procedures). As the history of Computer-Assisted Instruction (CAI) has amply demonstrated (Montague & Wulfeck, 1984), the end product often is not well targeted to the school's instructional needs, may be incompatible with the organization of the instructional program, and further may not be fully implemented or even used because those school officials having authority do not feel ownership or commitment to the project.

In contrast, the system view in Figure 1 recognizes the critical importance to technology development of communication through feedback loops between the various groups within the system. These loops provide a vital mechanism for ensuring that the resulting innovation is both objectively valuable and psychologically valued by the organizations involved with the project.

Principles drawn from General Systems Theory provide insights into conditions that will maximize the liklihood of achieving the overall goal of the system (in this case, successful implementation of an effective instructional technology innovation):

1. Synergism: As integration among system components increases, synergistic effects (i.e., synchronized energy) are more likely to occur. Extensive communication among system components, in this case, the funding agency, R & D organization, and school, are needed to develop the strong, well defined interrelationships required for integration to occur. This dialogue must yield mutual understandings and acceptance of relative roles and working relationships.

2. System structure: The clearer the vertical and horizontal structure, the more the system is in control and the more persons and activities within the system are vested with purpose and meaning. Again, the communication and feedback loops within the system properly used can serve to clarify the system structure and yield common understandings of the respective responsibilities of components within the system.

3. Optimization: If sub-systems operate to optimize their own individual performance, the net result will almost never be overall system optimization. Thus, in order to achieve overall success, system components must embrace the overall system goal, agree to cooperate, and devise strategies for accomplishing unique organizational goals within the larger context of the system goal. Joint planning sessions with mutual discussion and negotiation on project parameters can serve to create a sense of ownership by all parties and to adjust project plans so that the interests and needs of component organizations are incorporated.

4. Stability: All systems tend to have mechanisms for manning a steady state. This tendency leads to the well-recognized resistance to instructional innovations (e.g., Berman & McLaughlin, 1977) and implies a need for planning to anticipate the problem. Such planning should establish organizational supports for re-establishing stability after the introduction of the innovation.

5. Equifinality: The final state of an open system can be reached by starting from a number of different initial starting conditions and through a variety of means. As the number of alternatives increases, the system is more able to adjust to unforeseen changes in the environment and to accomplish its purposes. Thus, project planning should be open to multiple options for development and implementation to increase the adaptive flexibility of the effort.

The above systems view implies that formative evaluation should be structured to facilitate system functioning in order to achieve the system goal, to contribute to communication and feedback among system components, to provide information on the needs and concerns of groups within the system, and to gather information on program functioning and outcomes for revision purposes. Ideally, the evaluation should:

- o be conducted by a team external to the system. An independent team maximizes objectivity and minimizes self-involvement in the interests of organizations within the system. As such, it can serve to facilitate overall system functioning.
- o examine the functioning of system components as well as their interaction process from project development through full implementation. Recommendations for improvement should focus on promotion of cooperation among system components to achieve overall project goals.
- o capture ongoing communication and feedback mechanisms between all system components. Description and identification of strategies for improvement are critical since these mechanisms can serve to increase integration and synergism of system components and to clarify system structure.
- o address specific information needs of all organizations within the system. This information can serve to meet the needs and goals of individual components within the system and to resolve conflicts that may interfere with achieving the overall system goal. Explicit recognition of the needs of each organization also serves to establish the credibility and utility of the evaluation to all concerned parties.
- o be sensitive to the stage of development of the project. Information needs and concerns change as a project matures and the freedom to make changes diminishes. In the early stages of project planning, focal points should include articulation of instructional/training theory, operational plans for product development, initial implementation planning, and use of communication/feedback loops within the system. In the middle stages of project design and development, focal points should include product design and development, theoretical fidelity and quality of the instructional process, preliminary implementation (with supports reestablishing system stability), outcome measurement and ongoing communication/feedback processes. In the later stages of project implementation, focal points should include outcome assessment (both cognitive and affective), measurement of side-effects, project implementation, cost-effectiveness, and ongoing use of communication/feedback loops.

- capitalize on multiple measurement methods for collecting information. Diverse methods, both quantitative and qualitative, are helpful in acquiring the full range of relevant information.
- yield a full documentary base to inform subsequent decisions and to provide a full record of intentions and accomplishments. This record will contribute to an understanding of the functioning of the system and its components and will provide a vehicle for discussion in subsequent project planning and evaluation activities, including follow-on projects.

Insights into the Need for a Systems Approach

The systems-based evaluation approach described above was used to analyze experiences and problems encountered conducting evaluations of three ICAI projects. The projects differed in their stage of development from early to middle to late. The project in the early stages of development was designed to teach maintenance technicians to troubleshoot and repair a complex electronic machine. The project in the middle stages of development was intended to assist novices learn to program in PASCAL. The project late in its development cycle was directed toward teaching basic mathmatics and strategic thinking skills within a game context.

The analysis suggested that the systems-based evaluation approach was a useful tool for understanding the problems that emerged with conventional formative evaluation methods and for identifying promising strategies for future evaluations of such technical innovations. Foremost, strong incentives are required for full participation by all groups within the system. While the funding agency and the school generally perceived participation in evaluation activities as in their own self-interest, project developers were considerably more skeptical and resistant to participation; evaluation activities tended to be perceived as directed toward uncovering weaknesses and diminishing their control over the development process. Certainly one strategy for gaining full cooperation from the developer is to include external evaluation in the project funding agreement.

Second, the relevant range of expertise needs to be represented on the evaluation team in a manner that is credible to the funding agency, the developer, and the school. In one evaluation study, only limited cooperation was obtained from developers due, at least in part, to the fact that the evaluators belonged to a different professional community than the developers and were not viewed as knowledgable or capable of providing useful information. For ICAI projects developed for military training applications, the relevant range of expertise minimally includes: instructional psychology, military training, artificial intelligence, and the subject matter to be taught (e.g., electronic maintenance, PASCAL programming). In addition to constituting an evaluation team with skills in these areas, additional credibility would be accrued by convening a technical advisory panel of esteemed colleagues of developers (i.e., key figures in Artificial Intelligence).

It is also vital for conditions to be established that foster the development of informal networks and the growth of trust and credibility from the inception of the project. In these evaluations, trust and credibility developed slowly over the course of many months and was sometimes impeded by lack of knowledge about strategies and events occurring earlier in the development cycle. Supporting conditions for fostering productive working

relationships include involving all parties (including the evaluation team) at the beginning of the planning process and recognizing the need for long-term (i.e., multi-year) development and evaluation agreements. Such agreements are necessary for adequate communication and feedback to occur, for the extended development cycle encountered in projects with a high degree of technological complexity to run its course, and for mutual trust among individuals and groups to develop.

Finally, safeguards on the use and distribution of information collected by the evaluation team are required. Much of the difficulty encountered working with developers, and to a lesser extent, project implementers, arose from concerns about: (a) how and to whom the evaluation information would be disseminated; (b) whether and at what point they would have an opportunity to react to evaluation findings and to voice their interpretations of findings; and (c) how the evaluation results would be used. Thus, it is critical that all parties are assured that information will be treated appropriately and sensitively and procedures are in place to protect such assurances.

One such procedure for restricting information flow to guard such assurances is for the evaluation team to limit dissemination of its findings to organizations within the system, i.e. the funding agency, the R & D organization and the school. Groups should be provided an opportunity to react to the findings and to present their interpretations to the evaluation team. While ultimate incorporation of such input into the final evaluation report should be left to the professional judgment of the evaluation team, organizations should be provided the opportunity to prepare a "minority report", if they have serious disagreements with the final evaluation report. Finally, all publication and presentation of results to the technical and training community should be approved by organizations within the system studied.

It is hoped that a systems-based formative evaluation strategy will support the development of high quality instructional technology innovations, and, perhaps more ambitiously, yield a record that may provide guidance for establishing general policies related to training technology R & D.

## References


Anderson, J. R. & Reiser, B. J. (1985). The LISP tutor. Byte, 159-175.

Baker, E. L., & Atwood, N. K. (1985). Shaping the wind: Formative evaluation of intelligent computer-assisted instruction (ICAI). Paper presented at Annual Meeting of American Educational Research Association, Chicago, IL.

Berman, P. & McLaughlin, J. W. (1977). Federal programs supporting educational change, Vol. III: Implementing and sustaining innovations. R-15891 8-HEW. Santa Monica, CA: Rand.

Mattessich, R. (1982). The systems approach: Its variety of aspects. Journal of the American Society for Information Science, 383-394.

Montague, W. E. & Wulfeck, W. H. (1984). Computer-based instruction: Will it improve instructional quality? Training Technology Journal.

O'Neil, H.F., Jr. (1981). Computer-based instruction: A state-of-the-art assessment. New York: Academic Press.

O'Neil, H. F., Jr. (1979). Issues in instructional system development. New York: Academic Press.

von Bertalanffy, L. (1968). General system theory. New York: George Braziller.

# An Analysis of Attitudes Toward Instruction Among Vocational Education Instructors

George M. Usova, Head
Analysis and Evaluation Division
Shipyard Instructional Design Center, Atlantic
Portsmouth, VA 23709

Identifying attitudes toward instruction is important in determining program changes. An extensive study, conducted by George M. Usova sought to determine the instructional beliefs and practices held by skilled trades training personnel in naval shipyard apprenticeship programs.

The study surveyed the attitudes and practices of 256 key training personnel in apprentice trade training areas. The trades participating in the study were air conditioning and refrigeration, boilermaker, electrician, electronics, fabric worker, machinist, insulator, painter, pipefitter, rigger, sheetmetal, shipfitter, shipwright, and welder.

The trade trainers were asked to rate their feelings of agreement or disagreement toward educational practices by using a five point scale. The 17 item questionnaire measured attitudes on apprentice ability in physical and basic communication skills, use of self-paced and computer-managed instruction, adapting instruction to learner differences, and values held about inservice training. In addition, the questionnaire asked the trainers to indicate their primary method of instructional delivery, state how often they received inservice training, and tell how often apprentices used effective learning strategies to learn subject area content.

Background. Since the U.S. Navy is the single largest employer of apprentices in the nation, it has a substantial stake in assuring that its apprentice training programs are effective. One area of concern and investigation lies in the area of instructor training. Apprentices are trained by instructors who receive only 40-80 hours of preparatory instructor training.

The intent of the study, therefore, was to assess the attitudes and practices toward teaching and instruction among vocational trade instructors and other key training personnel in apprenticeship programs in naval shipyards. The importance of the study was to provide data from which to make better decisions for improving apprenticeship programs and identifying instructor training needs for inservice training.

Method. The Training Information Survey was developed as a two-part questionnaire to assess particular training attitudes and practices of training personnel (Group Superintendents, Shop Superintendents, Supervisory Training Instructors, and Trade Theory Instructors) involved in apprentice training at the naval shipyards (see Appendix A). The survey questionnaire is composed of two parts: (1) Part I: an attitude section that assesses agreement or disagreement with statements about apprentices' abilities, educational practices, and inservice training and (2) Part II: an "actual practice" section that assesses the extent and frequency with which instructional practices actually occur. Additionally, the questionnaire solicited biographical information (i.e., the respondent's trade, shipyard, and job position) that was matched with the questionnaire data.

In total, 472 questionnaires were sent to eight shipyards with 286 questionnaires returned (within the five-month period), which yields a return rate of 60.6 percent. A breakdown of useable questionnaires by respondent groups reveals the following figures: 19 Group Superintendents, 77 Shop Superintendents, 32 Supervisory Training Instructors, and 128 Trade Theory Instructors. The trade areas surveyed were air conditioning and refrigeration, boilermaker, electrician, electronics mechanic, fabric worker, machinist, insulator, painter, pipefitter, rigger, sheetmetal mechanic, shipfitter, shipwright, and welder.

Findings and Recommendations. As a result of the study, 11 separate findings were made from a data analysis of the study as a whole The findings and recommendations incorporate the results from Parts I and II of the questionnaire.

1. Apprentice hires do appear to possess the manual dexterity and coordination skills necessary to meet the demands of skilled trades training.

2. Apprentice hires do lack to a substantial degree basic skills in reading, mathematics, and communications to meet the demands of trade theory courses. Nearly 70 percent of training personnel believe apprentices are deficient in one or more basic skills. Verbal comments additionally support this finding.

3. Self-paced instruction is not well-accepted as an instructional delivery mode. This finding is further reinforced through other data showing that 82 percent of trade theory instruction is group-paced. Group-paced (predominantly lecture) is a traditional form of instruction that nearly everyone is familiar with. While group-paced instruction can be effective in achieving learning, it is widely recognized and established that there are a variety of alternative instructional methods (self-paced, individualized, simulation, grouping, team learning, and others) that can be more effective in achieving certain learning outcomes. It is the awareness, knowledge, and implementation of varied methodologies that can strengthen the instructor's teaching competency.

Recommendations. It is recommended that, through inservice training, training personnel be shown the educational values and benefits that can arise in student learning through self-paced instruction. The point here is not to attempt to force self-paced instruction upon trainers nor is it to be implied that self-paced instruction is the best delivery mode; rather, the intent is to recognize the use and role that self-paced instruction can play as a part of an eclectic (varied) approach to training.

4. There is wide agreement and support that a planned on-the-job instructional program could benefit shop training. Support, therefore, exists for a program that systematically permits and ensures that critical job skills are experienced and developed during on-the-job training.

Recommendations. It is recommended that all trade apprentice programs ensure that apprentices experience critical job skills through a planned sequence of experiences.

5. Overall, there is uncertainty over whether computer-managed instruction would assist learners.

Recommendations. As stated in recommendation number three, there is a need to educate trainers to the benefits inherent in using a computer-managed system. In an age of information explosion, instructional technology, and cost/time efficiency requirements, computer-managed instruction can perform an integral role.

6. There is widespread acceptance to the need for hands-on training removed from the production environment. This educationally sound concept permits learners to practice and develop skill proficiency on training aids, mock-ups, and simulators before actually "working a job". In this manner, theory and practice are closely tied together to ensure skill mastery and thereby reduce the possibility of costly errors that might occur on-the-job.

Recommendations. It is recommended that all trades explore additional sources and acquire, as needed, training aids, mock-ups, and/or simulators to develop improved hands-on training for vestibule training.

7. There is overwhelming support for the value, benefit, and need for instructor inservice among all trade training personnel. This belief lies in the generally accepted concept that if instructors become more competent in the area of training (learning theory and practice, objective and test-item development, instructional strategies, etc.), then students will achieve in greater proportions. Central to the issue of inservice training is the type and quality of that training. Inservice training must be directly responsive and relevant to the improvement of the teaching-learning process, i.e., the topics of inservice must address instruction-competency areas.

Recommendations. It is recommended that shipyards develop a systematic instruction training inservice program that is based upon instructor needs.

8. There is strong acceptance for the use of modern audiovisual aids to support instruction.

Recommendations. It is recommended that trade training personnel review their audiovisual acquisitions to determine whether they possess the proper type and amount of equipment necessary to support and facilitate learning of trade theory courses.

9. At the present time, strong agreement exists that shop trade theory courses meet the learning needs of apprentices.

10. Uncertainty exists about whether courses are presented in the different ways in which students learn. Supporting this uncertainty belief is survey data which show that 70 percent of respondents believe that instructors only "sometimes" adapt instruction to learner differences. Again, this is indicative of instructional rigidity or inflexibility to alternate methods and parallels the finding stated earlier. 82 percent of shop instruction is group-paced.

Recommendations. Launch a program of instructor inservice which includes topics on detecting and diagnosing learner abilities, individualizing instruction, classroom grouping patterns, peer learning techniques, and other topics as cited in recommendation number three.

11. Only 22 percent of students in trade theory courses demonstrate efficient learning strategies and study skills. Learning strategies and study skills, such as notetaking, listening, time management, memory techniques, and study methods, are the means through which to acquire and retain information. Responsibility for teaching those learning skills is a dual responsibility, requiring effort from both student and instructor. The instructor is responsible for showing students how to learn and better recall that specific information that is presented in the classroom; again, the instructor can be taught how to do this through inservice training. The student, on the other hand, can be taught the general application of learning strategies and study skills and must be personally responsible for applying those skills.

Recommendation. It is recommended that a course on learning-study skills be provided instructors as a component of an inservice training program and that a separate and distinct program on learning-study skills be offered to all apprentices during their first year in the program. Such a two-pronged approach will maximize learning efficiency.

## TRAINING INFORMATION SURVEY

Trade _____ Shop _____

Status: (Please check the one that applies to you.)

_____ Group Superintendent
_____ Supervisory Training Instructor
_____ Shop Superintendent
_____ Trade Theory Instructor

The Shipyard Training Support Center role is to support your effort to strengthen and improve apprenticeship training. We need your assistance in completing this survey form to help us define and clarify shipyard apprentice training needs for your shop. Please take a few minutes of your time to respond to the following statements. Return the completed form to your employee development representative, or if requested, your group superintendent. Thank you.

Directions, Part I

After reading and considering each statement, assign the number that reflects your feeling about the statement.

1 - Strongly Disagree
2 - Disagree
3 - Undecided
4 - Agree
5 - Strongly Agree

1. ___ New apprentice hires have the manual dexterity and coordination to meet the demands of skilled trades training.

2. ___ New apprentice hires possess the necessary basic skills in reading, mathematics, and communications to meet the demand of skilled trades training and trade theory courses.

3. ___ Self-paced instruction (that allows each apprentice to control his own rate of learning) could benefit the learners of skilled trades training in the shop.

4. ___ A planned on-the-job instructional program could benefit training in the shop.

5. ___ Computer enhanced and computer managed instruction would enhance the learning potential of students in the shop.

6. ___ Hands-on training, removed from the production environment, is needed to increase trade skills.

7. ___ Shop instructors could profit from an ongoing program of in-service training designed to improve teaching skills.

8. ___ The improvement of instructor teaching skills will improve student learning.

9. ___ Current and modern audiovisual aids and equipment are important in supporting student learning in skilled trades training.

10. ___ Presently, shop trade theory courses meet the knowledge and skill learning needs of apprenticeship students.

11. ___ Presently, shop trade theory courses are presented to meet the different ways in which people learn.

Directions, Part II

In order for us to more effectively understand your shop's training program, please respond to each statement as it relates to your shop's operation by checking ( ) the appropriate blank. Thank you.

1. Most of the instruction delivered in my shop is

_____ group-paced _____ self-paced
_____ others; specify _____

2. Instructors regularly receive in-service training in the improvement of teaching skills

_____ more than once per year _____ once per year
_____ none received
_____ other; specify _____

3. Records kept on student progress/skill levels are

_____ maintained regularly _____ not maintained
_____ maintained sketchily

4. As a group, instructors adapt instruction to account for differences in learner abilities, ex., slow learners vs. fast learners

_____ always _____ sometimes _____ rarely

5. As a group, students in trade theory courses demonstrate skill in efficient learning strategies and effective study skills

_____ always _____ sometimes _____ rarely

6. As a group, apprentice hires are lacking the necessary basic skills in (check as many that apply)

_____ reading _____ mathematics _____ communications
_____ other; specify _____

Direction, Part III

Please use the space below to comment further or explain any of your responses. Thank you.

Please return the completed survey to your employee development representative, or if requested, your group superintendent.

# RELATIONSHIP OF AN EXPERIMENTAL HISPANIC ENLISTMENT SCREENING TEST TO AFQT

John J. Mathews, Naval Training Systems Center

Carla M. French, University of Central Florida

Efforts are underway to increase the representation of Hispanics in the military. These efforts are not only founded on affirmative action goals but also in response to a decreasing pool of eligible youth. A large number of prospective recruits are excluded from military service due to low scores on the Armed Services Vocational Aptitude Battery (ASVAB). The potential aptitude level of many applicants is underestimated due to limited English skills of some recent immigrants and members of Hispanic subcultures. An aptitude test (Hispanic Enlistment Screening Test) written in Spanish has been developed by the Navy Personnel Research and Development Center (NPRDC) as an aid in identifying Hispanic youths whose ASVAB scores would likely be acceptable after remedial English training. This study is a preliminary evaluation of this test in the context of the Navy's Accession English Language Training (AELT) program. Participants are currently limited to high school graduates with English Comprehension Level Test scores at the marginal level of 45-80. They receive up to 22 weeks of AELT at the Defense Language Institute (DLI) prior to ASVAB retesting.

## METHOD

The subjects were Puerto Rican males who completed AELT and were retested on ASVAB by April 1985 (N = 205). Another 112 recruits were excluded from this study because they were still in training or had been discharged prior to ASVAB retesting. Predictor variables included scores on the experimental Hispanic Enlistment Screening Test (HEST) and the Armed Forces Qualification Test (AFQT) composed of subtests from the ASVAB. Like AFQT, the HEST has four subtests--Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), and Numerical Operations (NO). The ASVAB was administered to the subjects prior to Navy enlistment in Puerto Rico. The HEST was given at the DLI upon AELT entry. Those completing AELT (2 to 5 months) returned to Orlando to begin recruit training. Early in training, the subjects were retested on the ASVAB.

It is expected that as a result of AELT, posttraining AFQT scores, particularly on verbal subtests, will be higher than pretraining AFQT scores. The main hypotheses to be tested are that HEST scores will correlate higher with post-AFQT scores than will pre-AFQT scores and that HEST scores will add appreciably in predicting post-AFQT scores. Hispanic service applicant ASVAB data, supplied by the Air Force Human Resources Laboratory will be used to estimate reliability, correct for range restriction, and assess practice effects.

## RESULTS AND DISCUSSION

DESCRIPTIVE STATISTICS - Six subjects with post-AFQT percentile scores averaging 02, indicating extreme lack of motivation, were deleted from further analyses. This left a sample of 199. Computed means ($\bar{x}$) and standard

Table 1

Comparison of Means and Standard Deviations of
Pre and Posttraining ASVAB Measures

| Measure | | Pre | Post | t |
|---|---|---|---|---|
| AR | Mean | 41.9 | 42.6 | 1.7 |
| | SD | 4.8 | 6.0 | 3.3** |
| WK | Mean | 37.5 | 41.1 | 8.6** |
| | SD | 4.5 | 5.0 | 1.5 |
| PC | Mean | 38.0 | 43.3 | 7.7** |
| | SD | 6.9 | 7.0 | 0.2 |
| NO | Mean | 50.9 | 53.7 | 4.4* |
| | SD | 8.8 | 9.2 | 0.8 |
| AFQT (Centile) | Mean | 19.3 | 27.4 | 11.1** |
| | SD | 5.8 | 10.9 | 10.6** |

*Difference mostly due to practice effect.
**p less than .01, hypothesis that post is greater than pre.

Table 2

Sample Statistics for the Hispanic Enlistment Screening Test (HEST)

| Measure | Mean | SD | Number of Items | Percent Correct |
|---|---|---|---|---|
| AR | 27.9 | 5.2 | 35 | 80 |
| WK | 30.9 | 6.8 | 40 | 77 |
| PC | 13.2 | 3.4 | 20 | 66 |
| NO | 43.0 | 6.6 | 50 | NA* |
| Total | 115.0 | 14.2 | 145 | 79 |

*Not appropriate for speeded (NO) measure.

deviations (SD) for all variables are in Table 1. Subtests of AFQT are expressed in standard (T-score) metric with population mean of 50 and SD of 10. Initial ASVAB scores were low on all subtests except NO ($\bar{x}$ = 50.9). The lowest scores were on the most verbal of the subtests, WK and PC. These means were about 38, which corresponds to less than the 15th percentile. Pre-ASVAB SDs were considerably restricted due to the limited range (11-51 percentile) of AFQT in the sample. The mean AFQT percentile was 19.3. The raw score SD was 6.9. This is contrasted to an SD of 15.5 for a group of over 1,000 Hispanic service applicants (data from AF Human Resources Lab, 1984).

Retest scores were compared to pretraining scores via a test of correlated means (Guilford, 1965). Significant increases were found on the verbal subtests, WK and PC. Performance on PC was over five T-score points higher on retest. This indicates the AELT program had a measurable effect. The apparent increase on NO scores is misleading due to practice effects. Several studies (Sims & Maier, 1983, and unpublished study on Army applicants, 1982) have detected appreciable (about 2 T-score points) retest increases on NO but no increases on verbal subtests. When practice is statistically controlled, pre-post differences on NO are insignificant. There is no reason to expect AELT experiences to affect scores of nonverbal subtests such as NO. Differences between pre and postscore SDs were tested using a formula for correlated variances (Guilford, 1965). The only significant differences were increases on AR and AFQT.

Means for HEST subtests (Table 2) were high relative to the numbers of items. Percentages correct ranged from 66 on PC to 80 on AR. This indicates the subjects were generally able to comprehend and reason in written Spanish. Internal consistency reliability estimates were computed via Kuder-Richardson Formula 21. These estimates are shown in table 3 along with retest reliability data on like-named ASVAB measures for Hispanic applicants. The coefficients for corresponding HEST and ASVAB subtests are of comparable magnitude and are above .8 for AR and WK.

CORRELATIONAL ANALYSES - Correlations (r) of like-named HEST and pretraining ASVAB scores with posttraining ASVAB scores are given in Table 4. Low to moderate rs between pre and post-ASVAB scores were obtained. This was expected considering the range restrictions and low pretraining English verbal achievement in the sample. Correlations between similar pre and postsubtests were lowest for WK (.26) and PC (.15). In order to get an unrestricted population estimate of these rs, they were corrected for selection on AFQT (Guilford, 1965) using the raw score SD of 15.5 for Hispanic service applicants. Corrected rs for verbal scores were only .45 on WK and .30 on PC. Both uncorrected and corrected rs between HEST and corresponding posttraining ASVAB scores were higher than the rs involving pre-ASVAB on all subtests except NO. Only the comparison test (Guilford, 1965) for AR rs with postscores was statistically significant ( $p < .05$), however.

Uncorrected and corrected rs with post-AFQT scores were also higher for HEST AR, WK, and PC than for like-named pre-ASVAB subtests. Comparison tests for both WK (HEST r = .49 vs. .38 for pre) and PC (HEST r = .46 vs. .31 for pre) were significant.

Table 3

Reliability Estimates for Subtests and AFQT Total

| | ASVAB* (Retest) | HEST** (KR-21) |
|---|---|---|
| AR | .82 | .81 |
| WK | .85 | .89 |
| PC | .67 | .64 |
| NO | .71 | NA |
| Total | .90 | .89 |

*Based on Hispanic applicants for military service in 1983.
**Kuder-Richardson Formula 21 (Guilford, 1965).

Table 4

Correlations of Pre-ASVAB and HEST Subtests with Post-ASVAB Subtests and AFQT

| Variables | ASVAB Pre - ASVAB Post | | ASVAB HEST - ASVAB Post | | |
|---|---|---|---|---|---|
| | $r$ | $r_c$ | $r$ | $r_c$ | $t$ |
| AR vs. AR | .41 | .62 | .60 | .70 | 2.0* |
| WK vs. WK | .26 | .45 | .39 | .51 | 1.1 |
| PC vs. PC | .15 | .30 | .25 | .35 | 0.6 |
| NO vs. NO | .46 | .55 | .32 | .38 | ** |
| | ASVAB AFQT Pre - AFQT Post | | AFQT HEST - AFQT Post | | |
| | $r$ | $r_c$ | $r$ | $r_c$ | $t$ |
| AR vs. AFQT | .27 | .56 | .44 | .60 | 0.9 |
| WK vs. AFQT | .03 | .38 | .30 | .49 | 1.9* |
| PC vs. AFQT | .07 | .31 | .34 | .46 | 2.5* |
| NO vs. AFQT | .24 | .51 | .23 | .34 | ** |

*$p$ less than .05, hypothesis that HEST corrected $r$ ($r_c$) is greater than pre $r_c$.
**No difference hypothesized.

Multiple correlations (R) with post-ASVAB scores as criteria were computed using corrected intercorrelations. Table 5 shows results of tests for increases in R squared (Bottenberg & Ward, 1963) when different combinations of predictors are included. The best combination for predicting post-AFQT scores consisted of pre-AFQT, and HEST AR and PC (R = .71). With only subtests as predictors, the optimal set included pre-ASVAB AR and NO, and HEST AR and WK. Among HEST subtests, NO did not add appreciably to prediction of post-AFQT. Due to their high intercorrelation (.68), HEST WK and PC were virtually interchangeable as predictors.

Multiple Rs in predicting post-ASVAB subtest scores from like-named pre-ASVAB and HEST subtests are also given in Table 5. As hypothesized, each HEST subtest made a significant contribution to prediction of the corresponding ASVAB subtest. The smallest increase involved HEST NO. As expected, the verbal subtests were the least predictable (WK R = .55 , PC R = .41).

CONCLUSIONS AND RECOMMENDATIONS - If Hispanics with low English skills are given language training which improves AFQT scores, then the HEST appears to make a significant contribution in predicting posttraining AFQT performance. This contribution also seems to be of practical importance. The amount of AFQT variance uniquely accounted for by HEST is about 10 percent, which is an increase of 25 percent over the variance explained by pretraining AFQT scores (50 percent vs. 40 percent). The significance levels reported should be interpreted with caution because corrected correlations have inflated standard errors (Bobko & Rieck, 1980). This does not affect the magnitude of rs, however.

The HEST NO subtest did not add to the prediction function of AFQT scores. Possible reasons for this include: (a) the nonverbal content of NO, (b) the pre-ASVAB NO correlated .51 with post-AFQT, thus not leaving much margin for improvement, and (c) timing problems with this brief, speeded subtest.

Data are being collected on several hundred more Puerto Rican recruits. In addition to seeing if the results of this study replicate, optimal subtest weights, and cutoff scores on composites will be sought. Possible group administration effects, particularly on NO will be investigated. Unless its contribution dramatically increases, NO will be recommended for deletion from the HEST. Before recommending it for operational use, the HEST should be normed on a female and male group which includes Chicanos and Cuban-Americans. Given satisfactory validation of the HEST, it might be administered at selected processing stations to identify nonqualifying Hispanics who would likely attain acceptable AFQT scores after remedial English training.

Table 5

Results of Multiple Correlation (R) Comparisons
Using Corrected Intercorrelations

| Step | Predictors | Criterion | $R^2$ | R | F |
|---|---|---|---|---|---|
| 3 | AFQT pre, ARH, PCH | AFQT post | .50 | .71 | 5.8* (3 vs. 2) |
| 2 | AFQT pre, ARH | AFQT post | .48 | .69 | 20.3** (2 vs. 1) |
| 1 | AFQT pre | AFQT post | .40 | .63 | - |
| 4 | ARH, NO pre, WKH, AR pre | AFQT post | .46 | .68 | 5.3* (4 vs. 3) |
| 3 | ARH, NO pre, WKH | AFQT post | .44 | .67 | 9.7** (3 vs. 2) |
| 2 | ARH, NO pre | AFQT post | .40 | .63 | 12.2** (2 vs. 1) |
| 1 | ARH | AFQT post | .36 | .60 | - |
| 3 | ARH, WKH, PCH | AFQT post | .41 | .64 | 2.6 (3 vs. 2) |
| 2 | ARH, WKH | AFQT post | .40 | .63 | 9.9** (2 vs. 1) |
| 1 | ARH | AFQT post | .36 | .60 | - |
| 2 | AR pre, ARH | AR post | .53 | .73 | 38.8** (2 vs. 1) |
| 1 | AR pre | AR post | .38 | .62 | - |
| 2 | WK pre, ARH | WK post | .30 | .55 | 20.9** (2 vs. 1) |
| 1 | WK pre | WK post | .20 | .45 | - |
| 2 | PC pre, PCH | PC post | .17 | .41 | 13.7** (2 vs. 1) |
| 1 | | PC post | .10 | .32 | - |
| 2 | NO pre, NOH | NO post | .32 | .57 | 6.0** (2 vs. 1) |
| 1 | NO pre | NO post | .30 | .55 | - |

*p less than .05, hypothesis of increase in $R^2$.
**p less than .01, hypothesis of increase in $R^2$.

## REFERENCES


Bobko, P. & Rieck, A. (1980). Large sample estimates for standard errors of functions of correlation coefficients. Applied Psychological Measurement, 4, 385-398.

Bottenberg, R. & Ward, J. (1963). Applied Multiple Linear Regression (PRL-TDR-63-6). Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division.

Guilford, J. P. (1965). Fundamental Statistics in Psychology and Education (4th ed.). New York: McGraw-Hill.

Sims, W. & Maier, M. (1983). The Appropriateness for Military Applications of the ASVAB Subtests and Score Scale in the New 1980 Reference Population (Memorandum 83-3102). Alexandria, VA: Center for Naval Analyses.

# Attrition and Performance Ratings of ESL Soldiers in BT*

Harvey Rosenbaum

American Institutes for Research
Washington, D.C.

The Army's Basic Skills Education Program (BSEP) is an on-duty, job related, basic skills development program in the areas of literacy, math, and English-as-a-second language (ESL). One of BSEP's goals is to improve the ability of soldiers to perform satisfactorily in training and on the job. BSEP I programs are intended to provide soldiers with the basic competencies to complete Initial Entry Training (IET). BSEP II programs are to provide soldiers, E-1 through E-5, with the job related competencies necessary for their military duties. A fundamental notion underlying the BSEP program, for which there is some support (Krug, et al., 1984), is that soldiers with basic skill deficiencies are less likely to perform their job well, are less successful in the Army, and display a higher rate of attrition.

This paper presents additional support for the notion motivating BSEP using attrition data and performance rating data for a group of soldiers who completed the BSEP I ESL Program. The data show that level of English language proficiency is linearly related to attrition in Basic Training (BT) and tends to be related to drill sergeants' ratings of soldiers' BT performance. The attrition data also provide a basis for estimating the percent of attrition reduction in BT due to the ESL program.

## BACKGROUND

In the summer of 1982, the Army implemented a new six-week BSEP I ESL Program at eight TRADOC installations. The new program is functionally, or BT, oriented and intended to provide soldiers with information necessary for success in BT as well as improve their English language skills. The BT information was drawn from 25 tasks included in the Soldier's Manual of Army Testing or SMART book.

Between September 1983 and May 1984, the American Institutes for Research (AIR) conducted an evaluation of the new ESL program using a variety of program and follow-up measures (Rosenbaum and Stoddart, 1984). One of the pre/post measures used to assess the effect of the program on language proficiency was the English Comprehension Level Test (ECLT). The ECLT, developed by the

*This research was conducted for the U.S. Army Research Institute. Contract Nos. MDA-903-81-C-AA04 and MDA-903-84-C-0128. This paper does not necessarily express the official opinions or policies of the contracting agency.

Defense Language Institute English Language Center (DLIELC), is probably the most widely used English language test in the Armed Services. It is a standardized, group administered test of English comprehension with a strong reading component. Data gathered in previous studies by AIR (Holland, et al., 1982; Rosenbaum, 1983), indicate that as a measure of group performance, the ECLT is an effective indicator of oral comprehension and speaking proficiency.

Eligibility for the ESL program is largely determined by the soldiers' ECLT score. Soldiers with ECLT scores below 70 are automatically eligible for the program.

Included in the follow-up measures used to determine program effects, were data on BT attrition and performance ratings completed by drill instructors on an available sample of soldiers in BT. Data presented in the following sections were gathered as part of AIR's overall evaluation of the ESL program.

## METHOD

### Subjects

ECLT scores and BT attrition data were available for 582 soldiers. Supervisors' ratings of soldiers' performance in BT were also obtained for an available sample of 156 soldiers who had completed the program. Nearly 90 percent of the soldiers were native speakers of Spanish and more than 70 percent of these were from Puerto Rico. Approximately 75 percent of the sample graduated from high school and nearly 25 percent graduated from two- or four-year colleges.

### Procedure

Soldiers who appear potentially eligible for the ESL program are usually administered the ECLT when they arrive at their training base. Those with ECLT scores below 70 are usually placed in the program within a week or two of their ECLT testing. Soldiers are retested with the ECLT when they complete the program.

Installation education centers provided AIR with information on whether their students completed BT or were discharged from BT. Most education centers, however, do not routinely receive this information from the BT units and must make special efforts to obtain it. One education center did not provide any data on attrition and most of the other centers provided attrition data on only part of their student enrollment. Consequently, we were only able to obtain attrition data on a portion of the nearly 1800 soldiers involved in the overall evaluation.

Performance ratings were obtained for an available sample of soldiers during their last four weeks of BT. AIR researchers had drill sergeants rate soldiers' performance on 14 BT activities by

comparing their performance with all other soldiers. All activities required the use of language and were taken from the Program of Instruction for BT. The descriptions of the activities were pilot tested with drill sergeants at Fort Dix to insure that they were understandable and relevant. The rating system was a four-point scale with the values of better than most, as well as most, not as well as most but gets by, and performs inadequately. Examples of the BT activities are "responds correctly to questions," "uses challenge and password" and "reads SOP for inspections."

## RESULTS

### Attrition

Soldiers who were discharged from the Army during BT displayed poorer performance on the ECLT than those who completed BT as shown in Table 1. Discharged trainees had lower entry ECLT scores, lower exit ECLT scores, and made smaller ECLT gains.

Table 1
Comparison of Mean ECLT Scores

| Attrition category | Entry ECLT | Exit ECLT | ECLT gain |
|---|---|---|---|
| Completed BT | 42.9 | 57.6 | 14.8 |
| (n) | (505) | (507) | (505) |
| Discharged from BT | 37.4 | 46.3 | 9.8 |
| (n) | (78) | (75) | (75) |

The attrition rate is linearly related to exit ECLT scores with soldiers scoring below 30 having an attrition rate more than five times that of soldiers scoring above 69 as shown in Table 2.

Table 2
Comparison of Soldiers Completing BT with Those Who Were Discharged from BT

| | Exit ECLT score | | | | | | |
|---|---|---|---|---|---|---|---|
| Attrition Category | 0-29 | 30-39 | 40-49 | 50-59 | 60-69 | >69 | n |
| Completed BT | 64% | 75% | 84% | 91% | 92% | 93% | 507 |
| Discharged from BT | 36% | 25% | 16% | 9% | 8% | 7% | 75 |
| n | 36 | 63 | 95 | 123 | 119 | 146 | 582 |

## Performance Ratings

Sergeants tended to assign a soldier similar performance ratings for most of the 14 activities. Two sets of Pearson Product Moment correlations were conducted to determine the degree of similarity between the activity ratings. In order to conduct these correlations, ratings were given numerical or score values by assigning a value of one through four to the activity ratings: the value of one was assigned to performs inadequately, two was assigned to not as well as most, etc.

Soldiers' performance ratings for each activity were then correlated with each of the remaining 13 activities yielding 91 statistically significant correlations between .51 and .81. For the second set of correlations, each activity rating was correlated with the average of all 14 ratings for each soldier yielding a correlation range of .77 to .86.

The strong correlation of the 14 activity ratings justifies using the soldiers' average ratings as unitary performance ratings. To relate soldiers' average performance ratings to their exit ECLT scores, soldiers were divided into two categories: average performance ratings of 2.4 or less were assigned to the performs not as well as most category, and average ratings of 2.5 or more were assigned to the as well as most or better category.

Soldiers with lower exit ECLT scores tended to receive lower performance ratings. For example, soldiers with ECLT scores below 60 are nearly three times as likely to be in the performs not as well as most category than are soldiers with higher ECLT scores as can be inferred from Table 3. However, the distribution of soldiers in the category performs not as well as most by exit ECLT score is not clearly linear. Two factors that may be affecting these data are (1) the sample size below 50 ECLT is small, and (2) BT performance ratings may be less directly related to language proficiency. During interviews with AIR researchers, drill sergeants often stated that a soldiers' attitude and motivation are major factors in completing training.

Table 3
Comparison of Soldiers Rated as Performing Not as Well as Most With Soldiers Receiving Higher Ratings

| Performance Rating Category | Exit ECLT | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0-29 | 30-39 | 40-49 | 50-59 | 60-69 | >69 | n |
| As well as most or better | 67% | 50% | 79% | 69% | 89% | 90% | 124 |
| Not as well as most | 33% | 50% | 21% | 31% | 11% | 10% | 32 |
| n | 9 | 12 | 19 | 32 | 36 | 48 | 156 |

## DISCUSSION

Level of English proficiency is clearly a factor in soldiers' success in BT. The data indicate that limited proficiency in English contributes to attrition and poorer performance. Because it increases soldiers' proficiency in English, as measured by the ECLT, the BSEP I ESL program can be assumed to be reducing attrition and improving soldiers' performance in BT. The sample of nearly 1800 soldiers that completed the ESL program during the evaluation period showed a mean ECLT gain of 15.0. This mean gain is as good or better than the mean ECLT gains obtained from previous Army ESL programs.

The reduction in BT attrition attributable to the ESL program can be estimated by comparing the attrition of soldiers completing the ESL program with a hypothetical group of eligible soldiers receiving BT without the benefit of the program. Computing this estimate requires one conservative assumption in addition to the available data. The assumption is that the attrition rate at each exit ECLT level is also applicable for soldiers who have not participated in the ESL program. Even if incorrect, this assumption is most likely to result in an underestimate of the reduction in attrition due to the ESL program since soldiers in the attrition sample are different from a hypothetical group of new trainees with matching English proficiencies. The sample group started BT already knowing a great deal of technical information and should find BT easier than the hypothetical group. Interviews with soldiers during the BT follow-up phase of the evaluation support this view. Many soldiers stated that BT was easy because of the information they had learned in the program, others claimed that they would not be able to complete BT without the program.

We can estimate the BT attrition for 1000 eligible soldiers who were not enrolled in the ESL program and compare this with the attrition estimate for 1000 soldiers enrolled in the program. The process of estimating the attrition for 1000 not enrolled soldiers is presented in Table 4. Line one shows the pre-program distribution of ECLT scores for the sample of 1762 soldiers used in the ESL program evaluation. The next line indicates the number of soldiers per thousand at each ECLT range. The third line provides an index of attrition for each ECLT level based on the percent of attrition for that level as given in Table 2. The final line in Table 4 indicates the number of soldiers estimated to attrite at each ECLT level yielding a total of 185 per 1000.

Table 4
Procedure for Estimating BT Attrition of 1000 Eligible Soldiers Not Enrolled in ESL Program

| | ECLT Levels | | | | | |
|---|---|---|---|---|---|---|
| | 0-29 | 30-39 | 40-49 | 50-59 | 60-69 | Total |
| Distribution of SM by entry ECLT | 18% | 22% | 18% | 21% | 21% | 100% |
| No. of SM/1000 | 180 | 220 | 180 | 210 | 210 | 1000 |
| Index of attrition | .36 | .25 | .16 | .09 | .09 | - - |
| No. of SM attriting | 65 | 55 | 29 | 19 | 17 | 185 |

Using the same procedures, Table 5 shows the basis for the attrition estimate for 1000 soldiers enrolled in the program yielding an attrition total of 116. According to these estimates, the Army's ESL program saves 69 out of every 1000 soldiers from attrition in BT or reduces BT attrition by 37 percent.

Table 5
Procedure for Estimating BT Attrition of 1000 Soldiers Enrolled in ESL Program

| | ECLT Levels | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0-29 | 30-39 | 40-49 | 50-59 | 60-69 | >69 | Total |
| Distribution of SM by exit ECLT | 3% | 9% | 15% | 20% | 26% | 27% | 100% |
| No. of SM/1000 | 30 | 90 | 150 | 200 | 260 | 270 | 1000 |
| Index of attrition | .36 | .25 | .16 | .09 | .08 | .07 | 1.00 |
| No. of SM attriting | 11 | 23 | 24 | 18 | 21 | 19 | 116 |

## REFERENCES


Holland, V.M., Rosenbaum, H., Stoddart, S.C., & Redish, J.C. BSEP I/ESL programs - Volume one: Findings. Washington, D.C.: American Institutes for Research, October 1982.

Krug, R.E., Hahn, C.P., & Wise, L.L. Review of BSEP I and BSEP II programs. Task One Report. Washington, D.C.: American Institutes for Research, July 1984.

Rosenbaum, H., Hahn, C.P., and Holland, V.M. Testing the English language proficiency of soldiers who are not native English speakers. Paper presented at the 25th Annual Conference of the Military Testing Association, Gulf Shores: Alabama, October 1983.

Rosenbaum, H., & Stoddart, S.C. Evaluation of the Functional Pre-BT ESL Course. Washington, D.C.: American Institutes for Research, December 1984.

# STANDARD SETTING METHODS

# FOR SKILL QUALIFICATION TESTS (SQTs)

Dr. Allan L. Pettie

United States Army Training Support Center, Fort Eustis, Virginia

The purpose of this paper is to examine three different methods for setting passing scores for Skill Qualification Tests (SQTs). The Skill Qualification Testing Program is one component of the United States Army's Individual Training Evaluation Program. The other components, the Common Task Test and the Commander's Evaluation, are hands-on tests. The SQT component is a task-based, performance-oriented objective test. Enlisted soldiers are tested by Military Occupation Speciality and Skill Level.

Since the implementation of SQTs, the Army has used the same passing score of 60 for all SQTs. Experience has shown that the adoption of this common standard, which failed to consider test difficulty and content, and its application across the broad spectrum of diverse MOSs was unjustified. Starting October 1, 1986, this policy of a common standard for all SQTs will be discontinued. A separate passing score, based upon information about that SQT, will be set for each SQT. Prior to the implementation of this change, pilot studies of three methods, which were deemed applicable to the Army, were conducted. These three methods (untrained examinee, performer non-performer, and validation data) are described briefly in the following sections.

## UNTRAINED EXAMINEE METHOD

The rationale for the setting of a passing score by this method is that the trained soldier should be able to do at least as well as the atypical untrained examinee. Factors, not related to competence, but which allow the incompetent to pass, such as test wiseness, the effects of tests which measure primarily general aptitude, reading ability, and common sense could be minimized by this method. Scores of untrained examinees could be indicative of the extent to which the SQT measures general aptitude and reading ability.

The untrained examinees of choice were soldiers in the same or related MOS who were in the first week of Advanced Individual Training. It was presumed that these examinees would be similar in background and representative of the trained group before training. After the SQT was administered, the passing score was set at the eightieth percentile. A soldier who scored above the eightieth percentile was considered to be atypical of the untrained group; thus, minimum competence is defined on the basis of the untrained examinee.

## PERFORMER NON-PERFORMER METHOD

This adaptation of the contrasting group method (Livingston and Zieky, 1982) relies upon the judgment of supervisors to form a performer group and a non-performer group. Supervisors were instructed to base their judgments of MOS competence on criteria other than test scores. Costs of classification errors, performers who fail and non-performers who pass, were considered to be equally serious. The passing score was determined to be the point of intersection for the performer and non-performer distributions.

## VALIDATION DATA METHOD

Out of a desire to take advantage of existing soldier tryout data, this method was formulated. Soldier tryout data are normally collected at the task level, a mini-test of three or more questions. Soldiers, through self-ratings or supervisory ratings, are designated performers or non-performers on the particular task. Since the data for a performer group, consisting of the same performers, and a non-performer group, consisting of the same non-performers, do not exist, a composite performer group and a composite non-performer group were formed.

The performer group, actually task-level performers, thus resulted in testees who were performers on all tasks. Similarly, the non-performer resulted in a testee group which was made of non-performers on all tasks. Variances were calculated at the task-level and then summed across tasks. Since the scoring for SQTs is task-based, an estimate of the standard deviation was calculated by an adaptation of standard deviation estimation by items (Tinkleman, 1971).

Similar to the performer non-performer method, costs were equalized, and the passing score was set at the point of intersection for the performer and non-performer distributions.

## RESULTS

Since the pilot studies were conducted by the untrained examinee and validation data methods or the performer non-performer and validation data methods, passing scores set by the method pairs are presented in Table 1.

TABLE 1

PASSING SCORES FOR UNTRAINED EXAMINEE, PERFORMER NON-PERFORMER AND VALIDATION DATA METHODS

| SQT | UNTRAINED EXAMINEE | VALIDATION DATA | PERFORMER NON-PERFORMER |
|---|---|---|---|
| 15B1 | 36 | 61 | |
| 16P1 | | 48 | 72 |
| 16R4 | | 63 | 63 |
| 24G1 | | 70 | 70 |
| 24G3 | | 67 | 67 |
| 24A1 | 50 | 74 | |
| 24H3 | 50 | 76 | |
| 27E1 | | 61 | 53 |
| 27E3 | | 63 | 64 |
| 31V1 | 55 | 63 | |
| 31V2 | 58 | 65 | |
| 31V3 | 50 | 65 | |
| 31V4 | 49 | 67 | |
| 42E3 | 38 | 60 | |
| 43E1 | 42 | 68 | |
| 43E3 | 47 | 64 | |
| 64C1 | 73 | 63 | |
| 64C3 | 81 | 61 | |
| 71Q3 | 41 | 56 | |
| 71Q4 | 41 | 53 | |
| 71R1 | 55 | 67 | |
| 71R2 | 43 | 67 | |
| 71R3 | 47 | 61 | |
| 71R4 | 47 | 61 | |
| 73D1 | 64 | 62 | |
| 73P3 | 62 | 64 | |
| 75B2 | 40 | 61 | |
| 75B3 | 31 | 57 | |
| 76P1 | 39 | 61 | |
| 76P3 | 43 | 57 | |

Passing scores set by the untrained examinee method demonstrate greater range than the other two methods. The 75B2 SQT has the lowest passing score of 31, and the 64C1 SQT has the highest passing score of 81. Passing scores for the validation data method ranged from a low of 48 for 16P1 SQT to a high of 76 for the 24H3 SQT. Passing scores for the performer non-performer method ranged from a low of 53 for the 27E1 SQT to a high of 72 for the 16P1 SQT.

Comparisons of the untrained examinee method to the validation data method show that except for four SQTs (64C1, 64C3, 73D1, and 73D3) passing scores for the untrained examinee method are lower than the validation data passing scores. Passing scores set by the validation data and the performer non-performer methods, with the exception of 16P1 and 27E1, are very close in value. Upon analysis of the performer non-performer data, a possible explanation for the discrepancy was discernible. This data consisted of seven cases, four 25's and three 75's, for the non-performer group. Perhaps, the three soldiers who scored 75 were misclassified. In any event, the non-performer group appeared to be widely disparate in ability. No expalantion was discovered for the 27E1 discrepancy.

## CONCLUSIONS

There was no substantial agreement or pattern of disagreement between the passing scores set by the untrained examinee and the validation data methods. Untrained examinee passing scores were generally lower and considerably different from the validation data passing scores. The discrepancy between the two estimates ranged from a difference of two points (73D1 and 73D3) to a 36 score point difference (75B3). Given the different rationales for the two approaches, this lack of agreement should be expected. The crucial question is which rationale is best applicable to the Army's needs.

Better agreement was reached for the validation data and the performer non-performer methods. For four of six SQTs no difference was observed between the passing scores. Considerable differences were noted for two SQTs. A possible explanation for the 16P1 difference was noted above, but the 27E1 difference of eight score points is troublesome.

The validation data method was developed as a substitute for the performer non-performer method. It appears that the validation data method may be a viable alternative to the more resource intensive performer non-performer method. The validation data method is less resource intensive, because of the collection of the data does not represent a new requirement. The

validation data method is attractive, because of the existing requirement for the collection and maintenance of the data and because it represents no new requirement upon the test developers.

Examination of the setting of passing scores will continue along three different lines. Few pilot studies were performed for the performer non-performer method; therefore, further analysis with more SQTs needs to be conducted to examine the viability of the validation data method. Another form of analysis will be the application of multiple matrix sampling estimation procedures to estimate the mean and standard deviation for the performer non-performer groups from soldier tryout data. Although the random assignment of examinees to items assumption of this method will be violated, the application of this method should be examined. Finally, the proximity of the passing score to an acceptable level of competence and the reasonableness of passing scores set by the untrained examinee method will be examined.

## References

Livingston, Samuel A., and Zieky, Michael. Passing Scores. Princeton, N. J.: Educational Testing Service, 1982.

Tinkleman, Sherman N., Planning the Objective Test. In R. L. Thorndike (ed.). Educational Measurement. Washington, D. C., Americal Council and Education, 1971, pp 65-66.

# Assessing Tank Commander and Gunner Proficiency on U-COFT

Scott E. Graham
and
John A. Boldovici

Army Research Institute-Fort Knox Field Unit

The Armor community is striving to improve its selection and training of M1 Tank Commanders (TC) and gunners. If these goals are to be accomplished, valid performance measures must be established which assess the full range of tank crewmen duties. One of the most pressing needs in Army evaluation is the development of tests whose psychometric properties are known and which inspire confidence (Boldovici & Sabat, 1985).

The reliable measurement of hard, combat-oriented skills and soft, leadership skills has not been successfully accomplished. The lack of standardized and stable performance measures makes it difficult to compare individual and crew performance across time and units. As a result, NCO promotions may be yoked to Time in Grade rather than be performance based. Inconsistencies in training evaluation research have also resulted from the lack of valid performance criteria. A number of criterion measures have been developed for assessing gunnery-related performance. Continuing credence is, however, placed on scores obtained from live-fire gunnery exercises, and in particular, Table VIII. Unfortunately, problems affecting the reliability of these scores, e.g., varying weather and equipment conditions, make questionable the comparison of scores across days, ranges, and units.

Training simulators and other electronic training devices provide new opportunities for evaluating the performance of Armor crewmen. Their main advantages include precise presentation of target conditions with accurate scoring and timing. Graham and Black (1985) have identified a number of critical TC tasks which can be evaluated on the newly developed M1 Unit-Conduct of Fire Trainer (U-COFT). These target engagement tasks include laying the main gun, issuing fire commands, boresighting and degraded mode gunnery.

The research described here develops and evaluates a U-COFT proficiency test for M1 gunners. The primary goal is to assess the psychometric characteristics of the U-COFT as a testing device and to evaluate the feasibility of testing gunners alone, i.e., independent of TC performance.

## Method

U-COFT Test Development. The U-COFT Gunner's Test contained eight shortened exercises, each with four engagements, selected from the U-COFT's TC/Gunner's training matrix. One target was friendly (an M2) making a total of 31 target engagements. The engagements were selected to match target conditions found in Table VIII of the M1 tank combat tables. For example, half of the engagements in each were ownvehicle stationary (or moving), and half were long (short) range. Roughly equivalent numbers of single/multiple and stationary/moving targets were included. Table VIII has 20% battlesight engagements (as opposed to precision or "full up") while the U-COFT test has 25%. Half of the engagements also required thermal sights.

Scoring. A number of performance measures were obtained from each engagement. These included Hitrate which was defined as the number of hits divided by the number of targets presented. Other measures included First Round Hitrate, Azimuth and Elevation errors, Target Identification (ID) time,

which was the time from when the target appeared until the gunner said "identified," and Opening time, which was the time from target appearance until the first round was fired.

The U-COFT software package reports three composite performance scores which were also recorded. The Target Acquisition score measures target acquisition time and identification/classification errors (U-COFT Handbook, 1985). The Reticle Aim score assesses opening time, time to kill, and reticle aim error. Lastly, the System Management score counts pre-firing switch errors, ammunition errors, and excessive ownvehicle exposure times. Each of the scores is reported as a letter grade, A, B, C, or F with corresponding numerical values of 4.0, 3.0, 2.0, and 1.0.

Participants. The U-COFT gunner's test was administered at the completion of other research evaluating training transfer between the U-COFT and the MK-1 videodisc gunnery simulator (Witmer, in prep). The 32 soldiers used were M60A3 loaders and drivers from the 194th Armor Brigade at Ft Knox, KY. The majority had ranks of Private First Class. The soldiers, with few exceptions, had not served as gunners other than in Advanced Individual Training, and had no experience with the M1 tank.

Three TCs were predominately used with a fourth TC used for one session. One civilian TC was an ex-General Electric employee who had hundreds of hours of U-COFT experience as a U-COFT TC (COFT-experience). The other two were a Sergeant First Class from the Armor School Weapons Department (Sr NCO) and a Sergeant from the 2/6 Cavalry (Jr NCO), a training support unit of the Armor School. They had no prior U-COFT experience.

Procedure. The gunners received 1 1/2 hours of U-COFT training during the U-COFT/MK-1 transfer study. The U-COFT gunner's test began with eight practice engagements. The last four required use of the Gunner's Auxiliary Sight (GAS) as there was simulated failure of the laser rangefinder, stabilization system, Gunner's Primary Sight and computer system. The TC trained the gunner on use of the GAS and how to fire with manual lead and elevation.

The eight subtests were sequentially presented with a short pause between each. During this time, the U-COFT Instructor/Operator (I/O) had to terminate the standard 10 engagement exercise, dump the printouts, and enter the six-digit code for the next subtest. This procedure was awkward and a few printouts were missed. Following a break, a retest was presented, which consisted of a different subtest order. The practice, test, and retest took approximately 2 hours.

An attempt was made to minimize the effects of differential TC performance by having the I/O talk the TC onto the target. The I/O might have said, for example, "next target, a T-72, is left in 10 seconds." This modification theoretically reduced the variability of target identification times across TCs and minimized fire command errors. The standard U-COFT procedure also requires the gunner on defensive engagements to move his head out of the GPS, check the GAS to see that the gun has cleared the berm, say "driver stop", and then go back to the GPS. This procedure was omitted.

## Results and Discussion

For each of the dependent variables, test-retest reliability coefficients (Pearson r) were computed. The reliability coefficient for Hitrate (.80), First Round Hitrate (.72), Target ID Time (.87), Opening Time (.22), and the U-COFT Target Acquisition (.76) and reticle aim (.83) scores are encouraging.

Demonstrated reliability is a necessary component for valid tests and these U-COFT measures appear appropriate. The next logical step is to validate the measures against known and acceptable performance criteria, i.e., combat effectiveness measures. The typical inclination is to validate device-mediated gunnery tests against live-fire gunnery scores. Unfortunately, the poor psychometric properties of the live-fire measures virtually assure a weak relationship at best. The validation and acceptance of U-COFT tests will likely result from its convergence with various gunnery performance variables, sound military judgement, and face validity.

The data show poor reliability for Azimuth (.42) and Elevation (-.07) errors and the U-COFT System Management scores. The elevation and azimuth error unreliability appears to result from a tight distribution with some extreme scores. The mean elevation error of .57 mils is within the hit range of most targets, yet some individual subtest elevation errors exceeded 7.0. The U-COFT Reticle Aim score reduces the effects of extreme errors by scoring an "F" for any missed target. While the score does reflect time in addition to lay error, the reliability for Reticle Aim score is .83. This suggests Reticle Aim could be a reliable measure.

The unreliability of System Management errors (.11) may be artifactual of the U-COFT procedure. The gunners were instructed to leave the gun select switch on "main gun". On defensive engagements when the crew was "killed" as the result of being exposed too long, the gun select switch automatically reset to safe. Random switching errors may have then resulted from failure to put the switch back on main gun. Other system management errors resulted from incorrect ammo select switch settings.

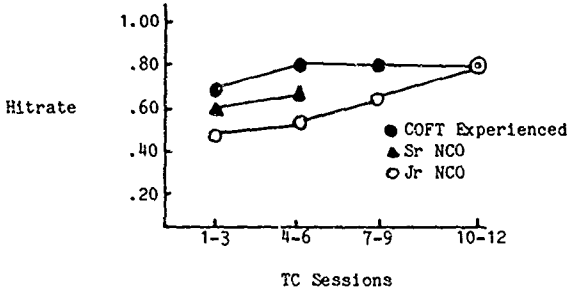
The feasibility of using U-COFT as a testing device depends, in part, on its ease of the administration and scoring. Three U-COFT scores, Reticle Aim, Target Acquisition, and System Management, could well be used as performance measures. The composite mean of these scores yielded a test-retest reliability of .82, and correlated .87 with Hitrate. These data substantiate the internal validity of the U-COFT scoring procedure.

Performance on the eight subtests varied considerably. Mean Hitrate for single long range stationary targets at night was .91, while the hitrate for multiple short range moving targets in degraded mode was only .19. Multiple and moving targets, not surprisingly, were most difficult, although the hitrate for short range single moving targets was .83.

One minor difficulty with the U-COFT for training and testing is that it includes dispersion rounds. A gunner may have a perfect sight picture, fire within the required time limits, and still miss the target. The opposite also occurred, although less frequently. While the dispersion rounds were likely included to match live ammunition characteristics, they result in greater unreliability of testing scores, and random bad feedback for trainees. Software updates should include the capability of turning off the dispersion rounds.

TC Performance Effects. Despite attempts to minimize TC effects, U-COFT performance differed as a function of TCs, i.e., COFT-experienced, Sr NCO, and Jr NCO. Gunners paired with the COFT-experienced TC had a combined test and retest Hitrate of 74%, while those paired with the Sr NCO shot 64%, and those with the Jr NCO shot 63%, $F(2,28) = 3.89$, $p < .05$. Similar significant differences for TCs were found for First Round Hitrate, ID time, and the Reticle Aim Target Acquisition, and System Management scores.

The data were recomputed to reflect these changes over sessions. The figure below shows changes in Hitrates for the three TCs over the duration of the experiment. Each TC session reflects the mean performance of three different gunners.




Hitrate is ostensibly a measure of gunner performance as the gunner makes the final lay of the gun and pulls the trigger. Hitrate is shown here to be also a function of TC performance. The first three gunners paired with the Jr NCO hit 48% of the targets while his last three gunners hit 77%. Each session gave the TC about 3 1/2 hours of U-COFT experience, including the U-COFT time in Witmer's experiment. Three TC sessions, therefore, correspond to about 10 hours of U-COFT time. The abscissa in the figure could alternatively be labled 10, 20, 30 and 40 hours of TC COFT time. These data show that the COFT-experienced TC's contribution to hit performance asymptoted between 10 and 20 hours, while the Jr NCO's contribution had not stabilized at 40 hours.

Several factors may be contributing to this pattern. First when target ID time is plotted as a function of TC sessions, improvement in ID time for the three TCs essentially mirrors Hitrate performance. Secondly, the TCs might becoming better trainers of the novice gunners over sessions. The TCs were learning what errors were typically being made by the gunners and were better instructing the later gunners in these areas, e.g., use of the GAS. While the observations are subjective and non-systematic, improvement in training was particularly noticeable for the Jr NCO. At first he nervously barked his fire commands, and paid little attention to the gunner. In later sessions, he readily assisted the I/O in the U-COFT orientation and praised the gunners during breaks. General Electric officials independently reported a similar phenomenon during their COFT tests. TCs and gunners were at first blaming each other for misses, but over time greater cooperation developed.

The results suggest at least three things. First, under the conditions used in this experiment, the U-COFT is not appropriate for testing gunner proficiency alone, as the gunner's performance was not independent of the TC. Confederate TCs might first be trained to stable performance levels, but this would require considerable time and resources, e.g., 40 plus hours. If a gunner alone test is desired, the soon-to-be-fielded Institutional-Conduct of Fire Trainer (I-COFT) may be more appropriate. The I-COFT will have essentially the same hardware as the U-COFT but software modifications will permit voice synthesized fire commands. The U-COFT has been developed as a TC/gunner training device, and for the most part, is better suited for testing the proficiency of TC/gunner pairs.

Second, the results reinforce the notion that crew performance is largely a function of the TC's ability to train his crew. More time might well be spent on training the TC how to train his crew.

Third, U-COFT performance results from a number of factors, in addition to the gunner's ability to quickly and accurately lay the gun. Some of the factors may be idiosyncratic to the COFT itself and not related to live-fire gunnery performance. The improvement of the Jr NCO over 40 hours of training may have resulted from a growing familiarity with the device and the computer-generated graphics, and not from changes in his gunnery ability. Had the TC been going through the U-COFT training matrix, he would likely have been half way through the matrix before his performance stabilized. Improvements in performance might be unrelated to the systematic progression of target conditions, but of general familiarity with the machine. The built-in training package is, nevertheless, better than most other training devices which have none. The validity of the U-COFT as a testing device is threatened by the fact it takes so long to get stable performance.

Modeling Gunnery Performance. The precise target presentation and response recording capabilities of the U-COFT result in accurate part-task gunnery data, which can be used in the development and refinement of combat models. Various attrition-rate models exist with factors such as time to acquire target and time to fire following a hit (or miss). The value of these complex models has been limited, however, by poor parameter estimates. The estimates have largely come from past battle data which are sketchy at best, as armies engaged in war are interested in a number of factors in addition to gathering quality data for combat models. The U-COFT has the capability of gathering reliable estimates under a number of conditions, including future battlefield conditions. As such, the U-COFT can be used to evaluate alternative future training strategies.

In the present research, stepwise regression analyses were performed to help understand what factors were underlying the various performance measures. Hitrate was found to be primarily determined by reticle aim accuracy, as target ID time and Opening time failed to significantly load into the equation. In the previous discussion of factors affecting changes in hitrate over TC sessions, changes in target ID time were shown to parallel changes in hitrate. The regression analysis suggests, however, that changes in target ID time are unrelated to changes in Hitrate, once the effects of aiming error are removed. This finding lends additional credibility to the explanation that improvement over TC sessions resulted from improved training of the gunner.

## General Discussion

The results demonstrate the U-COFT holds great potential as a device for assessing gunnery proficiency, although not necessarily for gunner-alone tests as tried in this research. Test-retest reliability coefficients exceeding .80 were found for a number of performance measures including Hitrate and target ID time. These measures are well within the acceptable psychometric range for military tests. In addition, U-COFT tests might be used as criteria against which other less expensive or portable tests could be validated.

The U-COFT proficiency test developed for this research mirrored target conditions in the M1 tank Table VIII. This may not be necessary. As Boldovici (1979) noted, the lowest level of "enabling skills" required for

gunnery performance, e.g., psychomotor skills and system procedural knowledge, are highly redundant across engagement conditions. As a result, it is unlikely that a crew who is relatively good at long range moving targets at night would be poor at short range stationary targets in daylight.

The high Hitrate for single stationary targets further demonstrates the relative ease of using the M1 fire control system when it is fully operational. Considering that the experimental gunners were M60A3 loaders and drivers with little gunners' experience, a ceiling effect would be expected if the easier target engagements were used with experienced TC/gunner pairs. Perhaps the best feature of the U-COFT is its ability to train and test under various degraded conditions. Future U-COFT tests might place even greater emphasis on evaluating degraded gunnery performance, e.g., manual control conditions.

Improved evaluation of tank gunnery skills with the U-COFT can lead to a stronger Armor force. Skills and abilities other than those measured by the U-COFT are, however, equally important and should not be overlooked. Graham and Black (1985) found that soft skills, e.g., ability to train, ability to communicate, and general leadership qualities, were predominately identified as the distinguishing characteristics of TC excellence. Likewise, the results described here show the TC's ability to train his gunner has a large effect on gunnery performance. Continuing development of evaluation batteries which assess both hard combat-oriented skills and soft leadership skills is needed, if the force is truly to become an Army of excellence.

## References


Boldovici, J.A. (1979). Analyzing Tank Gunnery Engagements for Simulator-Based Process Measurement. Army Research Institute Research Report 1227.

Boldovici, J.A. & Sabat, S.R. (1985). Measuring Transfer from Training Devices to Weapon Systems, Paper presented at NATO Symposium on the Transfer of Training to Operational Military Systems, Brussels, Belgium.

Graham, S.E. & Black, B.A. (1984). Defining and Assessing Tank Commander Excellence, Army Research Institute Research Report 1401.

Unit-Conduct of Fire Trainer Instructor's Utilization Handbook (1985). General Electric Company, Daytona Beach, FL.

Witmer, B.G. (in Prep). Unit-Conduct of Fire Trainer (U-COFT) and Videodisc Gunnery Simulator (VIGS) Cross-Training and Training Transfer. ARI-Ft Knox Field Unit Working Paper.

# Criterion Referenced Testing for the U.S. Navy's Nuclear Submarine Fleet

Jeffrey A. Cantor and Lee Walker
DDL-OMNI Engineering
McLean, Virginia

## Introduction

The design and development of criterion-referenced System Achievement Tests (CR SATs) for each of the Naval Enlisted Classifications (NECs) associated with the Fleet Ballistic Missile (FBM) Strategic Weapon System (SWS) Training Program is a goal of the U.S. Navy's Strategic System Program Office (SSPO). DDL-OMNI Engineering has been involved in the design, development and refinement of a method to accomplish this goal over the past two years. The methodology which will be described in this paper presentation has evolved as a result of the planning, trial and refinement on the part of both SSPO and DDL-OMNI Engineering.

## Background

The CR SAT is a test which is administered to all SWS technicians periodically to measure overall system knowledge and skill levels based on the specified jobs/tasks and procedures identified as essential for each NEC. The primary purpose of the SAT is to evaluate the readiness of the personnel to operate the equipments within their NEC, so as to accomplish the ship's mission; its secondary purpose is to assess the knowledge and skill of the personnel tested. Previous efforts at a test design specification for a SAT had generally been in terms of the format and content of a previous version of the SAT within a given NEC. This procedure did not allow for control in terms of overall coverage of all requisite jobs/tasks and procedures, nor did it allow for updating of information coverage necessary as a result of changes in technology of the equipment and changes in procedures. The new CR SAT program was initiated to allow for such changes.

## Test Design Specification

The basis for the construction of a CR SAT is the test design specification. This specification is an engineering blueprint of the knowledges identified as necessary to possess in order to effectively perform as a competent FBM/SWS technician in the specific NEC for which the CR SAT is written.
The project undertaken to produce a test design specification allows for inputs to ensure adequate content coverage and changes in technology and at the same time, be self-sustaining in a dynamic NAVY personnel system. The project team includes contractor subject-matter experts, training data analysts, and competent and knowledgeable NAVY technicians and supervisors representative of the NEC under development. The methodology devised for the test design specification is predicated on the use of a committee of technical experts with current knowledge of the jobs/tasks and procedures necessary for successful performance in the NEC for which the specification is to be constructed.

The contractor project leaders, in preparation for the test design specification workshop, prepare a comprehensive listing of those jobs/tasks and procedures identified as necessary to the function of the NEC. This listing of jobs/tasks and procedures comes from applicable Personnel Profile

Tables (PPPs), Training Level Assignments (TLAs), and other procedural documentation literature which supports the equipment systems operated by the respective NEC. These listings are then prepared in the categories of the applicable equipment system/subsystem or to the applicable area of administration/security, casualty operating procedures, or watch qualifications. The basic test will include approximately 360 question items, divided into sections of equipments (approximately 270 items), administration-/security (30 items), watch qualifications (30 items), and casualty operating procedures (30 items).

Test Design Workshop Process

The workshop takes the form of a five-day convening, with technician representatives from each of the Submarine Groups. Since two levels of tests will be constructed for each NEC, one at the watchstander level and one at the supervisor level, personnel representative of each of the levels are identified to participate. In actual operation, two concurrent workshop convenings are held. The technicians are first oriented to the purpose of the workshop. During the week-long workshop process, the technicians will be providing feedback regarding each of the above cited categories. In the case of the administration/security category, the technicians provide inputs via a Q-SORT process. In this Q-SORT, each technician prioritizes a stack of 3x5 index cards; each card containing one procedure relating to administration or security. After each technician completes the prioritizing of his stack of cards, the group of technicians assembles and rates that one group's set of cards which becomes the final listing of procedures for that section of the design specification. The technicians become familiar with the consensus-seeking process prior to completing this exercise. A short consensus-seeking exercise is provided in order to acquaint them with this process.

The equipment area of the test involves the technicians providing inputs via working in sets of prepared booklets. Each set of booklets represents an equipment subsystem identified as one for which the NEC is responsible. The purpose of this exercise is for the technician to provide insights into the relative importance and difficulty which an entry-level or minimally competent technician should find with that equipment. The group of technicians is presented with an overview of the description of the minimally competent technician, in terms of petrol experience, rating, and submarine/watchstander qualifications.

The first equipment booklet provides for a hierarchialy ranking of the equipment subsystems. It is here that the technician ranks the equipments according to categories of performance of maintenance procedures. Based on a summation of categories of performance of each of these ratings, a hierarchial listing of equipment subsystem is produced. The lowest of equipments might be deleted from the test, depending on the number of subsystems of equipment associated with that NEC. It is important to note that of the approximately 270 items allocated to equipments, a proportional number of items will be assigned hierarchially to each equipment.

Having completed the equipment hierarchial ranking, the technicians next move into the rating of individual equipment packages. This exercise will provide the necessary inputs to determine which phase of equipment

operation should be tested. The technicians will review each operating, standard maintenance, casualty, and corrective maintenance procedure associated with each equipment package. These ratings will be in terms of criticality of performance and difficulty of performance. Immediately after the packages are rated, a computerized scoring algorithm will process the data and produce a hierarchially ranked listing of jobs/tasks associated with each equipment subsystem, and its ranking with reference to the entire subsystem.

This listing will also contain the numbers of items to be assigned to each of the jobs/tasks or procedures. The next step for the group of technicians is to assign suggested cognitive levels for testing purposes to each of the jobs/tasks or procedures. This cognitive level exercise is preceded by a brief workshop in the process of determining appropriate cognitive levels.

Prototype Test Development

The finalized test design specification is produced on-site via a word processing computerized process, and is reviewed by the technician group prior to the end of the workshop. The next step is for the contractor to gain approval of the test design specification by SSPO and the Type Commander. After approval is granted, a prototype test is developed by the contractor. This test, which reflects 360 multiple-choice type questions, follows the design specification. Question items which are specified a cognitive levels above recall are often written in situation sets to elicit responses that would reflect understanding of problem solving associated with the equipments or administrative procedures.

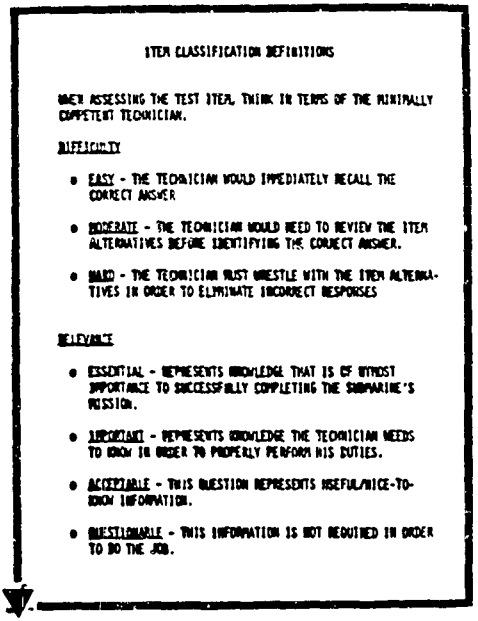
Performance Standard Development

A second week-long workshop is held after the prototype test is developed and after an internal quality control process is accomplished. This Q-C process involves reviewing the items for compliance with a NAVY OD specification developed for the FBM/SWS program. The second workshop involves another two groups of NAVY technicians representative of the NEC; one at the watchstander and one at the supervisor level. This second group of technicians, again working independently but concurrently with each other, will have the task of reviewing their respective test and determining acceptable performance standards. The task for the technician group is to ascertain the minimally acceptable performance level for the watchstander and supervisor respectively. The process adopted and adapted for this ascertaining of minimally acceptable performance standards is a modified Ebel Method. This modified Ebel was developed after consultation with Educational Testing Service consultants and other consultants, all of whom had extensive experiences in the testing and measurement sciences. The Ebel Method involves a jury panel of job incumbents and others with an expertise in the job under consideration, all of whom will make judgments concerning the ability of a minimally competent candidate to answer each question item or series of items correctly.

In practice, an identification of a committee of successful job incumbents and supervisors of incumbents is selected. The Submarine Groups have this as a tasking. Questionnaire instruments are prepared to elicit responses from each of the job incumbents. At the outset of the workshop each committee

member is asked to take the prototype test. The tests are then scored, and the results made known to the committee. This has the effect of providing a benchmark of performance to the committee. After all, they realize that they represent the higher performers in that NEC.

The next phase of the workshop involves each committee-member reviewing each area of the test, section by section, from administration/security through each equipment subsystem, which is prepared in the questionnaire booklets. The technician committee-member is asked to respond to the question item in terms of its difficulty (categories include easy, moderate, hard), and its relevance to the overall mission of the boat (categories include essential, important, acceptable and questionable). Relevance ratings represent the following: (a) essential - represents knowledge that is of the utmost importance to successfully performing a task; (b) important - represents knowledge that a technician needs to know in order to properly perform his duties; (c) acceptable - this question area represents useful/nice-to-know information; and (d) questionable - this information is not required in order to do the job.


ITEM CLASSIFICATION DEFINITIONS

WHEN ASSESSING THE TEST ITEM, THINK IN TERMS OF THE MINIMALLY COMPETENT TECHNICIAN.

DIFFICULTY

- EASY - THE TECHNICIAN WOULD IMMEDIATELY RECALL THE CORRECT ANSWER
- MODERATE - THE TECHNICIAN WOULD NEED TO REVIEW THE ITEM ALTERNATIVES BEFORE IDENTIFYING THE CORRECT ANSWER.
- HARD - THE TECHNICIAN MUST WRESTLE WITH THE ITEM ALTERNATIVES IN ORDER TO ELIMINATE INCORRECT RESPONSES

RELEVANCE

- ESSENTIAL - REPRESENTS KNOWLEDGE THAT IS OF UTMOST IMPORTANCE TO SUCCESSFULLY COMPLETING THE SUBMARINE'S MISSION.
- IMPORTANT - REPRESENTS KNOWLEDGE THE TECHNICIAN NEEDS TO KNOW IN ORDER TO PROPERLY PERFORM HIS DUTIES.
- ACCEPTABLE - THIS QUESTION REPRESENTS USEFUL/NICE-TO-KNOW INFORMATION.
- QUESTIONABLE - THIS INFORMATION IS NOT REQUIRED IN ORDER TO DO THE JOB.

For each item, the committee-member will assign one of the rankings from each of these two dimensions. After the committeemembers complete this exercise for a test section, which is included in one questionnaire booklet, the panel moderator, a contractor representative, will convene the entire group to discuss the findings and reach a consensus decision and overall ranking for the items in that test section. This consensus ranking is

accomplished using the matrix format depicted below.

| IMPORTANCE | DIFFICULTY: EASY | DIFFICULTY: MEDIUM | DIFFICULTY: HARD |
|---|---|---|---|
| ESSENTIAL | # OF ITEMS 10<br>PROPORTION .80 | # OF ITEMS 15<br>PROPORTION .85 | # OF ITEMS 8<br>PROPORTION .70 |
| VERY IMPORTANT | # OF ITEMS 17<br>PROPORTION .80 | # OF ITEMS 6<br>PROPORTION .75 | # OF ITEMS 3<br>PROPORTION .60 |
| NOT VERY IMPORTANT | # OF ITEMS 4<br>PROPORTION .65 | # OF ITEMS 8<br>PROPORTION .50 | # OF ITEMS 9<br>PROPORTION .50 |

CUT-OFF SCORE = (10 x .80) + (15 x .85) + (8 x .70) +
(17 x .80) + (6 x .75) + (3 x .60) +
(4 x .65) + (8 x .50) + (9 x .50) = 59.4

MAXIMUM POSSIBLE SCORE = 80.0

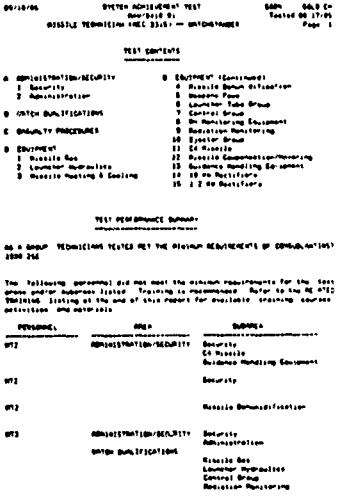
Pre-established percentages are assigned to each cell of the matrix to assist in converting the findings to a percentage "cut-score" for test reporting. The moderator begins a discussion of the test items and displays the findings offered by each committee member on the graphic. Outlyers, that is major discrepancies from any established pattern of response are discussed. A consensus finding is the goal of the exercise. This procedure is then repeated for each area of the test. After the entire test is reviewed in this manner, the committee has an opportunity to review the findings for each area of the test.

Test Scoring Format

After the cut-scores are established and the prototype test is approved, the SAT is released for administration. The SAT program is designed for at-sea administration. Therefore, a package of tests, proctor guides and ancillary testing materials is prepared and delivered to the crew prior to deployment. When the ship returns to home port, the answer sheets are optically scanned for scoring purposes. The individual SAT is scored by area/subarea of the test. The number of items correct is presented as a percentage correct of the total number of items included in that area/subarea of the test. This information is then presented graphically in a Test Report which is distributed to the SSPO, TYCOM and Crew Commander.

## Test Reporting Format

A typical Score Report will present the overall test contents in outline form. It will present, by personnel tested, the recommendations for training as a result of the test. These recommendations will be made by area/subarea of the test. The report will also flag potential problem areas identified as a result of a composite of the test findings, including the crew in question. This is accomplished by presenting both the mean (x) and the average scores, as well as the scores of the crew in question, for each area/subarea. This information is also graphically displayed. A summary sheet will also call out individual scores for each area/subarea for each technician who takes the test on that crew. A score falling below the cut-point will be bracketed for easy recognition.



TEST CONTENTS

A ADMINISTRATION/SECURITY
1 Security
2 Administration
B WATCH QUALIFICATIONS
C ORDNANCE PROCEDURES
D EQUIPMENT
1 Missile Gas
2 Launcher Hydraulics
3 Missile Heating & Cooling
D EQUIPMENT (Continued)
4 Missile Datum Alignment
5 Weapons Panel
6 Launcher Tube Group
7 Control Group
8 Gas Monitoring Equipment
9 Radiation Monitoring
10 Ejector Group
11 C4 Missile
12 Missile Communication/Monitoring
13 Ordnance Handling Equipment
14 10 Pin Rectifiers
15 2 2 Pin Rectifiers

TEST PERFORMANCE SUMMARY

AS A GROUP TECHNICIANS TESTED MET THE MINIMUM REQUIREMENTS OF COMPOSITE(S)?
YES NO

The following personnel did not meet the minimum requirements for the test shown and/or subareas listed. Training is recommended. Refer to the RE ATT/ TRAINING listing at the end of this report for available training courses activities and materials.

| PERSONNEL | AREA | SUBAREA |
|---|---|---|
| WT2 | ADMINISTRATION/SECURITY | Security<br>C4 Missile<br>Ordnance Handling Equipment |
| WT2 | | Security |
| WT2 | | Missile Communication |
| WT3 | ADMINISTRATION/SECURITY | Security<br>Administration |
| | WATCH QUALIFICATIONS | Missile Gas<br>Launcher Hydraulics<br>Control Group<br>Radiation Monitoring |

## Summary

This presentation has described a testing program currently in operation for the FBM/SWS Submarine fleet. It discusses the major iterations conducted to arrive at a performance based System Achievement Test. These steps include a test design workshop which involves incumbent technician personnel reviewing documentation for suitable information for testing; developing and acquiring test items at appropriate cognitive levels for inclusion into the test; developing a prototype test and subjecting this test to review; conducting a cut-score or performance standard determination workshop to ascertain appropriate performance levels; and scoring and reporting the findings in a meaningful manner for those officer personnel who must make personnel judgments about the state of readiness of the FBM program.

# A Study of Suggestopedia at the Defense Language Institute Foreign Language Center (DLIFLC)

Brian J. Bush
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

This report documents the results of a study conducted to evaluate the effectiveness of "Suggestopedia", a method proposed to accelerate language learning, as compared to the standard instructional methodology implemented by DLIFLC. An additional effort was made to evaluate a flexibly-scheduled methodology identified by DLIFLC as a target of opportunity.

## Methodological definitions

The Suggestopedia methodology is characterized by a variety of techniques emphasizing a relaxed and positive learning atmosphere. The instruction is delivered in situational contexts maximizing the use of the oral communicative skills of proficiency. The flexibly-scheduled treatment is similar to the standard DLIFLC methodology, except that the former uses a pacing of the presentation of material based upon group readiness rather than a fixed schedule for the presentation of materials. Both methodologies use the Progressive Skill Integration (PSI) approach which is a functional approach to language teaching that stresses the integration of the various components of language (i.e., pronunciation, grammar, vocabulary, writing systems, etc.) into communication skills. It is a progressive approach in that students progress through a number of stages beginning with the perception of new concepts and culminating with the acquisition of working communication skills.

## Objectives

The objectives of the study were to compare the effectiveness of the Suggestopedia method with the standard DLI method, as well as to provide an evaluation of the flexibly-scheduled pacing procedure.

## Method

### Subjects

The study included forty junior enlisted Army personnel scheduled to begin the Russian Basic Course (RBC) randomly selected and sorted into two sections for each of the Suggestopedia and standard DLIFLC courses.

One section of ten junior enlisted Army and Navy personnel comprised the flexibly-scheduled group. This group was previously identified and in place prior to its incorporation in this study.

Measurement instruments

Effectiveness was measured and evaluated by academic performance and student attitudes toward the methodologies. Three measures of academic performance were used. One was a set of achievement test scores. A second instrument used was a Proficiency Advancement Test (PAT), a combined measure of both achievement and proficiency. The distinction between achievement and proficiency is that achievement measures performance on course materials, while proficiency measures performance with the target language regardless of the course of instruction. The third measure of academic performance used was a face-to-face oral interview which is considered a measure of conversational proficiency only.

Measures of both proficiency and achievement were used in order to provide a better evaluation of the Suggestopedia methodology as compared to the standard and flexibly-scheduled DLIFLC methodologies.

Student attitudes were measured using a pre and posttest instrument and a weekly attitude survey. The pre and posttest addressed student attitudes toward learning Russian and toward learning foreign language in general. The weekly attitude surveys measured student attitudes about themselves while in class, their opinions about the class, and opinions about their instructor(s).

Administration of measurement instruments

The instruments designed to measure academic performance were administered at different intervals in accordance with completion dates projected by the Suggestopedia and standard DLIFLC methodologies. The standard DLI group had the normal fifteen weeks for completion of the same curriculum that the Suggestopedia group had for completion at the end of ten weeks. The flexibly-scheduled group found that they completed the same curriculum in fourteen weeks.

Attitudinal measures followed the same schedule as identified for the performance measures. For example, the posttest was administered at the end of weeks ten, fourteen, and fifteen for the Suggestopedia, flexibly-scheduled, and control groups respectively.

## Results

Demographic variables

Tests of equality between groups were conducted on demographic variables considered to have a possible relation to treatment outcomes. The variables examined were as follows: Military rank, military occupational specialty (MOS), age, years of military service, Defense Language Aptitude Battery (DLAB) scores, educational level, prior language training, General Technical (GT) scores, and gender. A pretest measure of attitudes toward learning Russian and toward learning languages in general was also used to check the equality of the three groups. No significant differences were found among the three groups on any of the demographic variables or on a comparison of the pretest results.

Achievement tests

An analysis of variance was conducted on the two main components of the achievement tests, written and oral. Results from the written component indicated significant differences among the Suggestopedia ($M$=44.63), standard DLIFLC ($M$=81.49), and flexibly-scheduled ($M$=80.87) groups, $F$ (2,46)=48.21, $p$ <.05.

Results from the oral component also indicate significant differences among the Suggestopedia ($M$=62.90), standard DLIFLC ($M$=78.23), and flexibly-scheduled ($M$=79.93) groups, $F$ (2,46)=11.43, $p$<.05.

Results of T-tests indicated that both the standard DLIFLC and flexibly-scheduled groups scored significantly higher than the Suggestopedia group on the oral and written components. There were no significant differences between the standard DLIFLC and flexibly-scheduled groups.

The results indicate that the greatest differences between the Suggestopedia and the two comparable DLIFLC groups are on the written portion of the achievement tests. This finding is expected, in part, because of the emphasis placed by the Suggestopedia methodology on the oral rather than written features of language learning. The comparability of scores for the standard DLIFLC and flexibly-scheduled group reflect both the similarity between their methodologies as well as their balance with written and oral instruction.

Proficiency Advancement Tests (PATS)

An analysis of variance was conducted on each component of the PAT, (i.e., listening, reading, and speaking). No significant difference was found on the listening component among the Suggestopedia ($M$=58.46), standard DLIFLC ($M$=66.23), and flexibly-scheduled ($M$=63.73) groups, $F$ (2,43)=2.83, $p$>.05.

Significant differences were found on the reading scores among the Suggestopedia ($M$=66.71), standard DLIFLC ($M$=73.47), and flexibly-scheduled ($M$=75.10) groups, $F$=(2,43)=6.16, $p$<.05.

The third component of the PAT, speaking, again reflected significant differences between the Suggestopedia ($M$=.64), standard DLIFLC ($M$=.89), and the flexibly-scheduled ($M$=.79) groups, $F$(2,43)=7.632, $p$<.05.

T-tests conducted between the groups on each component indicated that, with one exception, the standard DLIFLC and flexibly-scheduled groups scored significantly higher on each measure than the Suggestopedia group. On the listening component there were no significant differences between the flexibly-scheduled and Suggestopedia groups. In each test there were no significant differences between the standard DLIFLC and flexibly-scheduled groups.

As expected, the PAT results, while still showing significantly lower scores for the Suggestopedia group, indicated somewhat higher overall scores on the PAT then on the achievement tests. The PAT is the first measurement which begins to incorporate elements of proficiency, which are more closely in consonance with Suggestopedia's total emphasis towards language learning.

Face-to-face oral interview scores

The analysis of variance conducted on the face-to-face oral interview scores found no differences between treatments. This is most probably attributable to the small level of proficiency attained after only ten or fifteen weeks of study as well as the early emphasis Suggestopedia places upon proficiency as compared to the more gradual development of proficiency, by DLIFLC, through the integration of the various components of language into communication skills and proficiency.

Attitudinal results

Attitudinal measures failed to discriminate among treatment groups. However, it was interesting to note that the students had positive attitudes toward their particular methodologies. They also felt confident, across groups, in their ability with the target language.

## Discussion

Based upon achievement and proficiency measures, the Suggestopedia methodology did not accelerate learning in the context of this study. In fact, Suggestopedia's gains were significantly less than those found for the standard DLIFLC and flexibly-scheduled groups on both the oral and written measures of achievement, and on the three measures from the Proficiency Achievement Test (PAT), listening, reading, and speaking, with the aforementioned exception on the listening results.

There is some indication that Suggestopedia scores improved as measures of proficiency were increased. Suggestopedia scores were poorest as compared to the two other groups on the achievement only tests. On the PAT, which first began to incorporate some proficiency applications, Suggestopedia scores improved with regard to the other two groups. A further indication of this trend by Suggestopedia to do better as more measures of proficiency are included for evaluation is that there were no significant differences among the three groups on the face-to-face oral interview, a measure of proficiency only. Therefore it seems apparent that Suggestopedia scores would be lower on the achievement-oriented measures which are more heavily stressed during the earlier stages of language learning. A second point should be made regarding proficiency results, especially for the face-to-face oral interviews. Proficiency findings may not be able to provide clear discriminations between methods at the early stages of language training.

The combination of findings from this study, along with the research in related Suggestopedia areas indicate a potential use for at least some of the components found in the Suggestopedia method, for example, the use of incidental learning as a teaching technique; the emphases on relaxation and a positive attitude toward the target language. Martin and Schuster (1974) and Lipsitt and Lolordo (1963) reported a curvilinear relationship between learning and stress. Learning increased as stress increased up to an optimum level after which learning fell off as a function of increases in stress (demonstrating the Yerkes-Dodson Law).

If Suggestopedia were to be used as some form of enrichment adjunct to the established DLIFLC language training program, it should probably be used from one to five weeks. This may be the best time interval since positive student attitudes and the instructor energy required to implement a Suggestopedia program seemed to peak and then dimish after approximately five weeks of intensive application. However, specific components characteristic of the Suggestopedia methodology that are found useful in this study could be incorporated throughout the course of instruction.

## References


1. Martin, D.J., & Schuster, D.H. (1977). The interaction of trait anxiety and muscle tension in learning. Journal of Suggestive-Accelerative Learning and Teaching, 2, 63-67.

2. Lipsitt, L.P., & Lolordo, V.M. (1963 August). Interactive effect of stress and stimulous generalization on children's oddity learning. Journal of experimental Psychology, 66 (2), 210-214.

Effect, of Chemical Protective Clothing and Sustained Cognitive Performance

Bernard J. Rimland, Ph.D. and John L. Goodie
U.S. Army Research Institute of Environmental Medicine
Natick, Massachusetts

Previous research has shown that the relative impermeability of the current Army clothing system for protection against nuclear, biological and chemical (NBC) agents can result in potentially hazardous or incapacitating conditions within the system, particularly during strenuous physical work in the heat (3,4,5,6). There has been very little research on the psychological performance of soldiers in NBC clothing who are working on jobs not requiring strenuous physical activity, and no research on the effects of heat on them.

Heat has been shown to affect cognitive performance markedly in sedentary soldiers, particularly during sustained operations (2,3). The few studies that have investigated the performance of soldiers performing in NBC clothing (e.g.,7,8) are lacking in important scientific and methodological considerations, and have not been concerned with sustained performance.

The purpose of this study was to examine definitively the effects of moderate heat on the sustained cognitive performance of sedentary soldiers wearing NBC protective clothing.

## Method

### Subjects

Twenty-three male soldier volunteers, ages 18-27 (x̄=20.1), were studied. Only persons who could read without glasses were acceptable.

### Tasks

The tasks, which included aspects of those performed by members of artillery fire direction centers (FDC), forward observers and Army communications personnel, were as follows:

(1) Computation of "Site" - "Site" is an aiming adjustment used by FDC's when firing. Tape-recorded data were transmitted to the men over headsets in a format similar to that of artillery fire missions. The men recorded the data on standard form, performed arithmetic calculations, entered data into and read answers from an artillery slide rule and recorded the answers.

(2) Receiving and decoding grid coordinates - Pre-recorded, coded coordinate sets were transmitted as radio messages via headsets. The men recorded the alpha-coded coordinates on a form, chose the correct one of three decoding keys, translated the code into numeric format and recorded the answers.

(3) Receiving and decoding message - Pre-recorded, coded messages of from five to eight words were transmitted as radio messages over the headsets. The men recorded each message on an appropriate form, decoded it by referring to a simulated Army codebook and recorded the transcription on the form.

(4) Plotting targets and determining range and deflections - Each man was given a map with battery positions and deflection reference points. He also had a list of targets (grid coordinates) to plot. He plotted them, found their range and deflections from designated batteries and recorded the answers. He was also informed the time it will be on complete procedure each trial, thus providing assessment of number of pilots per unit of time.



Copy available to DTIC does not permit fully legible reproduction

The first three tasks were paced by the rate/frequency of the messages and were not under the men's control. The map task was at times paced by the situation requirements and at other times was "self-paced." The men did not know which of the three kinds of radio tasks they would be required to perform until a message arrived. Messages were designed to mimic real military radio transmissions, including a variety of voices and transient background noises.

In addition to the above tasks, a visual field surveillance task and an auditory perception task were performed during hours 2, 4 and 6. The results of these tasks will be reported elsewhere. A number of personality measures and vision tests also were administered, but results are not reported here.

Design and Procedure

The men arrived at three-week intervals in six-man groups. Each group completed its assignment before the next arrived. The groups underwent two weeks of intensive training followed by an "experimental" week to evaluate performance in the heat while wearing the MOPP clothing. The clothing system worn was MOPP IV (Mission Oriented Protective Posture; "IV" refers to total encapsulation - suit, worn completely closed, boots, gloves, mask and hood.)

The men trained in a classroom, six to seven hours a day, for two weeks; no weekends. Training in the code-wheel, codebook and site tasks began with simple written forms, and became progressively more complex until the men could handle the rapid, noisy "military" messages. During training, the men practiced several hundred messages with immediate feedback and discussion of errors. Emphasis was placed first on accuracy, then on speed of performance. Map plotting was practiced for hundreds of trials with immediate feedback of errors. Again, emphasis initially was placed on accuracy rather than speed.

During the first week, the men were briefed on proper procedures for donning the MOPP clothing and performed briefly while wearing single components of the system, e.g., gloves only or mask only. During the second week, they performed the tasks daily with and without MOPP IV. By the end of the week, each had worked in MOPP IV for about eight hours, spread over five days.

The "experimental week" proceeded as follows:

Monday:- Two one-hour "refresher" runs to bring the men up to pre-weekend performance levels on the various tasks (21.1 degrees C., 35% rh);

Tuesday:- Control Day, seven hours at 21.1 degrees C., 35% rh, battle dress uniform (BDU), referred to as "BDU-Control-1;"

Wednesday:- Mopp Control Day, seven hours at 32.8 degrees C., 35% rh, MOPP IV worn over BDU, referred to as "MOPP-Control;"

Thursday:- Control Day, same as Tuesday, "BDU-Control-2;"

Friday:- Stress Day, seven hours at 32.8 degrees C., (15% rh, MOPP IV worn over BDU, referred to as "MOPP-Heat-Stress."

The 32.8 degree C. temperature of the MOPP Control Day was calculated by using Equation (1) as the MOPP IV equivalent to the 21.1 degree C. condition used for the BDU for a seven-hour period. These conditions were also used where appropriate during the two-week training period.

The tasks were presented as one-hour blocks of messages, 20 per hour. [illegible] irrelevant (non-stressful to which the men had been trained not to respond), six were codebook, six code-wheel and six site tasks. Intervals between



Copy available to DTIC does not permit fully legible reproduction

m ssages ranged from 30 seconds to over two minutes according to a random o rder. There were no duplicate messages throughout the entire experiment.

Messages were sent in random order. Each of the six men had a different order, which was the same from hour to hour, e.g., if message #5 was codebook for man #1 in the first hour, it was codebook for him in all hours. Since messages and intervening intervals varied in length, work patterns differed; the men were forced to work independently, yet on identical material.

The radio messages were presented to the men four times on each of the four experimental days, as hours 1,3,5 and 7.

Concurrent with working the radio messages, the men did the map task. The radio messages had priority; map work was interrupted upon hearing a message, to be resumed only when what was required by the message was completed. Thus, for each of hours 1, 3, 5 and 7, the men were always engaged in cognitive work.

During each of hours 2,4, and 6, the men worked only on the map task, at their own pace and without interruption, for about 50 minutes. They also were tested for auditory perception, in groups of three, and, individually, for peripheral vision. A ten-minute break was given each hour.

On the MOPP-Heat-Stress Day, rectal temperatures were taken every five minutes or less. Regulations require removal from heat if temperatures exceed 39.2 degrees C. The men were encouraged to drink water ad lib; masks and canteens were equipped for drinking in MOPP IV. No lunch was eaten during the experiment week. Access to a portable toilet was permitted only on an emergency basis. Only three men availed themselves of this opportunity during the entire study.

## Results and Discussion

Twenty-three men actually arrived to participate. Of those, three were disqualified; one was allergic to the MOPP clothing, one had insufficient training due to illness and one went on sick-call on the final (stress) day. Therefore, all data analyses are based on n=20.

Responses to the messages were scored/verified by criteria established beforehand. Errors were classified into "omission" and "commission." Omission consisted of missing part of a message or translation. Commission included incorrect reception of messages and incorrect translations and computations.

On the MOPP-Heat-Stress Day, two men had to be evacuated for medical reasons. They were given the maximum number of omission errors possible for the missed hours and scored as having plotted no targets for the time they missed. The group averages reflect those decisions [see Fine & Kobrick (? page 121)].

Individual data for the codebook task are shown in Figure 1 for errors of omission and commission combined. An ANOVA yielded a significant F in effect for Conditions ($F=11.7$; $df=3,57$; $p<.001$). No significant differences between the two MOPP-control conditions or between hours of testing within either of the control conditions were evident at any time.

The MOPP suit by itself (MOPP-Control) seems to produce an increment in performance. Not significant statistically for the first three hours, the effect

Copyright ... permit fully ... reproduction

s ... stronger after five hours. However, after seven hours, two distinct patterns emerged: performance in the MOPP-Control condition improved to the level of the PLU-Control conditions, whereas the group showed a statistically significant decrement in excess of 20 percent when in the MOPP-Heat-Stress condition. The decrements in performance in both MOPP conditions were found to be due almost entirely to increases in errors of omission.

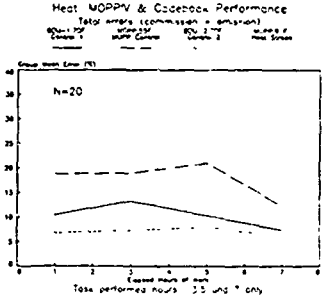


Figure 1

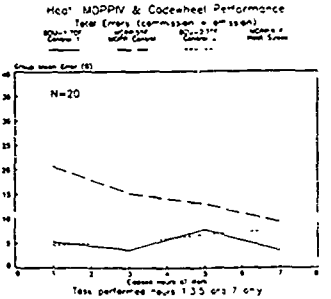


Figure 2

CODE BLOCK: Results for the codewheel task are shown in Figure 2. ANOVA yielded a significant main effect for Conditions ($F=18.2$;d.f.2,234;$p<.00001$), and a significant Conditions by Hours of Work interaction ($F=4.11$;d.f.8,234;$p<.0001$). Performance in the PLU-Control conditions showed remarkable consistency and stability. The opposite trends of the two MOPP conditions clearly were evident again. In the MOPP- Control condition, performance initially was adversely affected (significantly only after the first and third hours), only to gradually improve until, after seven hours, it approached the levels of the two PLU conditions. The MOPP-Heat-Stress condition showed no initial adverse effects, but dramatic and significant increases in error occurred from the third to the fifth and the fifth to the seventh hours. Virtually all of the increase occurred as errors of omission.

CODEBLOCK OF LETTERS: This was the only radio task in which the message was repeated; in both other tasks, the message was heard only once, hence the lower error rates for this task, as shown in Figure 3. ANOVA showed a significant main effect for Conditions ($F=.6$;d.f.2,234;$p<.01$) and for Elapsed Hours of Work ($F=2.2$;d.f.2,234;$p<.05$) and a significant Conditions by Hours interaction ($F=.06$;d.f.8,234;$p<.00$).

Performance in the two PLU-Control conditions again was highly stable, remarkably similar and quite free of errors. The group in these conditions averaged only 2-3 errors consistently for each of the four hours in which the task was performed. There was no significant decrement in performance in the MOPP-Control condition hour by hour. The MOPP-Heat-Stress condition showed significant increases in percent group error from the third to the fifth and the fifth to the seventh hours, culminating in about a 20% decrement, due entirely to an increase in errors of omission.

MEMORY OF WORDS PLOTTED, HOURS 1,3,5 AND 7: ANOVA yielded a significant main effect for Conditions ($F=11.6$;d.f.2,234;$p<.00001$), and a significant effect for Elapsed Hours of Work ($F=3.2$;d.f.3,234;$p<.10$). More targets tended

to b plotted in the PDU-Control-2 condition than in PDU-Control-1, but this was significant only for the third hour. More important was the tendency for target plotting productivity to be lower in the MOPP conditions, but this was not consistently statistically significant; it did not hold when both PDU control groups were compared with both MOPP groups. The effect was much more apparent when the data for self-paced hours 2,4, and 6 were considered.

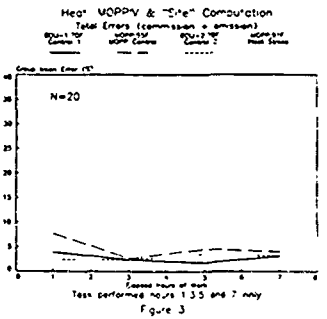


Figure 3

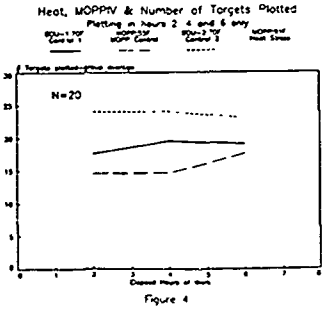


Figure 4

[illegible] PLOTTED, [illegible] 2,4, [illegible]. These data are shown graphically in Figure 4. The ANOVA resulted in a very significant Conditions effect ($F=11.4$; $df=3,23$; $p<.00001$). As in hour 2, the men were most productive in the PDU-Control-2 condition than in the other conditions. Here, however, the effect was significant for each of hours 2,4, and 6.

Concerning the effects of the NBC suit, by the fourth hour, the group, when in either MOPP condition, had significantly lower productivity than when in either PDU-Control condition. By the end of the sixth hour, however, productivity in the MOPP-Control condition had returned to the same level as the PDU-Control-1 condition, whereas productivity in the MOPP-Heat-Stress condition showed a significant deterioration (17.7 targets per hour, compared with 17.6 per hour for MOPP-Control, 19 per hour for PDU-Control-1 and 23 per hour for PDU-Control-2). The stability and magnitude of performance in the PDU-Control conditions should be noted, attesting both to the effective training procedures and the high level of motivation of the men.

[illegible]: Several other measures were obtained but were not used: (a) Number of plotting errors- After training, all men could plot with virtually no errors. This superior performance showed in the experimental data. During hours 1,3,5 and 7, the group averaged .3% errors per hour. During hours 2,4 and 6, it averaged .61 errors per hour. Because of the low incidence of errors, statistical analyses were not performed; (b) Number of ranges calculated- The number of ranges was the same as the number of targets since a range was required for each target. Thus, results for ranges and for targets are the same; (c) Range calculation errors- These were so infrequent that statistical analysis were meaningless; (d) deflection calculation errors- Due to an error in placement of deflection reference points, deflections for some targets could not be determined since those targets did not occur in systematic rows, and since each man worked at his own pace, meaningful [illegible] activity or accuracy information could not be determined.

3 (CLOTHE): After four to five hours of exposure to a moderately hot environment, the cognitive performance of a group of highly trained soldiers clad in the MOPP IV configuration of NBC protective clothing began to deteriorate markedly. By the end of seven hours of exposure, increases in percent group error ranged from 17% to 29% over control conditions on investigator-paced tasks. Virtually all of this decrement was due to increases in errors of omission. The productivity of the group on a self-paced task (map plotting) diminished by about 40% from control conditions after six hours in the heat, but accuracy of plotting did not appear to be markedly affected.

While generalization from a small sample should be made with caution, it should be noted that the control data in this study are very similar to control data from three previous studies using the same tasks (2,3,4). This consistency of baseline performance with near lends credence to the validity of the results presented here.

Finally, this study was done in a climatic chamber under conditions in which the men knew that they were being monitored for health and safety. The additional stress of being exposed to actual toxic agent or even the threat of it and/or other hostile situations in all likelihood will have further negative impact on group performance.

## References


1. Breckenridge, J.R. U.S. Army Research Institute of Environmental Medicine, Natick, MA. Personal communication, 4 Nov 81.
2. Fine, B.J. & Kobrick, J.L. Effects of altitude and heat on complex cognitive tasks. Human Factors, 1978, 20, 115-122.
3. Fine, B.J. & Kobrick, J.L. Effects of altitude and heat on complex cognitive tasks,II. Unpublished research.
4. Fine, B.J. & Kobrick, J.L. Effects of simulated rapid translocation and heat on sustained performance of cognitive tasks. Unpublished research.
5. Garrett, Jr., P. & Phillis, J. Human Engineering Laboratory Aviator Supply III/V (HEVAS III/V). Presented at Tri-Service Aeromedical Rsch. Panel Chemical Defense Meeting, San Antonio, TX, 1981.
6. Goldman, R.F. & Breckenridge, J.R. Current approaches to resolving the physiological heat stress problems imposed by chemical protective clothing systems. In Proc. 1976 Army Sci. Conf., West Point, N.Y. Vol. IV,447-465.
7. Hamilton, P.E., Simmons, R.F. & Kimball, K.A. Psychological effects of chemical defense ensemble imposed heat stress on Army aviators. U.S. Army Aeromedical Rsch. Lab., Ft. Rucker, AL, USAARL Rept. 82-7, 1982.
8. Hamilton, P.E. & Zapata, L. Psychological measurements during the wear of [illegible] Chemical Defense Ensemble. U.S. Army Aeromedical Research Lab., Ft. Rucker, AL, USAARL Rept. 82-7, 1982.
9. Shvartz, E., Goldman, R.F. & Goldman, R.F. Comparison of physical, biological and physiological methods of evaluating the thermal stress associated with [illegible] protective clothing. Ergonomics, 1976, 19, 47-342.
10. Yorke, E.D. Selection and physical conditioning requirement of rapid runway [illegible] personnel with the CB ensemble. Human Engr. Lab. Aviator Supply III/V (HEVAS III/V). Presented at Tri-Service Aeromed. Rsch. Panel Chemical Defense Technical Meeting, San Antonio, TX, 1981.


(Opinions presented in this paper are those of the author(s), and do not necessarily reflect the views or opinions of the Department of the Army [illegible] position, unless so designated by other official documentation.)

# Exceptional Recruits: A Look at High- and Low-Aptitude Personnel

**Anita R. Lancaster**
**Office of the Assistant Secretary of Defense**
**(Force Management and Personnel)**

The Department of Defense, on a quarterly and annual basis, reports on the "health of recruiting" by describing the quality and quantity of youth accessed. Recruit quality is defined in terms of aptitude test scores and educational level (high school graduation status). During periods of favorable recruiting, such as that experienced since 1981, DoD typically reports that accessions have above average aptitude and that their high school graduation status is as good or better than that of the American youth population.

Recently, however, DoD has reported recruiting men and women of unprecedented quality. For example, in Fiscal Year (FY) 1984, all four Services met or exceeded their accession objectives. Of the 333,400 young men and women entering the Services last year, 93 percent possessed high school diplomas and 93 percent scored average or above average on the enlistment aptitude test. This contrasts with previous high school graduate recruit levels of 91 percent in FY 1983, 86 percent in FY 1982, 81 percent in FY 1981, and 68 percent in FY 1980. Similarly, the FY 1984 aptitude level of 93 percent in the average or above category compares with 92 percent for FY 1983, 87 percent for FY 1982, 82 percent for FY 1981, and 69 percent for FY 1980. Although the trend has been upward, it appears that recruiting results for FY 1985 will be similar to those of FY 1984.

Seventy-five percent of American youth have high school diplomas and 69 percent of a nationally representative sample of American youth ages 18 to 23 scored average or above on the enlistment test. Given that 93 percent of FY 1984 recruits both have high school diplomas and scored average or above on the enlistment test, it would seem that the Services have taken in a disproportionate share of high quality American youth into their ranks over the past five years.

Because of the exceedingly high quality levels achieved in recent years, there has been considerable speculation that the Services' abilities to successfully enlist such youth, over time, simply cannot continue. In fact, some contend that an improving economy, a decreasing manpower pool, and keen competition with higher education and employers for bright young people may result in a significant decline in the quality of recruits in the future.

There is no doubt that if recruiting becomes more difficult, the costs of competing with the private sector for exceptionally high quality youth must be weighed against the consequences of incrementally lowering recruit quality levels. Someday, for example, we may have to consider the option

of lower aptitude standards against increased military pay or reenlistment bonuses. There also, undoubtedly, is some point of diminishing returns - where a decline in recruit quality would adversely affect force readiness. Current state-of-the-art, however, does not provide us with any precise way to calculate the effects of varying such options.

It is incumbent upon the Department of Defense to begin to plan for potential recruiting difficulties now. It would be unconscionable for us to believe that current recruit quality will continue in an unmitigated flow. Similarly, we would be remiss in not exploring the development of methodologies for determining appropriate quality levels so that we do not increase recruiting resources to maintain inflated levels. We would not want to make these decisions capriciously. We need to know how much it costs to recruit, train, supervise and retain personnel at all aptitude levels in order to make informed decisions.

Fortuitously, we are in a position to study the effects of having accessed both high- and low-aptitude recruits. At last year's Military Testing Association conference, Lieutenant General E. A. Chavarrie, Deputy Assistant Secretary of Defense (Military Manpower and Personnel Policy) outlined the research efforts we then were initiating in this area, and I am pleased to report that we have made considerable progress since then.

With regard to researching the effects of enlisting high-aptitude personnel, Ms. Janice H. Laurence, from the Human Resources Research Organization (HumRRO), will present a paper she co-authored with Ms. Elizabeth Schneider, also from HumRRO, on the "Enlistment and Utilization of High-Aptitude Recruits." This competitively bid project recently was awarded to HumRRO and, through this research, we will be learning about the demographic characteristics and pre-service experiences of high-aptitude recruits. In addition, HumRRO researchers will be examining the occupational assignment of high-aptitude personnel and their performance, relative to lower aptitude personnel; thus, we hope to be able to develop a model for estimating the incremental benefits of enlisting higher aptitude youth versus those of somewhat lower aptitude. If we able to do so effectively, we will be in a position in the future to answer "What...if" types of questions with regard to the effects of lowering or raising recruit standards.

Another HumRRO researcher, Dr. Barbara M. Means, will be presenting a paper she co-authored with Ms. Arti Nigam and Ms. Jane G. Heisey, also from HumRRO, entitled, "When Low-Aptitude Recruits Succeed." These researchers have been studying the characteristics and performance of two cohort groups - those recruits accessed during Project 100,000, when low-aptitude personnel deliberately were enlisted, and the recruits who were unintentionally accessed during the misnorming of the enlistment test in the late 1970s. These analyses involve not only an examination of the military performance of low-aptitude recruits, but an assessment of the impact of military service on the post-military lives of these recruits. The data that Dr. Means will share on the military performance of low-ability personnel is illuminating.

Finally, Dr. W. S. Sellman, Director for Accession Policy, Office of the Assistant Secretary of Defense (Force Management and Personnel), will provide us with discussant remarks. Dr. Sellman was one of the original Project 100,000 researchers under Mr. I. M. Greenberg, then Director of Project 100,000 for Defense Secretary Robert S. McNamara and later a Deputy Assistant Secretary of Defense. Today, Dr. Sellman has responsibility for setting policy for recruitment and utilization of high- and low-aptitude personnel.

# Using Military High- and Low-Aptitude Recruits Wisely

Wayne S. Sellman
Office of the Assistant Secretary of Defense
(Force Management and Personnel)

Time and time again the Department of Defense has expressed its interest in and need for quality personnel. We try to maximize the enlistment of youth whose aptitude and education level ensure the highest probability of successful military performance. Nongraduates and lower-aptitude personnel are less adaptable to military life and are harder to train than their high school graduate, higher-aptitude counterparts.

Despite DoD's preference to enlist the bright and well educated, recruiting market trends, conditions, and constraints in the past have led to enlistment of substantial numbers of low-aptitude personnel and less than optimal use of high-aptitude personnel. Though currently there is no recruiting "red alert" in terms of attracting adequate numbers of bright young people into the military, prudent management calls for planning against a time when this might not be true.

With a declining manpower pool, the Services will be in heavier competition with the private sector for young "employees," particularly those of high-aptitude. Because high-quality personnel may become more difficult to recruit and retain, it is a propitious time to begin studying in more depth both the performance of low- and high-aptitude personnel so that we may use our resources wisely.

## Lessons from the Past

As mentioned previously, the Defense Department has experience with accessing large numbers of relatively low-aptitude personnel. From 1966 through 1971, for instance, 320,000 low-aptitude men entered service under "Project 100,000." This program, as part of President Lyndon Johnson's War on Poverty, admitted low-aptitude men, including many who would have been disqualified under previous aptitude (test score) standards, into the military so that they might learn useful skills. Though Project 100,000 was initiated to provide the military's educational and training opportunities to our nation's culturally disadvantaged youth, it also served as a partial solution to the quantitative manpower demands imposed by increasing American involvement in Vietnam.

Between 1976 and 1980, a second large-scale accession of low-aptitude individuals occurred. This occasion was not in response to a specific, planned program; rather these large numbers of low-aptitude recruits resulted from the inadvertent misnorming of the Armed Services Vocational Aptitude Battery (ASVAB), the test given to select and classify recruits (Sellman & Valentine, 1981). Had the test been properly calibrated, many of these "Potentially Ineligibles," as Greenberg (1980) named them, would not have qualified for enlistment or would have been assigned to less demanding

occupational specialties. This "unfortunate" occurrence may have also been fortuitous in forestalling major manpower shortages during this difficult recruiting period.

## Looking Toward the Future

Many people believe that the Defense Department's role as an institution in American society goes beyond that of "providing for the common defense." Because the military is this country's largest employer of youth, many social engineers believe that DoD should have responsibilities for training and imparting lasting skills and values for society's disadvantaged youth. Military Service, they believe, should provide job-related training opportunities for valuable social and growth experiences, and increased interaction and cooperation with a wide variety of different kinds of people. Disadvantaged youth should be afforded the opportunity to "grow-up" in a relatively structured environment--learning to be responsible productive citizens.

One does not have to agree with this role for DoD to understand that the Services have accessed low-aptitude youth many times to come to the nation's defense and to provide needed manpower. In the event that the Defense Department is called upon in the future to admit large numbers of low-aptitude recruits, the lessons learned from the past may prove invaluable. DoD should understand more fully the effects of low-aptitude personnel on the military and the military's effect on such individuals.

## From Low to High: Developing Human Resources

The Department of Defense's role as a developer of human capital has been directed primarily at those recruits who are of average or above average ability. Less is known about developing the skills and abilities of low and very high-aptitude recruits so that they might reach their respective full potential in the military and carry with them the benefits of their military experience into civilian life.

Three research studies are currently being conducted which address these questions. Our project dealing with high-aptitude enlistees will describe their enlistment motives, performance across jobs of varying complexity, and reasons for leaving military service. This study will also compare the costs and benefits associated with accessing high- versus average-aptitude enlistees. This is the most comprehensive attempt to investigate performance of high-aptitude personnel.

The second study aims at investigating the ability of low-aptitude individuals to function effectively in the military--an issue of concern for years. Though research indicates that low-aptitude enlistees do less well in military training, there is a scarcity of empirical data concerning their actual job performance. Regardless of DoD's role as a social institution, information is needed concerning the costs of training and supervising these individuals when placed in various military occupational categories.

The third study examines the effect of military service on the post-military lives of low-aptitude veterans. Advantages provided to low-aptitude enlistees (i.e., maturation, identification with an established organization, enhanced self worth, responsibility, training opportunities) are obvious, according to social observers. However, actual data documenting employment and social experiences of low-aptitude veterans has not been available generally. This project allows for detailed comparisons of low-aptitude veterans accessed during Project 100,000 (1966-1971) and the ASVAB misnorming (1976-1980) with their low aptitude nonveteran counterparts.

DoD's current pioneering efforts, aimed at studying the utilization and performance of youth at both ends of the ability distribution, will enable us to use our precious resources wisely. Knowing the strengths and weaknesses of both high- and low-aptitude personnel will enable DoD to assign them more efficiently and better meet the challenges of the changing demographics of the next decade.

## References


Sellman, W.S., & Valentine, L.D. (1981, August). Aptitude testing, enlistment standards, and recruit quality. Paper presented at the 89th Annual Convention of the American Psychological Association, Los Angeles, CA.

Greenberg, I.M. (1980). Report on mental standards for enlistment: Performance of Army personnel related to AFQT/ASVAB scores. In Department of Defense, Implementation of new Armed Services Vocational Aptitude Battery and actions to improve the enlistment standards process. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, & Logistics).

# When Low-Aptitude Recruits Succeed

Barbara M. Means, Arti Nigam, and Jane G. Heisey

Human Resources Research Organization

> The CNO.. does not want enlistment standards lowered and he certainly does not want another "Project 100,000," the infamous effort of a generation ago to bring into the armed services individuals below minimum enlistment standards and "train them up" so they could perform adequately. "Project 100,000" is almost universally remembered by military people as a disaster that caused the military services tremendous grief.
>
> John Burlage, The Navy Times, February 25, 1985

> ...there was little, if any, measurable adverse effect on the services and very substantial gains, in my opinion, to the individuals and to society.
>
> Robert S. McNamara, in interview with T. Sticht, April 26, 1985.

The extent to which the military services should accept individuals with low aptitude test scores for military enlistment has been, and remains, a controversial topic. Both during Project 100,000, when the acceptance of large numbers of young men previously ineligible because of their aptitude scores was instituted as a social welfare program, and during the misnorming of the enlistment examination when thousands of low-aptitude individuals were accidentally accepted for military service, emotions have run high. Military commanders argue that increased training costs, poorer military performance,

Paper presented at the Military Testing Association, San Diego, October 1985.

This research was performed under contract No. N66001 with the Navy Personnel Research and Development Center and was sponsored by the Directorate for Accession Policy, Office of the Assistant Secretary of Defense (Force Management and Personnel). The views, opinions, and findings expressed in this paper are those of the authors and should not be construed as Department of the Navy or Department of Defense position or policy unless so designated by other documentation.

The authors wish to extend their gratitude to the Defense Manpower Data Center (DMDC) and particularly to Les Willis, who performed the statistical tabulations of DMDC Cohort File data.

and a tarnishing of the military's image result from accepting large numbers of such individuals. Social critics point to the military's capability to provide a positive force for improvement in the lives of the economically and educationally disadvantaged, and note that when operational aptitude standards are raised in good recruiting years, smaller proportions of women and minorities (who score lower on the enlistment examination) are accepted for military service.

For enlistment purposes, intellectual aptitude is measured by Armed Forces Qualification Test (AFQT) scores, which are reported in terms of percentiles and are grouped into five broad categories--Categories I through V--with Categories III and IV further divided into subgroups denoted by letters. Numerically higher groups signify lower aptitude. Category IVA, for example, includes percentiles 21 through 30, and was the lowest eligible aptitude category for high school graduates in the Navy, Air Force, and Marine Corps; the Army accepted high school graduates in Category IVB, which includes percentiles 16 through 20. Higher aptitude standards were set for non-high school graduates (except in the Marine Corps).

Rather than join either side of the "I loved it" - "I hated it" debate, we have chosen to reframe the question. When low-aptitude individuals are brought into military service, where do they perform best? The services, of course, have their own ideas about this question, and these are reflected in the minimum aptitude scores set for entry into various jobs and in quotas set for the maximum number of Category IV personnel that can be recruited for a given job. A July 1984 memo from the Army Deputy Chief of Staff for Personnel (DCSPER) to the Army Recruiting Command, for example, stipulated that the specialties for which the most Category IV personnel could be recruited that year were infantry (14.4% of Category IV recruits), motor transport (8.4%), cannon crew (6.1%), and food service (5.1%). Many jobs, such as intelligence analyst, radar controller, and computer repairer were not to get any Category IV personnel.

The largest, recent data set concerning the performance of low-aptitude military personnel is found in the Defense Manpower Data Center (DMDC) records for those individuals who entered service between January 1976 and October 1980 as a result of an enlistment test miscalibration. The particular aptitude score range of those individuals let in by mistake varies depending on the service entered and the individual's education level, since required scores vary for these groups. The analyses to be reported in this paper contrast the highest of the ineligible aptitude categories (which contained most of the inadvertent acceptees) for each service/education group with the lowest of the aptitude-eligible groups. Because military jobs differ considerably in terms of their duties and required intellectual aptitude, all performance measures were examined separately for the nine DoD occupational codes.

Although there is considerable variation within each occupational code in terms of the aptitude composite scores stipulated for entry into various jobs within that code, the occupational codes do vary considerably in terms of average score required and thus can be classified according to aptitude demands as perceived by military managers. The codes with high requirements are electronic equipment repair (1), communications and intelligence (2),

medical and dental technicians (3), and other technical and allied (4). Medium requirements are set for functional support and administrative (5) electrical/mechanical equipment (6), and craftsmen (7) specialties. Low requirements are set for infantry, gun crew, and seamen (0) and for service and supply handlers (8). These job demand categories were used in our analyses of military performance.

Probably the most basic--and the most often quoted--military performance index is attrition. To look at our comparison groups, which contain men who entered in different years and enlisted for different terms of service, we counted as attrition cases all those who had separated from enlisted service for reasons other than entry into an officer program or completion of their term of service. These data are shown for the aptitude comparison groups, by education and occupation type, in Table 1. The data clearly show that if completion of one's term of enlistment is the criterion for success, the inadvertent enlistees did as well as those in the higher-aptitude comparison groups. In fact, the lower-aptitude groups generally had somewhat lower attrition rates, at least in the Army and Navy. Their rates were significantly lower ($p<.01$) in the Army among nongraduates in low- or medium-requirement jobs. Similarly, in the Navy, the aptitude-ineligible group had less attrition in both low- and medium-requirement ratings, regardless of education level. (It should be noted, however, that in the Navy many individuals do not have an assigned occupation, and lower-aptitude individuals were more likely to be in this no-occupation group.) The usual 2:1 difference between nongraduates and graduates in terms of attrition rates is apparent in this table, and the education differences completely overwhelm those attributable to aptitude category. These findings are compatible with those of Shields and Grafton (1983), who found no difference between these aptitude groups in terms of first-term completion for Army enlistees in 18 specialties.

Shields and Grafton examined other first-term performance measures as well. The only measure to show any differences between these aptitude groups--and the differences were modest--was Skill Qualification Test (SQT) score. The Army's SQTs are job-specific tests containing hands-on, ratings, and written test measures of task performance. Most of the tests' variance derives from the written portion of the exam, and hence a correlation with AFQT score is to be expected. Nevertheless, differences were not impressive: averaged over 17 jobs, the mean standardized SQT score was 96 for Category IVA-B (eligible) and 91 for Category IVC (ineligible) accessions among high school graduates; among nongraduates, it was 99 for Category IIIB (eligible) and 95 for Category IVC (ineligible). (The test is standardized with $\bar{X}=100$ and SD=20.) In an earlier report, Greenberg (1980) reported a similar pattern of SQT scores among FY 1977 Army accessions across nine jobs.

With the passage of time, we can now look at indications of career performance for many of the individuals who entered during the ASVAB misnorming. Since most of these measures, such as retention and promotion, are dependent upon date of entry, we decided to look at a single cohort--FY 1977 entrants. (This cohort includes an extra quarter because of the change in government fiscal year, and hence is the largest single-year cohort from that period.) Using records that were updated as of December 1984, we could examine military records for over seven years of military service. To permit an unbiased examination of promotion rates, those who entered with an advanced pay grade (anything other than E-1) were deleted from the data set.

Table 1

Percent Attrition for FY 1976 - FY 1980 Male Accessions in Aptitude Comparison Groups, by Occupational Requirement Level, Service, and Education
(N appears in parentheses)

| Occupational Requirement | ARMY | | | | NAVY | | | | AIR FORCE | | | | MARINE CORPS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | |
| | IVC | IVB | IVA | IIIB | IVB | IVA | IVA | IIIB | IVB | IVA | IIIA | II | IVB | IVA | IVB | IVA |
| Low | 19* (18,172) | 20 (20,444) | 41** (27,595) | 43 (25,526) | 5** (2,235) | 8 (3,357) | 13** (1,001) | 20 (2,865) | 24 (408) | 22 (3,015) | 41 (1,304) | 41 (1,896) | 19 (6,819) | 19 (10,009) | 30 (944) | 31 (3,649) |
| Medium | 14 (12,581) | 14 (14,703) | 30** (15,275) | 33 (16,901) | 4** (7,287) | 6 (13,755) | 12** (3,019) | 15 (8,562) | 20 (1,385) | 18 (8,794) | 40 (3,678) | 38 (5,513) | 13 (2,217) | 13 (4,479) | 18 (256) | 21 (1,065) |
| High | 11 (4,312) | 12 (7,287) | 30* (7,870) | 32 (10,470) | 7 (1,097) | 8 (2,990) | 15 (427) | 17 (2,102) | 16 (141) | 15 (1,003) | 31 (1,006) | 28 (2,941) | 14 (656) | 13 (1,269) | 22 (83) | 26 (427) |

*Significantly different from higher-aptitude comparison group at $p < .05$.
**Significantly different from higher-aptitude comparison group at $p < .01$.



Table 2

Percent Retention as of December 1984 for FY 1977 Male Accessions in Aptitude Comparison Groups, by Occupational Requirement Level, Service, and Education
(N appears in parentheses)

| Occupational Requirement | ARMY | | | | NAVY | | | | AIR FORCE | | | | MARINE CORPS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | |
| | IVC | IVB | IVA | IIIB | IVB | IVA | IVA | IIIB | IVB | IVA | IIIA | II | IVB | IVA | IVB | IVA |
| Low | 18** (4,236) | 21 (4,554) | 9 (7,030) | 9 (5,534) | 36* (658) | 30 (1,013) | 25 (483) | 24 (868) | 20 (64) | 25 (512) | 15 (157) | 20 (310) | 11 (1,612) | 11 (2,304) | 3 (319) | 5 (1,164) |
| Medium | 26 (3,403) | 26 (3,867) | 15 (5,191) | 14 (4,817) | 34** (1,737) | 30 (3,634) | 26 (1,110) | 23 (2,181) | 35 (170) | 35 (1,668) | 24 (475) | 24 (971) | 30 (405) | 29 (754) | 16 (86) | 17 (327) |
| High | 34 (1,028) | 33 (1,784) | 19 (2,157) | 18 (2,474) | 42 (243) | 37 (673) | 37* (145) | 28 (497) | --[a] (29) | 38 (209) | 30 (141) | 30 (456) | 33 (114) | 27 (214) | -- (23) | 10 (133) |

*Significantly different from higher-aptitude comparison group at $p < .05$.
**Significantly different from higher-aptitude comparison group at $p < .01$.
[a]Fewer than 50 cases per occupation category.

Table 2 shows the percentage of men from each aptitude comparison group who were still in service as of December 1984. From that table it's apparent that the inadvertently enlisted men are just as likely to now be in the career force as are those in the appropriate higher-aptitude comparison group. Although retention rates are generally higher in the more demanding occupations (which are regarded as more desirable), no consistent aptitude-by-requirements interaction emerges.

A more sensitive index of career performance is paygrade achieved. Early paygrades are awarded pretty much in lock-step fashion (although there are considerable differences across services as to when the steps are taken), but promotions above E-4 are increasingly competitive. For those FY 1977 entrants in our aptitude comparison groups who were still in service in December 1984, Table 3 shows the proportion that had attained E-5 or better, and Table 4 shows the proportion that had attained E-6 or better. (Because of service differences in promotion policies, a floor effect in the Air Force E-6 data makes E-5 attainment more useful for examining performance within that service.)

These paygrade data are the first we have examined that show consistent mean differences favoring the higher-aptitude groups. Even here, the differences are not large. Analyses of the proportion of accessions still in service that attained a paygrade of E-5 or higher found no significant difference at $p<.01$ and only a single difference (for Marine Corps high school graduates in low requirement occupations) significant at $p<.05$. E-6 differences attain significance only (1) in the Army among high school graduates in occupations with low demands ($p<.01$) and (2) in the Navy for graduates in medium-demand occupations ($p<.05$) and nongraduates in high-demand occupations ($p<.01$). Moreover, it should be recognized that advancement is not based on on-the-job performance alone; individual scores on written tests for advancement are weighed and, in some cases, minimum scores on aptitude tests such as the AFQT are required as well. In this light, observed differences in advancement appear real, but surprisingly small.

Overall, we can make only a few generalizations at this point about the type of jobs in which low-aptitude individuals perform best. They are less likely than higher-aptitude men to become attrition cases within low- and medium-demand occupations, but there is no consistent pattern with regard to occupation type for the retention and promotion measures. What we can say about "when" low-aptitude recruits perform well is that they look quite acceptable when one looks at attrition or retention figures within an education/occupation group. They look less impressive when the analyst examines attainment of higher NCO ranks, but even here advancement appears to depend more on occupation type within a service than on aptitude category.

## References


Greenberg, I. M. (1980). Report on mental standards for enlistment. In Department of Defense, Implementation of new Armed Services Vocational Aptitude Battery and actions to improve the enlistment standards process. Washington, DC: OASD (Manpower, Reserve Affairs, & Logistics).

Shields, J. L., & Grafton, F. C. (1983). A natural experiment: Analysis of an almost unselected Army population. Alexandria, VA: Army Research Institute.

Table 3

Percent of FY 1977 Male Accessions Still in Service as of December 1984 Who Had Achieved Paygrade E-5 or Higher, by Occupational Requirement Level, Service and Education, for Selected Aptitude Groups (N appears in parentheses)

| Occupational Requirement | ARMY | | | | NAVY | | | | AIR FORCE | | | | MARINE CORPS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | |
| | IVC | IVB | IVA | IIIB | IVB | IVA | IVA | IIIB | IVB | IVA | IIIA | II | IVB | IVA | IVB | IVA |
| Low | 85 (774) | 83 (959) | 79 (660) | 79 (494) | 70 (235) | 68 (308) | 63 (120) | 67 (210) | --[a] (13) | 74 (127) | -- (23) | 83 (63) | 81* (185) | 89 (263) | -- (8) | 79 (58) |
| Medium | 80 (891) | 83 (1,021) | 74 (760) | 77 (697) | 73 (590) | 76 (1,092) | 66 (288) | 72 (511) | 73 (60) | 82 (587) | 85 (115) | 87 (232) | 81 (122) | 84 (216) | -- (14) | 82 (54) |
| High | 82 (354) | 83 (590) | 82 (409) | 84 (439) | 76 (103) | 79 (250) | 76 (53) | 81 (139) | -- (9) | 75 (79) | -- (42) | 86 (135) | -- (38) | -- (58) | -- (5) | -- (13) |

*Significantly different from higher-aptitude comparison group at $p<.05$.
[a]Fewer than 50 cases per occupation category.



Table 4

Percent of FY 1977 Male Accessions Still in Service as of December 1984 Who Had Achieved Paygrade E-6 or Higher, by Occupational Requirement Level, Service and Education, for Selected Aptitude Groups (N appears in parentheses)

| Occupational Requirement | ARMY | | | | NAVY | | | | AIR FORCE | | | | MARINE CORPS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | | HSG | | NHS | |
| | IVC | IVB | IVA | IIIB | IVB | IVA | IVA | IIIB | IVB | IVA | IIIA | II | IVB | IVA | IVB | IVA |
| Low | 33** (774) | 42 (959) | 37 (660) | 40 (494) | 12 (235) | 12 (308) | 9 (120) | 11 (210) | --[a] (13) | 2 (127) | -- (23) | 5 (63) | 21 (185) | 25 (263) | -- (8) | 22 (58) |
| Medium | 15 (891) | 18 (1,021) | 18 (760) | 20 (697) | 20* (590) | 25 (1,092) | 18 (298) | 20 (511) | 0 (60) | 1 (587) | 2 (115) | 6 (232) | 27 (122) | 31 (216) | -- (14) | 31 (54) |
| High | 24 (354) | 30 (590) | 25 (409) | 31 (439) | 22 (103) | 26 (250) | 9** (53) | 27 (139) | -- (9) | 1 (79) | -- (42) | 4 (135) | -- (38) | 34 (58) | -- (5) | -- (13) |

*Significantly different from higher-aptitude comparison group at $p<.05$.
**Significantly different from higher-aptitude comparison group at $p<.01$.
[a]Fewer than 50 cases per occupation category.

# Enlistment and Utilization of High-Aptitude Recruits

Janice H. Laurence and Elizabeth F. Schneider
Human Resources Research Organization

High-aptitude personnel are an important element in the overall manning of our military forces. In fiscal year (FY) 1984 alone, over 123,000 persons scoring in AFQT Categories I and II (the upper 35 percent of the distribution) were accessed into the active enlisted forces.

Concerns over the cessation of banner recruiting times, the quality demands imposed by increasing military specialization and complex weaponry, and increased competition from the private sector for quality youth lead to questions as to the efficient utilization of high-aptitude military personnel.

Most studies addressing issues of manpower quality have focussed on persons scoring within the upper half of the ability distribution as a single entity. This research represents a departure from this tactic by examining the enlistment, assignment, and performance of individuals within the various aptitude levels traditionally labeled as "quality" personnel. More specifically, this paper compares the historical proportions enlisted, assignment patterns, and military performance of high-(AFQT Categories I and II) and average-aptitude (AFQT categories IIIA and IIIB) personnel.

**Historical Enlistment Trends.** Approximately five percent of the current 18 to 23 year old male youth population, is in AFQT Category I, 35 percent is in Category II, and 34 percent is in Category III (Department of Defense, 1982). As Table 1 shows, from the early 1950s through the present, the Defense Department has exceeded the national percentage within Categories I through III combined. However, DoD's proportion of very high-aptitude personnel (particularly within Category I) has not remained consistently better than the national average. Very high quality it seems, was easier to come by when the draft was in effect. Prior to the All-Volunteer Force (AVF), DoD was exceeding the national percentage of Category I male youth by as much as four percent.

Table 1

**Percent Distribution of Male Recruits (All Services Combined) by AFQT Category, FY 1952-83**

| Percent Distribution of Male Recruits[a] | | | | | | | |
|---|---|---|---|---|---|---|---|
| Fiscal Year | Category I | Category II | Category III | Fiscal Year | Category I | Category II | Category III |
| 1952 | 6.4 | 22.0 | 32.3 | 1973 | 3.7 | 30.1 | 52.1 |
| 1953 | 7.2 | 24.1 | 31.5 | 1974 | 3.0 | 32.3 | 54.5 |
| 1954 | 8.2 | 25.3 | 38.9 | 1975 | 3.5 | 34.0 | 56.3 |
| 1955 | 7.8 | 25.3 | 38.1 | 1976 | 3.9 | 33.9 | 51.7 |
| 1956 | 7.1 | 25.9 | 40.2 | 1977 | 4.2 | 28.2 | 39.6 |
| 1957 | 7.8 | 25.2 | 42.8 | 1978 | 3.6 | 27.3 | 42.1 |
| 1958 | 8.7 | 26.2 | 47.1 | 1979 | 3.0 | 23.6 | 41.8 |
| 1959 | 9.1 | 27.8 | 47.7 | 1980 | 2.8 | 23.8 | 41.0 |
| 1960 | 8.2 | 26.9 | 51.3 | 1981 | 2.6 | 30.2 | 47.8 |
| 1961 | 6.1 | 31.3 | 49.7 | 1982 | 3.1 | 33.4 | 49.4 |
| 1962 | 6.2 | 31.8 | 45.7 | 1983 | 3.7 | 36.7 | 50.1 |
| 1963 | 6.0 | 32.5 | 47.8 | 1984 | 3.5 | 36.8 | 51.1 |
| 1964 | 6.3 | 32.1 | 47.1 | | | | |
| 1965 | 5.5 | 31.3 | 48.8 | | | | |
| 1966 | 6.4 | 33.5 | 43.5 | | | | |
| 1967 | 6.6 | 33.1 | 39.7 | | | | |
| 1968 | 6.0 | 31.8 | 37.6 | | | | |
| 1969 | 6.2 | 31.7 | 37.7 | | | | |
| 1970 | 5.3 | 30.5 | 41.0 | | | | |
| 1971 | 5.1 | 30.0 | 43.1 | | | | |
| 1972 | 4.2 | 30.2 | 48.1 | | | | |

All Volunteer Force Transition[b]

Sources: Data for FYs 1952-70 were calculated from Annual Reports of the Quantitative Distribution of Military Manpower. Data for FYs 1971-83 were provided by the Defense Manpower Data Center.

[a]Male recruits include persons without prior military service who were inducted or enlisted and entered active duty (all Services combined) during the indicated fiscal year.

[b]The off-ramp end of the draft occurred on 30 June 1973. The drawdown began in July 1972, with the last draft call issued in December 1972.

With the advent of the AVF and continuing today, there are proportionally fewer very high-aptitude youth in the military than in the general population. The percentage of male recruits scoring in AFQT Category I has dropped to as low as 2.6 in FY 1981. In FY 1984 DoD's percentage was 3.8, still below that of the male youth population.

Assignment of High and Average Aptitude Recruits. The assignment patterns of FY 1979 through 1983 male high- and average-aptitude recruits across jobs of low, medium, and high skill requirements are presented below in Tables 2A-2D. Nine of the 10 DoD occupation codes were placed in the various skill requirement or job difficulty categories according to their mean ASVAB composite requirements (Sticht & Caylor, 1982) as follows:

| Low | Medium | High |
|---|---|---|
| Infantry | Administrative | Electronics Repair |
| Service/Supply | Electrical/Mechanical Repair | Communications |
| | Craftsmen | Medical Corps |
| | | Other Technical |

The "Non-Occupational" category was omitted from the present analyses.

Overall, across all Services, there was an inverse relationship between the proportion of male recruits and AFQT category for low skill level jobs. That is, there were 17% Category Is, 27% Category IIs, 32% Category IIIAs, and 39% Category IIIBs in low difficulty jobs. This pattern held among education groups. This inverse relationship was also found among medium difficulty jobs, though the differences between the four aptitude groups were not as pronounced. For high skill requirement jobs, the relationship was direct. That is, the higher the aptitude category, the greater the proportion of recruits assigned within those occupational areas. The proportions of Category IIIB to Category I recruits within high cognitive skill requirement jobs were 18%, 27%, 33%, and 48%, respectively.

Interesting findings emerged when the data were examined across job skill requirements within AFQT categories. First, as one might expect, with increasing job skill demands, there was a higher proportion of Category I recruits, with 17% in low, 35% in medium, and 48% in high difficulty jobs. This pattern did not hold for any other aptitude group examined.

Most AFQT Category II males were assigned to medium skill jobs (40%), followed by high (33%) and then low (27%). Yet another pattern was shown among job difficulty levels for AFQT Categories IIIA and IIIB recruits. Most of these average-aptitude personnel served in medium skill requirement occupational areas followed by low and then high skill jobs.

Although the above comments are generalizations drawn across Services, Tables 2A-2D show that the Services differed somewhat in their assignment patterns for high- and average-aptitude personnel. Furthermore, there were dramatic Service differences in the proportions of recruits of various aptitude categories assigned among low, medium, and high skill occupations. Such Service differences, perhaps, can be attributed in large part to the variation in Service specific missions, job structure, and aptitude mixes. The Army and Marine Corps, for example, have more combat or low skill requirement jobs while the Navy and Air Force have a greater proportion of highly technical jobs.

Table 2A (ARMY)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

| | | Occupational Area Technical Skill Requirement[a] | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Low | | Medium | | High | | Total[b] |
| AFQT Category | Education | N | % | N | % | N | % | N |
| CAT I | HSG | 3865 | 35.5 | 2146 | 19 8 | 4841 | 44.7 | 10832 |
| | NHS/GED | 258 | 34 5 | 171 | 22 9 | 319 | 42.6 | 748 |
| | Total | 4103 | 35.4 | 2317 | 20.0 | 5160 | 44.o | 11580 |
| CAT II | HSG | 36322 | 40.5 | 21385 | 23.9 | 31891 | 35.6 | 89598 |
| | NHS/GED | 8946 | 44.4 | 5336 | 26.5 | 5848 | 29.1 | 20130 |
| | Total | 45268 | 41.3 | 26721 | 24.4 | 37739 | 34 4 | 109728 |
| CAT IIIA | HSG | 24875 | 43.2 | 16032 | 27 8 | 16710 | 29 0 | 57617 |
| | NHS/GED | 13394 | 48.8 | 7733 | 28.2 | 6334 | 23.1 | 27461 |
| | Total | 38269 | 45 0 | 23765 | 27.9 | 23044 | 27.1 | 85078 |
| CAT IIIB | HSG | 36128 | 43.6 | 28518 | 34 4 | 18264 | 22 0 | 82910 |
| | NHS/GED | 22635 | 52.2 | 12417 | 28.6 | 8298 | 19.1 | 43350 |
| | Total | 58763 | 46.5 | 40935 | 32.4 | 26562 | 21.0 | 125260 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

[b] Those not in job status have been excluded

Table 2B (NAVY)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

| | | Occupational Area Technical Skill Requirement[a] | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Low | | Medium | | High | | Total[b] |
| AFQT Category | Education | N | % | N | % | N | % | N |
| CAT I | HSG | 264 | 2.3 | 6398 | 55 8 | 4809 | 41.9 | 11471 |
| | NHS/GED | 68 | 7 3 | 296 | 31 9 | 565 | 60.8 | 929 |
| | Total | 332 | 2.7 | 6940 | 56.0 | 5374 | 43.3 | 12400 |
| CAT II | HSG | 3925 | 4.9 | 39475 | 49.1 | 36934 | 46.0 | 80334 |
| | NHS/GED | 1590 | 10.0 | 7662 | 48.2 | 6652 | 41.8 | 15904 |
| | Total | 5515 | 5.7 | 47137 | 49.0 | 43586 | 45 3 | 96238 |
| CAT IIIA | HSG | 3671 | 9.0 | 21455 | 52.5 | 15550 | 38 0 | 40875 |
| | NHS/GED | 2314 | 16.1 | 8328 | 58.1 | 3703 | 25.8 | 14345 |
| | Total | 5985 | 10.9 | 29783 | 54.1 | 19253 | 35.0 | 55021 |
| CAT IIIB | HSG | 6142 | 14.7 | 24880 | 59.4 | 10837 | 25.9 | 41859 |
| | NHS/GED | 2617 | 21.0 | 7687 | 61.7 | 2162 | 17.3 | 12466 |
| | Total | 8759 | 16.1 | 32567 | 60.0 | 12999 | 24.0 | 54325 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June) Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

[b] Those not in job status have been excluded.



Table 2C (MARINE CORPS)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

| | | Occupational Area Technical Skill Requirement[a] | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Low | | Medium | | High | | Total[b] |
| AFQT Category | Education | N | % | N | % | N | % | N |
| CAT I | HSG | 854 | 28 8 | 971 | 32 7 | 1140 | 38.5 | 2965 |
| | NHS/GED | 116 | 39.2 | 175 | 59.5 | 89 | 30.1 | 296 |
| | Total | 970 | 29.7 | 1062 | 32.6 | 1229 | 37 7 | 3261 |
| CAT II | HSG | 12195 | 36.8 | 12140 | 36.6 | 8843 | 26.7 | 33178 |
| | NHS/GED | 3549 | 55.0 | 1824 | 28.3 | 1073 | 16.7 | 6447 |
| | Total | 15744 | 39.7 | 13965 | 35.2 | 9916 | 25.0 | 39625 |
| CAT IIIA | HSG | 11136 | 46.5 | 9009 | 37 6 | 3827 | 16.0 | 23972 |
| | NHS/GED | 4814 | 64.6 | 1714 | 23 0 | 923 | 12.4 | 7451 |
| | Total | 15950 | 50 8 | 10723 | 34.1 | 4752 | 15.1 | 31425 |
| CAT IIIB | HSG | 19860 | 57 0 | 11326 | 32.5 | 3705 | 10 6 | 34898 |
| | NHS/GED | 6735 | 73 8 | 1625 | 17.8 | 772 | 8 5 | 9132 |
| | Total | 26595 | 60.4 | 12951 | 29.4 | 4484 | 10.2 | 44030 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical

[b] Those not in job status have been excluded.

Table 2D (AIR FORCE)

Number and Percent of FY 1979-1983 Male Recruits by Occupational Area Technical Skill Requirement, AFQT Category and Education Level

| | | Occupational Area Technical Skill Requirement[a] | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Low | | Medium | | High | | Total[b] |
| AFQT Category | Education | N | % | N | % | N | % | N |
| CAT I | HSG | 800 | 8.6 | 2771 | 29.7 | 5754 | 61.7 | 9325 |
| | NHS/GED | 81 | 12.7 | 248 | 39.0 | 307 | 48.3 | 636 |
| | Total | 881 | 8.8 | 3019 | 30.3 | 6061 | 60.8 | 9961 |
| CAT II | HSG | 14288 | 16.7 | 32357 | 37.8 | 38998 | 45.5 | 85643 |
| | NHS/GED | 1723 | 17 2 | 4876 | 48 7 | 3409 | 34 1 | 10008 |
| | Total | 16011 | 16.7 | 37233 | 38.2 | 42407 | 44.3 | 95651 |
| CAT IIIA | HSG | 11711 | 22 5 | 27021 | 51.8 | 13391 | 25.7 | 52123 |
| | NHS/GED | 1367 | 20 7 | 4315 | 60.7 | 1235 | 18.7 | 6617 |
| | Total | 13078 | 22 3 | 31036 | 52.8 | 14626 | 25 0 | 58740 |
| CAT IIIB | HSG | 14858 | 27.1 | 33231 | 60.6 | 6710 | 12.3 | 54820 |
| | NHS/GED | 978 | 23 6 | 2777 | 67.1 | 385 | 9 3 | 4140 |
| | Total | 15836 | 26.9 | 36008 | 61.1 | 7116 | 12.1 | 58960 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J S. (1982, June). Low includes Infantry, Service and Supply Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical

[b] Those not in job status have been excluded.

As Table 2A shows, for the Army, there was no linear relationship between the proportion of Category I recruits and job difficulty. While most (45%) Category Is were in high cognitive demand jobs, low rather than medium demand jobs had the second highest proportion. For Category II through IIIB Army recruits, the low skill demand requirements had the highest proportion of recruits. High demand jobs had the second highest proportion of Category II recruits, though for the average-aptitude recruits their secondary concentration was in medium skill jobs.

The Marine Corps was somewhat similar to the Army in terms of assignment of recruits within jobs within various skill requirements. Category I recruits were concentrated in high jobs (38%), followed by medium (33%), and then low (30%). For Category II through IIIB recruits reverse trends were found with the greatest proportion in low, then medium, then high difficulty jobs.

The Navy showed the greatest proportion of recruits, in all examined categories, in medium skill requirement jobs (i.e., 56% of Category Is; 49% of Category IIs; 54% of Category IIIAs and 60% of Category IIIBs). The smallest proportion of all categories were found in low skill jobs (i.e., 3% of Category Is; 6% of Category IIs; 11% of Category IIIAs; and 16% of Category IIIBs).

Finally, the Air Force showed an easily comprehendable assignment pattern within aptitude categories. For both Category I and II recruits the pattern was, most in high jobs (61% and 44%, respectively) and least in low skill jobs (9% and 17%, respectively). The order of job difficulty types among Category IIIAs was medium (53%), high (25%) and low (22%); for Category IIIBs it was medium (61%), low (27%) and high (12%).

**Attrition and Job Difficulty.** Twelve-month attrition rates for FY 1979-1983 male recruits were calculated (by education) within each AFQT category and technical skill requirement cell and are presented below in Tables 3A-3D. Chi-squares (1x3s) were computed across job difficulty levels for each AFQT category. Additional chi-squares (1x4s) were computed across AFQT categories within each job difficulty level. All analyses were run separately by education. Alpha was set at the .05 level.

Overall, a trend toward the highest attrition rates in low skill requirement occupation areas and the lowest attrition rates in high skill requirement areas emerged for all aptitude categories and education levels. Significant $\chi^2$s however were not found consistently across Services or education levels. For the Army, significant $\chi^2$s were found for AFQT Categories I through IIIB among high school graduates, nongraduates, as well as for both education groups combined. For the Navy, a significant relationship between attrition rate and job difficulty was found only for Category I nongraduates. There were no such significant relationships for the Marine Corps. Significant relationships emerged again for the Air Force for nongraduates in all four aptitude categories.

Generally, aptitude did not have an overwhelming influence on attrition within job difficulty levels. Significant differences were found, however, for Army nongraduates within each skill requirement level and for the Army total education group within the high and low job demand areas. Though the Army showed an inverse linear relationship between AFQT category and attrition, the actual rate differences between contiguous categories were small. (Of course one must keep in mind that very small percentage differences could actually

Table 3A (ARMY)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

| AFQT Category | Education | Occupational Area Technical Skill Requirement[a] Low | Medium | High |
|---|---|---|---|---|
| CAT I | HSG | 7.1 | 4.7 | 3.1 |
| | NHS/GED | 12.8 | 5.5 | 7 5 |
| | Total | 7.4 | 4.9 | 3.4 |
| CAT II | HSG | 8.4 | 5.0 | 4.7 |
| | NHS/GED | 16.0 | 10.7 | 11.4 |
| | Total | 9.9 | 6 1 | 5.7 |
| CAT IIIA | HSG | 9.1 | 5.9 | 5.4 |
| | NHS/GED | 17.6 | 12.1 | 12.1 |
| | Total | 12.1 | 7.7 | 7.2 |
| CAT IIIB | HSG | 9.7 | 6 5 | 6.2 |
| | NHS/GED | 16.7 | 12 2 | 11.7 |
| | Total | 12.4 | 8.2 | 7.9 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T G , & Caylor, J S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Electrical/Mechanical Repair, and Craftsmen. High includes Administrative, Electronics Repair, Communication, Medical Corps, and other Technical.

Table 3B (NAVY)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

| AFQT Category | Education | Occupational Area Technical Skill Requirement[a] Low | Medium | High |
|---|---|---|---|---|
| CAT I | HSG | 2.3 | 1 7 | .8 |
| | NHS/GED | 4.4 | 1 4 | 1.1 |
| | Total | 2.7 | 1.7 | .8 |
| CAT II | HSG | 2.4 | 1.5 | 8 |
| | NHS/GED | 4.2 | 2.0 | 2 0 |
| | Total | 2.9 | 1 6 | 9 |
| CAT IIIA | HSG | 2.8 | 9 | 1.2 |
| | NHS/GED | 4.1 | 2.0 | 3 0 |
| | Total | 3 3 | 1 2 | 1 5 |
| CAT IIIB | HSG | 2.2 | 7 | 1.? |
| | NHS/GED | 3 o | 1.5 | 3.1 |
| | Total | 2 6 | 1 0 | 1.6 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.



Table 3C (MARINE CORPS)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

| AFQT Category | Education | Occupational Area Technical Skill Requirement[a] Low | Medium | High |
|---|---|---|---|---|
| CAT I | HSG | 2.0 | 1.1 | .4 |
| | NHS/GED | 9 | 1.1 | 1.1 |
| | Total | 1.9 | 1 1 | 5 |
| CAT II | HSG | 2.0 | .8 | .5 |
| | NHS/GED | 3.3 | 1.6 | 1.7 |
| | Total | 2.3 | 9 | 6 |
| CAT IIIA | HSG | 2.0 | .8 | 1.2 |
| | NHS/GED | 3.6 | 1.4 | 1.7 |
| | Total | 2 5 | 9 | 1.3 |
| CAT IIIB | HSG | 1.9 | 8 | 1.0 |
| | NHS/GED | 3.3 | 2.2 | 1.3 |
| | Total | 2.3 | 1.0 | 1.0 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T G., & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative, Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical.

Table 3D (AIR FORCE)

Percent Attrition for FY 1979-1983 Male Recruits
by Technical Skill Requirement, AFQT Category and Education Level

| AFQT Category | Education | Occupational Area Technical Skill Requirement[a] Low | Medium | High |
|---|---|---|---|---|
| CAT I | HSG | 3.3 | 2.6 | 1.0 |
| | NHS/GED | 11.1 | 8.1 | 5 2 |
| | Total | 4.0 | 3.0 | 1.2 |
| CAT II | HSG | 2.9 | 3.2 | 1.6 |
| | NHS/GED | 12.8 | 9.9 | 7.6 |
| | Total | 4.0 | 4.1 | 2.0 |
| CAT IIIA | HSG | 3.3 | 3.7 | 2.3 |
| | NHS/GED | 12.4 | 10.9 | 8.4 |
| | Total | 4.0 | 4 7 | 2 8 |
| CAT IIIB | HSG | 3.7 | 4.1 | 2 7 |
| | NHS/GED | 17.7 | 12.5 | 10 1 |
| | Total | 4.6 | 4 8 | 3.1 |

[a] Skill requirement was defined according to the mean ASVAB composite entry requirements of jobs within the DoD occupational areas as presented in Sticht, T.G , & Caylor, J.S. (1982, June). Low includes Infantry, Service and Supply. Medium includes Administrative Electrical/Mechanical Repair, and Craftsmen. High includes Electronics Repair, Communication, Medical Corps, and other Technical

account for large numbers of "attritees"). The significant differences in attrition among aptitude categories most likely resulted from the relatively greater attrition differences between extreme aptitude categories (i.e., between AFQT Category I and IIIB).

Not only was there a lack of statistical significance between attrition and aptitude category within the various job skill demand levels but inspection for "non-significant" trends showed little as well. Similar to the Army, Air Force data showed an AFQT category-attrition rate inverse linear relationship. Statistical significance, however, was found only among nongraduates. Again, though $\chi^2$s were computed across all four aptitude categories it appears that the differences were mainly between the extremes--Categories I and IIIB.

**Concluding Note.** Since the study of high-aptitude military personnel is yet in its beginning stages, extensive generalizations cannot (and should not) be drawn from these analyses. Despite this caveat, the assignment differences and attrition rates of high- and average-aptitude personnel among jobs of various skill requirements have been an informative first step in the investigation of the utilization of high-aptitude personnel. Service differences in assignment patterns and attrition rates within job difficulty levels abound. Regardless of such differences in the Services and their job requirements, high- and average-aptitude personnel are assigned differentially within the various skill demand occupational areas. Furtnermore, for high- and average-aptitude recruits there are some tendencies toward higher attrition rates in low skill jobs. A clearer picture of the relationship between job difficulty or complexity may emerge by using a taxononomy independent of Service selection and classification standards and the broad DoD occupational area categorizations. These and further analyses may have important implications for a more efficient distribution and matching of available talent within jobs, and overall more effective use of limited resources.

**References:**


Department of Defense (1982, March). Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics).

Sticht, T. G., & Caylor, J. S. (1982, June). Evaluation of the Literacy Assessment Battery (LAB) as a predictor of success in the Armed Services (FR-ETSD-82-10). Alexandria, VA: Human Resources Research Organization.

# Mapping Predictors to Criterion Space: Overview

Norman G. Peterson
Personnel Decisions Research Institute

## Introduction

Our applied problem is to expand the presently measured predictor space for the ultimate purpose of accurately selecting persons for the U.S. Army and appropriately classifying those persons into jobs or Military Occupational Specialities (MOS). In this paper, I describe the strategy we have adopted, the thinking behind the strategy, and some of the progress that has been made following our strategy. A fuller description can be found in Peterson, 1985.

As you all know, the U.S. Army presently has a lot of jobs and hires, almost exclusively, inexperienced and untrained persons to fill those jobs. One implication of these obvious facts is that a highly varied set of individual differences' variables must be put into use to stand a reasonable chance of improving the present level of accuracy of predicting training performance, job performance, and attrition/retention in a substantial proportion, if not all, of those jobs. Much less obvious is the particular content of that set of individual differences variables, and the way the set should be developed and organized; or put another way, how the predictors should be mapped onto the criterion space.

## Theoretical Approach

We have approached this problem by adopting a construct-oriented strategy of predictor development, but with a healthy leavening from the content-oriented strategy. Essentially, we endeavored to build up a model of predictor space by (a) identifying the major, relatively independent domains or types of individual differences' constructs that existed; (b) selecting measures of constructs within each domain that met a number of psychometric and pragmatic criteria, and (c) further selecting those constructs that appeared to be the "best bets" for incrementing (over present predictors) the prediction of the set of criteria of concern (i.e., training/job performance and attrition/retention in Army jobs). Ideally, the model would, we hoped, lead to the selection of a finite set of relatively independent predictor constructs that were also relatively independent of present predictors and maximally related to the criteria of interest. If these conditions were met, then the resulting set of measures would predict all or most of the criteria, yet possess enough heterogeneity to yield powerful, efficient classification of persons into different jobs. The development of such a model also had the virtue that it could be at least partially "tested" at many points during the research effort, and not just at the end, when all the predictor and criterion data are in. For example, we could examine the covariance of newly developed measures with one another and with the present predictors, notably the ASVAB. If the new measures were not relatively independent of ASVAB and measures from other domains as predicted by the model, then we could take steps to correct that. Also, by constructing such a visible model, we thought that modifications and improvements could be much more straightforwardly implemented.

Figure 1 presents an illustrative, construct-oriented model and is presented in order to represent the model in abstract. Note that both the criterion and predictor space are depicted. A great deal of the work of Project A is devoted to describing and defining the job performance criterion and we, on the predictor side, have made much use of the information coming from those efforts.

If this illustrative model were to be developed and tested with data, then the network of relationships on the predictor side, the criterion side, and between the two could be confirmed, disconfirmed, and/or modified. It goes without saying, but I will say it anyway, that the development of such models must be done very carefully and conservatively, and subjected frequently to reality testing. We have kept this firmly in mind, Note, however, that the possession of such a model enables one to state fairly clearly why such a predictor is being researched, and to check quickly, at least rationally, whether or not the addition of a predictor is likely to improve prediction.

Finally, the model is depicted as a matrix with a hierarchical arrangement of both the rows and columns. We have found it very useful to employ this hierarchical notion, since it allows us to think in terms of appropriate levels of specificity for a particular problem as we do the research, or for future applications of measures.

We began our research with a general kind of model, very much like the one presented in Peterson and Bownas (1982). That is, we conceived of the predictor space as divided into several domains with major, relatively independent constructs falling into each domain. At this early point in the research, we were most concerned with thinking about the predictor space in a way guided by past research that would also provide "handles," if you will, for us to approach our particular applied problem. We formed

| | | Criteria | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Training Performance | | | Job Task Performance | | Attrition/ Retention | | |
| Predictors | | Pass/ Fail | Test Grades | Attendance | Common Tasks | Specific Tasks | Finish Term | Reenlist | Early Discharge |
| Cognitive | Verbal | H* | H | L | H | H | L | L | L |
| | Numerical | H | H | . | . | . | . | | |
| | Spatial | | | | | | | | |
| Psychomotor | Precision | | | | | | | | |
| | Coordination | | | | | | | | |
| | Dexterity | | | | | | | | |
| Temperament | Dependability | | | | | | | | |
| | Dominance | | | | | | | | |
| | Sociability | | | | | | | | |
| Interests | Realistic | | | | | | | | |
| | Artistic | | | | | | | | |
| | Social | . | . | . | H | H | H | L | L |

FIGURE 1. Illustrative Construct-Oriented Model
*Denotes expected strength of relationship, High, Medium, Low.

three domain teams to be responsible for broad pieces of this predictor space model, to wit: a "non-cognitive" team for temperament, biographical data, and vocational interest variables; a "cognitive" team for cognitive and perceptual variables; and a "psychomotor" team for psychomotor variables.

Literature Review. The domain teams began with a large-scale literature review. Within each area, the teams carried out essentially the same steps. These were: 1) compile an exhaustive list of possibly relevant reports, articles, books or other sources; 2) review each source and determine its relevancy for the project by examining the title and abstract (or other brief review); 3) obtain the sources identified as relevant in the second step; and 4) for relevant materials, carry out a thorough review and transfer relevant information onto special review forms developed for the project.

Within the first step, several activities were carried out to insure as comprehensive a list as possible. Several computerized searches of relevant data bases were done. In addition to the computerized searches, we obtained reference lists from recognized experts in each of the areas, emphasizing the most recent research in the field. We also obtained several annotated bibliographies from military research laboratories. Finally, we scanned the last several years' editions of research journals that are frequently used in each ability area as well as more general sources such as textbooks, handbooks, and appropriate chapters in the Annual Review of Psychology.

The vast majority of the sources identified as described above were not relevant to our purpose. These non-relevant sources were weeded out in Step 2. After obtaining the relevant sources, these were reviewed and two forms were completed for each source: an Article Review form and a Predictor Review form (several of the latter form could be completed for each source.) These forms were designed to capture, in a standard format, the essential information from the reviewed sources, which varied considerably in their organization and reporting styles. The output of the literature search, in the form of the completed review forms and copies of the actual sources, served as input to several later steps.

Expert Judgments. One of these steps was the identification of a set of predictor constructs that met a number of psychometric and practical criteria. There were twelve such criteria used to evaluate constructs, like reliability, criterion-related validity, robustness and ease of administration procedures, etc. At least two researchers evaluated each construct on these twelve factors, using five point scales, and these evaluations guided the selection of 53 predictor constructs.

Definitions of these selected constructs were written and descriptive materials (psychometric data, validity evidence, and illustrative items) were prepared. These materials were used in an expert judgment process wherein 35 experienced personnel and research psychologists estimated the "true validity" of each of the 53 predictor constructs for each of 72 Army enlisted criteria. These 72 criterion descriptions were prepared by Project A researchers who were focusing on describing the job performance of Army enlisted ranks. (See Wing, Peterson, and Hoffman, 1984, for a complete description of this expert judgment process.)

These expert judgments proved to be highly reliable (the reliability of the pooled raters' estimates of validity of each construct for each criterion was over .90), and factor analysis of their ratings provided our first model of the predictor space. Figure 2 shows that model. This

| CONSTRUCTS | CLUSTERS | FACTORS |
|---|---|---|
| 1. Verbal Comprehension<br>5. Reading Comprehension<br>16 Ideational Fluency<br>18 Analogical Reasoning<br>21. Omnibus Intelligence/Aptitude<br>22. Word Fluency | A. Verbal Ability/ General Intelligence | |
| 4. Word Problems<br>8. Inductive Reasoning: Concept Formation<br>10. Deductive Logic | B. Reasoning | |
| 2. Numerical Computation<br>3. Use of Formula/Number Problems | C. Number Ability | COGNITIVE ABILITIES |
| 12. Perceptual Speed and Accuracy | H. Perceptual Speed and Accuracy | |
| 49. Investigative Interests | U. Investigative Interests | |
| 14. Rote Memory<br>17. Follow Directions | J. Memory | |
| 19. Figural Reasoning<br>23. Verbal and Figural Closure | F. Closure | |
| 6. Two-dimensional Mental Rotation<br>7. Three-dimensional Mental Rotation<br>9. Spatial Visualization<br>11. Field Dependence (Negative)<br>15. Place Memory (Visual Memory)<br>20. Spatial Scanning | E. Visualization/Spatial | VISUALIZATION/ SPATIAL |
| 24. Processing Efficiency<br>25. Selective Attention<br>26. Time Sharing | G. Mental Information Processing | INFORMATION PROCESSING |
| 13. Mechanical Comprehension | L. Mechanical Comprehension | MECHANICAL |
| 48. Realistic Interests<br>51. Artistic Interests (Negative) | M. Realistic vs. Artistic Interests | |
| 28. Control Precision<br>29. Rate Control<br>32. Arm hand Steadiness<br>34. Aiming | I. Steadiness/Precision | |
| 27. Multilimb Coordination<br>35. Speed of Arm Movement | D. Coordination | PSYCHOMOTOR |
| 30. Manual Dexterity<br>31. Finger Dexterity<br>33. Wrist-Finger Speed | K. Dexterity | |
| 39. Sociability<br>52. Social Interests | Q. Sociability | SOCIAL SKILLS |
| 50. Enterprising Interests | R. Enterprising Interest | |
| 36. Involvement in Athletics and Physical Conditioning<br>37. Energy Level | T. Athletic Abilities/Energy | VIGOR |
| 41. Dominance<br>42. Self-esteem | S. Dominance/Self-esteem | |
| 40. Traditional Values<br>43. Conscientiousness<br>46. Non delinquency<br>53. Conventional Interests | N. Traditional Values/Conventionality/Non delinquency | |
| 44. Locus of Control<br>47. Work Orientation | O. Work Orientation/Locus of Control | MOTIVATION/ STABILITY |
| 38. Cooperativeness<br>45. Emotional Stability | P. Cooperation/Emotional Stability | |

FIGURE 2. Hierarchical Map of Predictor Space

| PILOT TRIAL BATTERY | CLUSTERS | FACTORS |
|---|---|---|
| ASVAB | A. Verbal Ability/ General Intelligence | |
| Reasoning 1 and 2 | B. Reasoning | |
| Number Memory (c) | C. Number Ability | COGNITIVE ABILITIES |
| Perceptual Speed and Accuracy (c)<br>Target Identification (c) | H. Perceptual Speed and Accuracy | |
| AVOICE | U. Investigative Interests | |
| Short Term Memory (c) | J. Memory | |
| Reasoning 1 and 2 | F. Closure | |
| Assembling Objects<br>Object Rotation<br>Shapes<br>Mazes<br>Path<br>Orientation 1, 2, and 3 | E. Visualization/Spatial | VISUALIZATION/ SPATIAL |
| Simple Reaction Time (c)<br>Choice Reaction Time (c) | G. Mental Information Processing | INFORMATION PROCESSING |
| ASVAB | L. Mechanical Comprehension | MECHANICAL |
| AVOICE | M. Realistic vs. Artistic Interests | |
| Target Tracking 1 (c)<br>Target Shoot (c) | I. Steadiness/Precision | |
| Target Tracking 2 (c)<br>Target Shoot (c) | D. Coordination | PSYCHOMOTOR |
| - | K. Dexterity | |
| ABLE/AVOICE | Q. Sociability | SOCIAL SKILLS |
| AVOICE | R. Enterprising Interest | |
| ABLE | T. Athletic Abilities/Energy | VIGOR |
| ABLE | S. Dominance/Self-esteem | |
| ABLE | N. Traditional Values/Conventionality/Non-delinquency | MOTIVATION/ STABILITY |
| ABLE | O. Work Orientation/Locus of Control | |
| ABLE | P. Cooperation/Emotional Stability | |
| Cannon Shoot (c) | Movement Judgment | |

(c) = Computerized Measures

FIGURE 3. Pilot Trial Battery Measures of the Modeled Predictor Space

model represents the predictor structure in terms of their covariances with each other based on their judged validity relationships to dimensions of Army enlisted criteria.

Test Construction. Figure 2 served as a blueprint of sorts for our test construction efforts. The three domain teams set about writing tests and inventories to measure the constructs shown there. We went through a fairly extensive process of writing (or, in the case of computerized tests, programming) instruments, trying them out at Army sites (MEPS and/or Army forts), then revising the instruments based on the tryout results. After about four such iterations (at Minneapolis MEPS, Fts. Carson, Campbell, and Lewis), we possessed a set of instruments collectively labeled the Pilot Trial Battery. That set of measures is shown in Figure 3.

Note that the measures are slotted into the cluster and factor space, insuring that we adequately operationalized the model. Note also that one measure, "cannon shoot", is included and it measures Movement Judgment, a variable that was not originally included. It was added because it seemed to be a variable that was important for a variety of combat arms MOS, but had escaped our notice because of a dearth of research on such a variable.

This Pilot Trial Battery consumed approximately six and one-half hours of testing time and the entire battery was administered to a sample of about 250 soldiers at Ft. Knox. Test-retest data were also collected. Analyses of these data were used to further revise the measures and to reduce the battery in size so that it could be administered in four hours. The reduction in the size of the battery was accomplished by deleting some tests entirely and by deleting items from other tests. (The tests deleted were Reasoning 2, Shapes, Path, and Orientation 1.) The existence of the predictor model proved especially helpful to those of us faced with the hard decision of deleting tests and items. The impact of various decisions in terms of coverage of the "predictor space" could readily be seen and, along with the tryout data, empirically evaluated.

This revised and reduced battery was labeled the Trial Battery and is presently being administered to a large sample (N=11,000) of soldiers in the U.S. and Europe in a concurrent validity study. In terms of testing time, 34% of the battery is devoted to the computerized perceptual/psychomotor measures, 50% to cognitive paper-and-pencil measures, and 16% to non-cognitive, paper-and-pencil inventories. Once the concurrent validity data are in hand, we will be able to make some fairly definitive tests of our model--in terms of its factorial structure, validity, and classification efficiency.

## References


Peterson, N. G., & Bownas, D. A. (1982). Skill, task structure, and performance acquisition. In Marvin D. Dunnette and Edwin A. Fleishman (Eds.), Human performance and productivity (Vol. I). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Peterson, N. G. (1985). Overall strategy and methods for expanding the measured predictor space. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Wing, H., Peterson, N. G., & Hoffman, R. E. (1984). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto.

# Using Microcomputers for Assessment: Practical Problems and Solutions

Rodney L. Rosse and Norman Peterson
Personnel Decisions Research Institute

## Introduction

"History repeats itself" is an adage that probably does not apply to the advances of microprocessor developments. Given the frantic rate of development, it is difficult to imagine that circumstances could ever again occur in just the way that they did at the outset of this effort in the Fall of 1983. It would seem, however, that any 1986 project might be enhanced by consideration of both the occasional wisdom and sometime folly of our beginning efforts.

Initially, even the goals to be accomplished were far from obvious and may have remained beyond our vision except for the valuable help obtained through visits to several research centers doing advanced work in computerized testing: (1) Air Force Human Resources Laboratory at Brooks Air Force Base, Texas, (2) Army Research Institute Field Unit at Fort Rucker, Alabama, (3) Naval Aerospace Medical Research Laboratory, Pensacola, Florida, and (4) Army Research Institute Field Unit at Fort Knox, Kentucky. Experimental testing projects using computers at these sites had already produced impressive developments which stimulated the ideas of the project at hand and have continued to influence our work.

In this paper, we focus primarily on the process we followed and some problems we encountered in hardware and software acquisition and development for the purpose of developing new predictor tests of abilities that could best be administered via microprocessors.

## Hardware Acquisition and Development

Much of the detail of the planned products was yet to evolve at the point of acquisition of the first six machines so that we had to focus upon more general objectives. It was clear that we wished to accomplish several things which were either difficult or impossible to accomplish with paper-and-pencil testing. Specifically, we required the ability to have a very high degree of precision in stimulus presentation and a high degree of control of respondent behavior. Dependent variables were specifically expected to include precision in timing of stimulus presentation and response speed.

Microprocessor. The choice of which microprocessor to use for the preliminary development was not obvious. The arrays of available microcomputer devices were, at the time, in transition from earlier machines which used the first popular microprocessor chips (i.e., 8080 or Z-80) into a newer variety of options created by the influence of IBM's entry into the market with their "PC" employing the newer 8088, 8086-7 chips. With the newer machines came more flexible operating systems (e.g., DOS 1 or DOS 2).

A computer designed for portable use was deemed to be a highly desirable characteristic because the machines were to be frequently disassembled, carried to new locations, and reassembled by non-technical personnel. Such portable machines had been available only briefly so that little reported experience with them was available.

We acquired six machines made by Compaq (TM) which appeared to suit the need. They were among the "newer" types of machines which used a variation of the MS-DOS operating system. They were equipped with standard game adapters which permitted the analog inputs from "off-the-shelf" joysticks and boolean input from game button switches.

The choice was specifically made to avoid using color in the visual displays for at least two reasons: (1) the certainty of individual differences in color vision among military recruits, and (2) dread of the prospects of attempting to calibrate video colors for standardization of presentation. Accordingly, we precluded the possibility of directly investigating the value of stimulus effects in color presentation.

The graphics capability of the Compaq microcomputer proved to be minimally acceptable for the applications which were to come. In graphics mode, the pixels (or dots) on the screen are organized into 200 rows and 640 columns. More recently, several computers of the "personal" computer type are offering 400 rows with 640 columns which should provide improved resolution.

Very accurate timing of events occurring in the testing process was essential. Initially, timing was accomplished by two means: (1) accessing the calendar clock that is available in any machine which uses MS-DOS (or the variations of MS-DOS that are sold under computer tradenames), and (2) use of calibrated software loops. Without delving too far into technical details, those two options eventually presented some difficulties because of time consumption in the process of obtaining the time. For instance, the computer CPU often had to be tied up with timing events when other work required being done in the timed interval.

A wonderful solution to the timing problem eventually presented itself in what the computer people call a "real-time-clock" which can be added to the "IBM-type" microcomputers for as little as $50. Operating on a small battery it maintains the correct date and time even when the computer is turned off. With appropriate software, the "real-time-clock" device allows the timing of events accurately to the nearest 1/1000-th of a second with negligible loss of computer time in the reading. (The sub-program used in our projects will read the time in approximately 1/3000-th of a second.)

Peripheral Devices for Response Acquisition: Response Pedestal. The initial choices in the hardware configuration for a "testing station" proved satisfactory for the "stimulus side", i.e., the controlled presentation to the subject. The standard keyboard and the "off-the-shelf" joysticks were hopelessly inadequate for the "response side." Computer keyboards leave much to be desired as response acquisition devices--particularly when response latency is a variable of interest. Preliminary trials using, say, the "D" and "L" keys of the keyboard for "true" and "false" responses to items was troublesome with naive subjects. Intricate training was required to avoid individual differences arising from differential experience with keyboards. Moreover, the software had to be contrived so as to flash a warning when a respondent accidentally pressed any other key. The "off-the-shelf" joysticks were sadly lacking in precision of construction such that the score of a respondent depended heavily upon which joystick she/he was using.

We came up with a plan for a "response pedestal" which consisted of readily available electronic parts. A prototype of the device was obtained from a local engineer. (See Figure 1.) It had two joysticks, a horizontal and a vertical sliding adjuster, and a dial. The two joysticks allowed either left or right hand usage. The sliding adjusters permitted two-handed coordin-

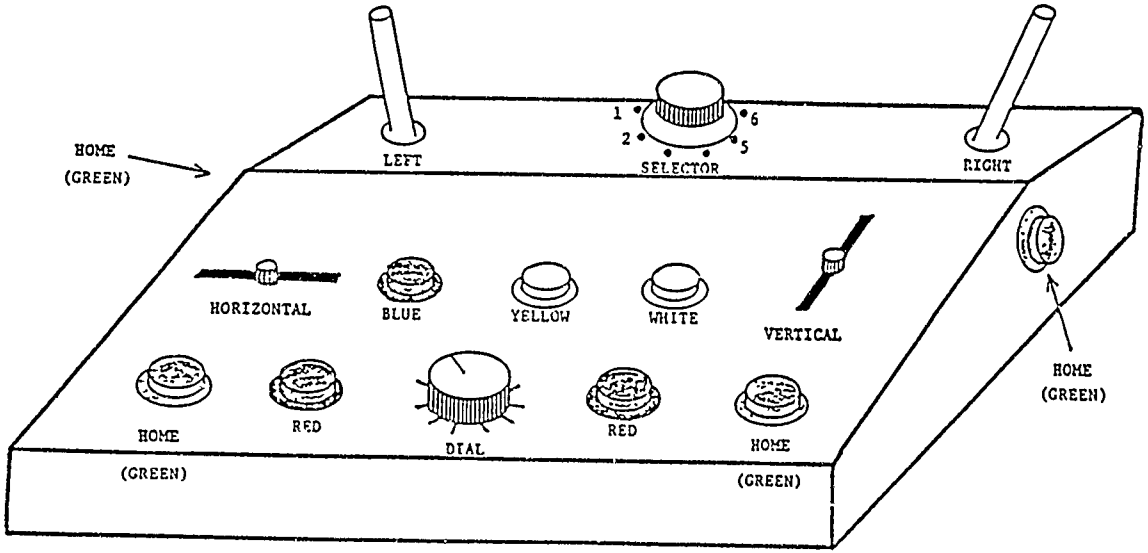



FIGURE 1. Custom-designed response pedestal

ation tasks. The dial permitted respondent selections in a manner similar to the now popular "mouse" devices that are sold for "personal computers."

The response pedestal had nine button-switches, each of which was to be used for a particular purpose. Three buttons (BLUE, YELLOW, and WHITE) were located near the center of the pedestal and were used for registering up to 3-choice alternatives. Also near the center were two buttons (RED) which were mostly used to allow the respondent to step through frames of instructions and, for some tests, to "fire" a "weapon" represented in graphics on the screen.

Of notable interest was the placement of the button-switches which were called "HOME" with respect to the positions of other buttons used to register a differential response. The "HOME" buttons required the respondent's hands to be in the position of depressing all four of the "HOME" buttons prior to presentation of an item to which (s)he would respond. This, it is believed, offered advantages of control of attention and control of hand position for measurement of response latency. Using appropriately developed software, we were able to measure total response time but also to break it down into two parts: (1) "decision time" which is defined as the interval between onset of stimulus and release of the "HOME" keys, and (2) "movement" time which is the subsequent interval to the registering of a response. It was possible, where of interest, to even tell quite reliably whether the respondent used a left hand or a right hand to respond since (s)he almost invariably would release the "HOME" buttons on the side of the preferred hand first.

The rotary switch marked "SELECTOR" in Figure 1 was an inconvenience that was required by our initial choice of "game-adapter" for reading analog input. The game adapter initially chosen allowed only four inputs and the response pedestal had seven analog outputs: 2 inputs for each of two joysticks, two sliding adjusters, and one rotary adjuster called the "DIAL." The "SELECTOR" was used to select which analog devices were to be operative for a particular test item. The final design for the response pedestal included a game-adapter with the capability of eight analog inputs and the "SELECTOR" switch was happily omitted.

Joysticks. Perhaps the greatest difficulty regarding the response pedestal design arose from the initial choice of joystick mechanisms. We soon discovered that joystick design is a complicated and, in this case, a somewhat controversial issue. Variations in tension or movement can defeat the goal of standardized testing. While "high-fidelity" joystick devices are available, they can cost thousands of dollars apiece which was prohibitively expensive in the quantities that were to be required for this project. The first joystick mechanism that was used in the response pedestals was an improvement over the initial "off-the-shelf" toys that predated the pedestals. It had no springs whatsoever so that spring tension would not be an issue. It had a small, light weight handle so that enthusiastic respondents could not gain sufficient leverage to break the mechanism. It was inexpensive.

Unfortunately, this joystick had a "wimpy" feeling which was greatly lacking in "face-validity" (or, as Hilda Wing dubbed it, "fist-validity") from the Army's point of view. It was felt that the joystick was so much like a toy that it would not command respect of the respondents. It was the contention of a minority of us that our "wimpy" device had "construct fidelity" in that it would do a perfectly adequate job of testing the constructs that were targeted.

The joystick mechanism had to be changed. Joysticks of every conceivable variety and type of use were considered. We learned about viscous dampening,

friction, tension, and even something called "stiction." Ultimately, a joystick device was fashioned with a light spring for centering and a sturdy handle with a bicycle handle-grip. It had sufficient "fist-validity" to be accepted by all (or almost all) and it was sufficiently precise in design that we were unable to detect any appreciable "machine" effects in fairly extensive testing.

## Software Development

We wish to turn attention now to the issues of software development. There were no "package programs" available to administer computerized tests. The selection of strategy for organizing and programming the needed software was to fall upon ourselves. We had three general, operational objectives in mind for the software to be produced: (1) as far as possible, it should be transportable to other microprocessors; (2) it should require as little intervention as possible from a test administrator in the process of presenting the tests to subjects and storing the data; and, (3) it should enhance the "standardization" of testing by adjusting for hardware differences across computers and response pedestals.

Primary Language. We chose to prepare the bulk of the software using the Pascal language as implemented by Microsoft, Inc. There were certain advantages to this in that Pascal is a common language and it is implemented using a compiler that permits modularized development and software libraries. As computer languages go, Pascal is relatively easy for others to read and it can be implemented on a variety of computers.

Some processes, mostly those which are specific to the hardware configuration had to be written in IBM-PC assembly language. Examples of these include the interpretation of the response pedestal inputs, reading of the real-time-clock registers, calibrated timing loops, and specialized graphics and screen manipulation routines. For each of these identified functions, a Pascal-callable "primitive" routine with a unitary purpose was written in assembly language. The functions were designed to be simple and unitary in purpose so as to be easily reproducible for other machines.

Strategy. The overall strategy of the software development is worth discussing. It quickly became clear that the direct programming of every item in every test by one person was not going to be very successful either in terms of time constraints nor in terms of quality of product. For the sake of making it possible for each researcher to contribute his/her judgment and effort to the project, it was necessary to plan so as to take the "programmer" out of the step between conception and product as much as possible.

The testing software modules were designed as "command processors" which interpreted relatively simple, problem oriented commands. These were organized in ordinary text written by the various researchers using word processors. Many of the commands were common across all tests. For instance, there were commands that permitted writing of specified text to "windows" on the screen and controlling the screen attributes (brightness, background shade, etc). A command could hold a display on the screen for a period of time (measured to 1/100-th second accuracy). There were commands which caused the program to wait for the respondent to push a particular button on the pedestal. Some of the commands were specific to particular item types. These commands were selected and programmed according to the needs of a particular test type. For each item type, we would decide upon the relevant stimulus

properties to vary and build a command that would allow the item writer to quickly construct a set of commands for items which she/he could then inspect on the screen.

Thus, entire tests were constructed and experimentally manipulated by psychologists who could not program a computer.

The strategies for developing commands have evolved and improved over the period of development. Eventually, the commands became almost "language-like" with syntax forms analogous to some of the common statistical packages like SPSS or SAS that are available on "main-frame" computers.

Hardware Testing and Calibration. One of the most useful software developments relates to the testing and calibration of the hardware, necessary for purposes of standardization. A complete hardware testing and calibration process can be undertaken by test monitors each time a machine is powered up. It checks the timing devices and screen distortion, and calibrates the analog devices (joysticks, sliding adjusters, dial) so that measurement of movement will be the same across machines. It also permits the software adjustment of the height to width ratio of the screen display so that circles do not become ovals or, more importantly, the relative speed of moving displays remains under control regardless of vertical or horizontal travel.

## Concluding Remarks

In the end, we were able to put together a portable, complete testing session lasting approximately 1-1/2 hours where very naive respondents can complete the test with little or no intervention from a test monitor. The data is automatically stored and "backed-up" on diskettes in a form readily transferrable to a "main-frame" for analysis. Except for occasional calibration or contingencies, the test monitor needs only to turn the computers on and put the respondents in front of them.

Finally, and perhaps most gratifying, we have found that the soldiers tested via this method have generally preferred computerized testing to paper-and-pencil testing. We have not gathered hard data on this aspect, but base our conclusions on observation of the soldiers while taking the battery and their comments to us after completing the battery. Perhaps this is due to novelty alone, but we feel it may also be due to the nature of the tests themselves plus the fact that the soldier, in large part, is in control of the testing process her/himself. They control the pacing of instructions for the tests and, for some tests, the pacing of item presentation. No administrator tells them when to begin and when to stop, and they are not in "lock step" with a larger group. We view this state of affairs as highly desirable for personnel selection testing.

# Computerized Assessment of Perceptual and Psychomotor Abilities

Jeffrey J. McHenry and Jody L. Toquam

Personnel Decisions Research Institute

One of the main goals of the Army Research Institute's (ARI's) Project A is to develop new predictor measures to supplement the Armed Services Vocational Aptitude Battery (ASVAB). In this paper, we describe 10 new computerized perceptual and psychomotor predictor tests that were pilot tested last fall and are currently being validated in a large-scale concurrent validation study.

## The Computer Battery

Toquam, Dunnette, Corpe, and Houston, (1985) have described the procedures used to identify target constructs for cognitive-perceptual predictor test development, and to determine which of these constructs would be measured via paper-and-pencil tests and which would be measured via computer. Following a similar procedure, members of the Project A research team working in the psychomotor ability domain identified two psychomotor ability constructs for predictor test development. Since measurement of both of these constructs required that subjects be presented with a moving stimulus object, it was decided that all psychomotor tests would be presented on the computer.

In total, computer tests were developed for seven constructs (i.e., five cognitive-perceptual ability constructs and two psychomotor ability constructs). To measure these seven constructs, 10 new computer tests were developed. The constructs and tests are listed in Table 1. As Table 1 shows, two tests each were developed to assess reaction time, perceptual speed and accuracy, and precision/steadiness, while one test each was developed to assess the remaining four constructs. (Complete descriptions of each test are available from the authors upon request.)

TABLE 1

Target Constructs and Computer Tests

| Target Construct | Definition | Test(s) |
|---|---|---|
| Reaction Time | The ability to detect a simple stimulus quickly | Simple Reaction Time<br>Choice Reaction Time |
| Perceptual Speed and Accuracy | The ability to compare two stimuli and to determine quickly and accurately whether they are the same or different | Perceptual Speed and Accuracy<br>Target Identification |
| Memory | The ability to encode and store information, and then retrieve that information quickly and accurately | Short Term Memory |
| Number Facility | The ability to perform simple numerical operations (e.g., addition, subtraction, multiplication, division) quickly and accurately | Number Memory |
| Movement Judgment | The ability to judge the movement speed and direction of an object and to determine when (or whether) that object will reach a given point in space | Cannon Shoot |
| Multilimb Coordination | The ability to coordinate the use of two or more limbs (e.g., two hands, two feet, a hand and a foot, etc.) to perform a task | Target Tracking 2 |
| Steadiness/ Precision | The ability to make fine coordinated movements in response to a moving stimulus object | Target Tracking 1<br>Target Shoot |

Pilot Testing

During test development, several pilot tests of portions of the computer battery were conducted at Ft. Carson, Ft. Lewis, and the Minneapolis Military Enlistment Processing Station. A more extensive pilot test of the entire battery was then conducted last fall at Ft. Knox.

The purpose of the pilot testing was to ensure that the tests satisfied three criteria for administration in the Project A concurrent validation study. First, we wanted to ensure that the 10 tests were reliable. Second, we wanted to make certain that the tests did not overlap greatly with the ASVAB. Finally, we wanted to ensure that the computer tests themselves are not highly intercorrelated, since our goal is to measure seven distinct ability constructs with these 10 tests.

## Method

Subjects

Subjects included 256 first-term Army enlisted personnel stationed at Ft. Knox. Subjects were drawn from a wide range of MOS. All subjects had been in the service between one and two years at the time of testing.

Procedure

When subjects arrived in the computer testing room, they were asked to take a seat at a testing station. They were told that the computer tests were self-administering so they could work at their own pace. They were instructed to read the instructions carefully, ask questions if they encountered any problems, and try their hardest.

Two weeks later, 121 of the subjects returned for retesting. They were given the same instructions that they had received two weeks earlier and asked to complete the entire computer battery a second time.

## Results

Scoring

Responses on computer tests may be used to compute numerous scores. For example, responses to Perceptual Speed and Accuracy items, may be summarized using average decision time, average movement time and average total response time across all items or across only those items in which the subject responds correctly. The average response for each of these may consist of the mean, the median or a trimmed mean computed by deleting the fastest and slowest response times. Other dependent measures derived from this test include the slope and intercept which are computed by regressing the subject's response time against some specified item parameter such as item length. Finally, percent correct can be used as a dependent measure for each subject.

In total, for the 10 tests, 168 different test scores were computed. Preliminary analyses of the reliability of each score and the intercorrelations among the various scores within each test were used to reduce this list to 19 test scores (see Table 2). These 19 scores received more extensive analyses.

Reliability

Table 2 contains the split-half and test-retest reliability for each test score. The majority of split-half reliabilities exceeded .80, and only two are less than .70. As expected, the test-retest reliabilities are lower than the split-half reliabilities. Five test scores have test-retest reliabilities less than .55. In general, those test scores with low test-retest

TABLE 2

Characteristics of the 19 Coaputer Test Scores

| Test Score | Reliability $r_{sh}$ | Reliability $r_{tt}$ | Overlap with ASVAB SMC | Overlap with ASVAB Uniqueness |
|---|---|---|---|---|
| COGNITIVE-PERCEPTUAL TESTS | | | | |
| Simple Reaction Time - Mean Rt | .90 | .37 | .07 | .83 |
| Choice Reaction Time - Mean Rt | .89 | .56 | .09 | .80 |
| Perc Speed & Acc - Pct Correct | .83 | .59 | .14 | .69 |
| Perc Speed & Acc - Mean RT | .96 | .65 | .06 | .90 |
| Perc Speed & Acc - Slope | .88 | .67 | .09 | .79 |
| Perc Speed & Acc - Intercept | .74 | .55 | .11 | .63 |
| Target Ident - Pct Correct | .84 | .19 | .05 | .79 |
| Target Ident - Mean RT | .96 | .67 | .16 | .80 |
| Short Term Memory - Pct Correct | .72 | .34 | .10 | .62 |
| Short Term Memory - Mean RT | .94 | .78 | .06 | .88 |
| Short Term Memory - Slope | .52 | .47 | .01 | .51 |
| Short Term Memory - Intercept | .84 | .74 | .11 | .73 |
| Number Memory - Pct Correct | .63 | .53 | .40 | .23 |
| Number Memory - Mean Oper RT | .95 | .98 | .33 | .62 |
| Cannon Shoot - Time Score | .88 | .66 | .02 | .86 |
| PSYCHOMOTOR TESTS | | | | |
| Target Tracking 1 - Mean Log Dist | .97 | .63 | .23 | .74 |
| Target Tracking 1 - Mean Log Dist | .97 | .77 | .17 | .80 |
| Target Shoot - Mean Time to Fire | .91 | .48 | .06 | .85 |
| Target Shoot - Mean Log Dist | .86 | .58 | .11 | .75 |

reliability are percent correct scores or scores with low split-half reliability.

Overlap with the ASVAB

The squared multiple correlation (SMC) between each test score and the 10 ASVAB subtests is also displayed in Table 2. These SMCs have been adjusted for shrinkage. Only for one test, Number Memory, does the SMC exceed .25. The median SMC across all 19 test scores is .10.

Table 2 also shows the uniqueness for each test score. This value represents an index of the unique (i.e, uncorrelated with the ASVAB) reliable variance of each test score. It is computed by subtracting the SMC with the ASVAB from the split-half reliability. All but two of the uniquenesses in Table 2 exceed .60. This information indicates that these 10 tests have much unique, reliable variance that may contribute to the prediction of job performance.

Overlap among the Computer Tests

Table 3 contains the intercorrelations among the 19 computer test scores. Well over half the intercorrelations between scores on different tests are less than .25, indicating that the various tests are measuring several different abilities.

To determine how we had fared in measuring our target constructs, a principal axis factor analysis was executed. Variables included 17 of the computer test scores (two variables, Perceptual Speed & Accuracy Mean RT and Short Term Memory Mean RT were withheld from the analysis since they correlated .82 and .83 with Perceptual Speed & Accuracy Slope and Short Term

TABLE 3

Intercorrelations among the ASVAB and Pilot Trial Battery (PTB) Tests
Ft. Knox Sample (N=168)

| | | SRT-RT | CRT-RT | PS&A-PC | PS&A-RT | PS&A Slp | PS&A-Int | Targ ID-PC | Targ ID-RT | STM PC | STM-RT | STM-Slp | STM-Int | Can Shoot | No Mem-PC | No Mem-RT | Trk 1-Dist | Trk 2-Dist | TSht-Time | TSht-Dist |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PTB - | SPT-RT | | | | | | | | | | | | | | | | | | | |
| Computerized | CRT RT | 53 | | | | | | | | | | | | | | | | | | |
| Cognitive | PS&A-PC | 17 | 17 | | | | | | | | | | | | | | | | | |
| Perceptual | PS&A-RT | 19 | 31 | 50 | | | | | | | | | | | | | | | | |
| Tests | PS&A-Slp | -03 | 09 | 52 | 82 | | | | | | | | | | | | | | | |
| | PS&A-Int | 32 | 31 | -27 | -08 | -61 | | | | | | | | | | | | | | |
| | Targ ID-PC | 11 | 07 | 40 | 33 | 32 | -13 | | | | | | | | | | | | | |
| | Targ ID-RT | 23 | 42 | 20 | 47 | 32 | 13 | 16 | | | | | | | | | | | | |
| | STM-PC | -06 | 04 | 50 | 17 | 26 | 23 | 25 | 03 | | | | | | | | | | | |
| | STM-RT | 23 | 40 | 26 | 49 | 25 | 23 | 27 | 47 | 0A | | | | | | | | | | |
| | STM-Slp | -06 | 03 | 28 | 09 | 26 | -11 | 18 | 13 | 32 | 39 | | | | | | | | | |
| | STM-Int | 28 | 42 | 10 | 35 | 11 | 31 | 18 | 42 | -11 | 83 | -19 | | | | | | | | |
| | Can Shoot | 13 | 10 | 00 | 08 | -01 | 11 | 09 | 25 | 02 | 25 | 14 | 18 | | | | | | | |
| | No Mem PC | -16 | -09 | 29 | 02 | 15 | -20 | 14 | -12 | 23 | 00 | 02 | 01 | -16 | | | | | | |
| | No Mem-RT | 21 | 24 | 11 | 34 | 21 | 03 | 10 | 27 | 07 | 18 | 15 | 11 | 08 | -45 | | | | | |
| PTB | Trk 1-Dist | 14 | 25 | -12 | 08 | 00 | 11 | 04 | 42 | 29 | 25 | 15 | 35 | 27 | 14 | 00 | | | | |
| Computerized | Trk 2-Dist | 11 | 19 | -01 | 11 | 04 | 09 | 02 | 39 | -19 | 25 | -01 | 27 | 30 | 14 | 02 | 81 | | | |
| Psychomotor | TSht-Time | 08 | 16 | 16 | 22 | 09 | 12 | 12 | 32 | 09 | 22 | 15 | 15 | 12 | -10 | 15 | 23 | 19 | | |
| Tests | TSht-Dist | 08 | 16 | 07 | 03 | -00 | 09 | -11 | 32 | -12 | 27 | -16 | 38 | 25 | -08 | 02 | 60 | 55 | 15 | |

Memory Intercept, respectively), scores from the 10 paper-and-pencil tests described by Toquam et al. (1985), and scores from the 10 ASVAB sub-tests. The sample included only those 168 subjects for whom complete data from all three sets of tests were available. Factor solutions were rotated using the VARIMAX method.

The 7-factor solution was judged the most interpretable. Significant loadings (i.e, greater than .35) for each test score on each factor are shown in Table 4. Based on the factor loadings, we named Factors I-VII general ability, spatial ability, psychomotor ability, general accuracy, basic processing speed, number facility, and a response style factor, respectively. For four of the seven factors (psychomotor ability, general accuracy, basic processing speed, and the response style factor), no paper-and-pencil tests load significantly on these factors. All but one of the tests with significant loadings on the spatial ability factor were paper-and-pencil tests. Both the ASVAB and the computer battery included tests with significant loadings on the other two factors, general ability and number operations; however, the only computer test scores with significant loadings on these factors was Number Memory. Thus, once again, Number Memory appears to be the only computer test that overlaps significantly with the ASVAB.

Some of the factors that include computer tests are moderately similar to the target constructs that we set out to measure with the computer battery. Basic processing speed, for example, contains measures from three target constructs: reaction time, perceptual speed and accuracy, and memory. The number facility factor includes Number Memory test scores, as we had hoped, and also includes the Coding Speed and Number Operations sub-tests from the ASVAB. Finally, the psychomotor ability factor includes

measures of both our target psychomotor ability constructs, multilimb coordination and steadiness/precision.

As Table 4 shows, the time score from Cannon Shoot failed to load significantly on any of the five factors. This indicates that the movement judgment ability tapped by this test differs from the abilities assessed by the other computer tests. This provides indirect evidence that the movement judgment test is measuring a unique perceptual ability, as we had hoped it would.

TABLE 4

Results from a Principal Components Factor Analysis of Scores on the ASVAB, Cognitive Paper and Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer Tests*
(N = 168)

| Variable | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Factor 7 | $h^2$ |
|---|---|---|---|---|---|---|---|---|
| ASVAB GS | 75 | | | | | | | 59 |
| ASVAB AR | 75 | | | | | | | 73 |
| ASVAB WK | 77 | | | | | | | 62 |
| ASVAB PC | 62 | | | | | | | 47 |
| ASVAB NO | | | | | | 84 | | 77 |
| ASVAB CS | | | | | | 62 | | 44 |
| ASVAB AS | 62 | | | | | | | 50 |
| ASVAB MK | 77 | | | | | | | 70 |
| ASVAB MC | 63 | 58 | 30 | | | | | 68 |
| ASVAB EI | 72 | | | | | | | 65 |
| Assemb Obj | 35 | 69 | | | | | | 66 |
| Ob, Rotation | | 61 | | | | | | 49 |
| Shapes | | 66 | | | | | | 51 |
| Pa.es | | 70 | | | | | | 67 |
| Path | | 67 | 30 | | | | | 65 |
| Reason 1 | 37 | 58 | | | | | | 54 |
| Reason 2 | 37 | 47 | | | | | | 44 |
| Orient 1 | 37 | 64 | | | | | | 58 |
| Orient 2 | 40 | 46 | | | 30 | | | 52 |
| Orient 3 | 60 | 52 | | | | | | 67 |
| Ski RT | | | | | 63 | | | 44 |
| CRT PT | | | | | 61 | | | 50 |
| PSBA PC | | | | 67 | 31 | | | 70 |
| PSBA Slope | | | | 88 | | | | 81 |
| PSBA Inter | | | | 65 | 50 | | | 74 |
| Target ID PC | | | | [illegible] | | | | 25 |
| Target ID RT | | 41 | 37 | | 30 | | | 57 |
| STM PC | | | | 39 | | | 34 | 41 |
| STM Slope | | | | | | | 41 | 25 |
| STM Int | | | 36 | | 51 | | | 47 |
| Cannon Shoot | | | 32 | | | | | 19 |
| NM PC | 53 | | | | | 37 | | 52 |
| NM RT | 37 | | | | | 64 | | 54 |
| Tracking 1 | | | 86 | | | | | 82 |
| Tracking 2 | | | 77 | | | | | 66 |
| Target Shoot TF | | | | | | | 42 | 23 |
| Target Shoot Dist | | | 64 | | | | | 48 |
| Variance Explained | 5 65 | 4 70 | 2 84 | 2 37 | 1 92 | 1 87 | 1 17 | |

*Note that the following variables were not included in this factor analysis AMQT, PSBA Reaction Time and Short Term Memory Reaction Time

(Please also note that decimals have been omitted.)

## Discussion

All of the tests except Simple Reaction Time yielded at least one test score with split-half reliability in excess of .80 and test-retest reliability in excess of .55. Thus, we met our first goal, which was to ensure that all the computer tests attained adequate levels of reliability.

Our second goal was to ensure that the new computer tests were not redundant with the ASVAB. SMCs between the 19 test scores and the ASVAB tended to be quite low. Uniquenesses indicated that the computer tests had the potential to contribute a great deal of unique, reliable variance to the prediction of job performance. Thus, we also met our second goal.

Analyses designed to evaluate the intercorrelations among the new tests showed that the various tests generally shared little common variance. Results from a factor analysis indicate that there were at least five (and probably six) different ability factors underlying performance on the 10 tests; these factors are moderately similar to the target constructs we set out to measure. It is important to note here that results from the factor analysis must be considered tentative at best because the sample size includes only 168 subjects. Data obtained from the ongoing concurrent validity study with over 10,000 subjects will provide us with more stable information about our constructs and the relationships among those constructs.

Generally, we felt that the results of the pilot testing indicated that only minor modifications were required in the tests prior to concurrent validation testing. Our observations of subjects during pilot testing suggested a number of changes in the instructions for virtually all of the tests. The split-half reliability data indicated that several of the tests could be shortened without any significant impact on test reliability. Finally, there was some evidence (not discussed in this paper, but noted in McHenry & McGue, 1985) that the two Target Tracking Tests should be made more difficult and that the Target Shoot Test should be made easier. Aside from these, few modifications were made in the computer battery prior to concurrent validation testing. (See Toquam, Dunnette, Corpe, McHenry, Keyes, McGue, Houston, Russell & Hanson, 1985, for more detailed information regarding changes in the computerized perceptual tests.)

Presently, concurrent validation testing is winding down. By the middle of next month, we will have collected predictor and criterion data on almost 10,000 first-term Army enlisted personnel in 19 MOS. It is our hope that at this time next year, we will be able to present some initial validity data for our 10 new computerized perceptual and psychomotor tests.

## References


McHenry, J. J., & McGue, M. K. (1985). *Problems, issues, and results in the development of computerized psychomotor measures.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., McHenry, J. J., Keyes, M. A., McGue, M. K., Houston, J. S., Russell, T. L., & Hanson, M. A. (1985). *Development of cognitive/perceptual measures: Supplementing the ASVAB.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., & Houston, J. S. (1985). *Adding to the ASVAB: Cognitive/perceptual measures.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

# ADDING TO THE ASVAB: COGNITIVE PAPER-AND-PENCIL MEASURES

Jody L. Toquam, Marvin D. Dunnette, VyVy A. Corpe, and
Janis Houston

Personnel Decisions Research Institute

## Introduction

The purpose of this paper is to (1) identify the cognitive/perceptual ability constructs that supplement or provide information about Army applicants' abilities not currently tapped by the Armed Services Vocational Aptitude Battery, or ASVAB; (2) describe the measures developed for paper-and-pencil administration and the cognitive/perceptual constructs they are designed to tap; (3) describe test development issues and the factors used to evaluate the psychometric quality of the new paper-and-pencil measures; and (4) report the relationships between scores on the ASVAB and scores on the new paper-and-pencil tests. Information about the cognitive/perceptual measures designed for computer administration are described in McHenry and Toquam (1985).

Before describing the new tests, we first examine the content of the current military selection and classification battery, the ASVAB, and then provide a brief review of the process involved in identifying the constructs for inclusion in the Pilot Trial Battery. (The Pilot Trial Battery is the term used for the battery of experimental tests administered at Minneapolis MEPS, Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. This battery includes twelve paper-and-pencil measures - ten cognitive and two non-cognitive, and ten computerized measures - seven cognitive/perceptual and three psychomotor.)

The current military selection and classification battery, the ASVAB, contains ten subtests. Scores on four of these are used to calculate the Armed Forces Qualification Test (AFQT) score which is used to determine qualification for entrance into the Army. Scores on the ten subtests are used in different combinations to determine applicants' qualifications for different military occupational specialties (MOS). Results from a factor analysis of ASVAB scores indicate that the battery assessed verbal ability, speeded performance, quantitative ability, and technical knowledge (Kass, Mitchell, Grafton & Wing, 1982).

Peterson (1985) describes the activities involved in identifying ability constructs that supplement information obtained from the ASVAB. Those activities included a review of the literature which was used to impose structure on the domain (i.e., establish a cognitive/perceptual abilities taxonomy) and then to summarize validity data for the different types of ability constructs. This information was input to the expert judgment task. All of this information was used to identify cognitive/perceptual ability constructs that tap abilities relatively independent of those measured by the ASVAB and that may be used to improve the Army's selection and classification decisions process.

Cognitive/perceptual ability constructs selected for inclusion in the Pilot Trial Battery and their designated priorities (in parentheses) are: (1) Spatial Visualization - Rotation and Scanning;

(2) Spatial Visualization - Field Independence; (3) Spatial Orientation; (4) Induction - Figural Reasoning; (5) Reaction Time - Processing Efficiency; (6) Memory - Number Operations; (7) Memory - Short Term Memory; (8) Perceptual Speed and Accuracy.

## Determining the Method of Administration

In this section, we review the factors that influenced our decision to measure a particular construct via paper-and-pencil or via computer. The first factor concerns the construct definition and the dependent measures suggested by that definition. For example, definition of the construct, processing efficiency, indicates that the dependent measure involves the time required to respond to simple stimuli. Such information can only be obtained on a computer because a precise measure of reaction time is required. Hence, those constructs that involve a reaction time component, such as Processing Efficiency, Perceptual Speed and Accuracy, and Memory were slated for computer administration. McHenry and Toquam (1985) provide a detailed description of measures developed for computer administration.

The second factor involves the cost related to adapting items to the computer. For example, test items for such constructs as spatial visualization and figural reasoning involve detailed figures and objects. To adapt these items to the computer would require high resolution graphics. The cost for hardware capable of supporting such graphics at the time was prohibitive. Thus, we determined that measures of spatial visualization, spatial orientation, and induction would be assessed via paper-and-pencil. We focus on the development activities and pilot-test results of the new paper-and-pencil measures in the remainder of this paper.

## Paper-and-Pencil Measures: Construct and Test Descriptions

In this section, we provide definitions of the constructs, describe criterion job performance areas or tasks that we expect measures of the constructs to predict and finally identify the tests designed to measure each construct. Detailed descriptions of the individual tests are available from the authors.

Spatial Visualization--Rotation

This involves the ability to mentally restructure or manipulate parts of a two- or three-dimensional figure. It serves as a potentially effective predictor of success in MOS that involve mechanical operations, construction and drawing or using maps. Two tests developed to measure this construct include Assembling Objects and Object Rotation.

Spatial Visualization--Scanning

This includes the ability to visually survey a complex field and to find a pathway through it. According to our expert judges, measures of this construct are potentially effective as predictors of success for Army MOS involving electrical or electronics operations, using maps in the field, and controlling air traffic. The two measures designed to assess this construct in the Path Test and the Maze Test.

Spatial Visualization--Field Independence

This includes the ability to find a simple form when it is hidden in a complex pattern. A measure of this construct is expected to predict success in MOS that involve detecting and identifying targets, using maps in the field, planning placement of tactical positions, air traffic control and troubleshooting operating systems. The Shapes Test was developed to measure this construct.

Spatial Orientation

This involves the ability to maintain one's bearing with respect to points on a compass and to maintain appreciation of one's location relative to landmarks in the environment. From job observations conducted in the field, we expect measures of this construct to predict success in combat MOS that involve maintaining directional orientation using features of landmarks in the environment. Three tests involving different orientation tasks were developed to assess this construct, Orientation 1, Orientation 2, and Orientation 3.

Induction - Figural Reasoning

This includes the ability to generate hypotheses about principles governing relationships among several objects. According to the panel of experts, measures of this construct are effective predictors of success in MOS involving troubleshooting, inspecting, and repairing electrical, mechanical, or electronic systems, analyzing data, controlling air traffic, and detecting and identifying targets. We developed two tests involving different tasks to assess abilities in this construct area. These were titled Reasoning 1 and Reasoning 2.

## Test Development Issues

Two issues impacted on our approach for developing the new paper-and-pencil measures. These include the target population completing the new tests for selection and classification purposes and the power versus speed components of each new test. We discuss each in turn below. The population completing these tests is the same population that completes the ASVAB to qualify for entrance into to the Army. This is, very generally speaking, a population composed of predominantly recent high school graduates, not entering college, from all geographic sections of the United States. For our purposes the target population was, practically speaking, inaccessible during the test development phase. We were constrained to using enlisted soldiers to try out the newly developed tests. The development group, enlisted soldiers, of course, represents a restricted sample because they have passed enlistment standards and often have completed basic and advanced individual training.

Differences between the target population and the sample available to us, lead to two major implications that served as general guidelines for test development and pilot testing activities. First, the target population includes a broad range of abilities, therefore we attempted to develop test with a broad range of item difficulties. And second, the the test development group, first-term enlistees, would be of generally higher in ability than the target population. Therefore, the overall difficulty level of the test should be somewhat higher (i.e.,the test should be somewhat easier) than what it would have been if we had access to an unrestricted sample of the target population.

Another decision to be made about each test was its placement of the power vs. speed continuum. Most psychometricians would agree that a "pure" power test is a test administered such that all persons taking the test are allowed enough time to attempt all items on the test, and that a "pure" speeded test is a test administered such that no one or very few taking the test has enough time to attempt all items. In practice, there appears to be a power/speed continuum, most tests fall somewhere between the two extremes on this continuum.

During the preliminary test development stage, we categorized each test as a power test, speeded test, or combination of the two using our construct definitions. For example, using our definition of Induction, we designed the test items to represent a very wide range of difficulty levels and established a generous time limit such that most subjects would have time to complete all items. Thus, measures of induction were designed to fall on the power end of the continuum. Our plan for measures tapping Spatial Visualization -Rotation and Scanning differed from this in that all items were constructed to be moderately easy but more restrictive time limits were imposed. Thus, these measures were intended to fall toward the speeded end of the continuum.

For the remaining constructs, Spatial Visualization-Field Independence and Spatial Orientation, we designed the measures using the construct definitions to determine the range of item difficulties and to establish time limits. Following each pilot-test we examined completion rates and item difficulty levels to assess how closely performance on each new test matched the corresponding construct definition with regards to speed and power components.

## Evaluating the Paper-and-Pencil Tests

Four pilot test or tryout sessions were conducted at Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. In the first pilot-test at Fort Carson, about 38 soldiers completed each paper-and-pencil test. The number at Fort Campbell was 57 and at Fort Lewis it was 118. At Fort Knox the numbers were 290 for time one and 97 to 126 for time two. Factors used to evaluate each test at one or more of these pilot-test sessions include the following: construct validity, test item characteristics, and test reliability. Below we present some general findings for all paper-and-pencil tests.

One goal of the the pilot-test sessions was to verify the construct validity of the new measures. Therefore, we identified published tests that measure constructs similar to our construct definitions. These published measures were included in the first three pilot-tests. It is important to note that, in general, most published tests or marker tests differed from the new tests in item difficulty levels and in the specific task required. Therefore, we did not expect a one-to-one correspondence between the new test and its published marker test.

Very few of the newly developed tests correlated above .65 with the designated marker; most correlations between new measures and marker tests fell between .45 and .60. These values were as expected given the differences in task requirements and in item difficulty levels between the new and marker tests. Basically this information suggested to us that although the tests did not duplicate their respective marker tests, they captured the essence of the target con-

struct.

Another goal of the pilot-test sessions was to assess the psychometric characteristics of each new test. Following the pilot-test sessions, then we computed item difficulty levels and item-total correlations for each test. These data were used to modify test items and to adjust time limits.

Results from the first pilot test indicated that all tests required some modification. For example, completion rates, item difficulty levels and raw total test scores suggested that some of the new measures may suffer from ceiling effects. Thus, for Assembling Objects, Object Rotation, Path Test, and Orientation 1, we constructed new items and adjusted time limits accordingly to obtain the desired difficulty level. For the Shapes Test and Maze Test, we modified test items to increase difficulty levels and to reduce the possibility of ceiling effects.

The reverse situation appeared on one of the orientation tests, Orientation 1. That is, item difficulty levels were low or the test was more difficult than desired. We modified this test by adding four "easy" items and by exparding the time limit.

For the remaining measures, Orientation 3, Reasoning 1, and Reasoning 2 very few changes were required. For example, item analysis data revealed that for some of the items, item-total correlations were higher for a distractor than for the correct response. These items were either modified or replaced.

Subsequent pilot tests indicated that the tests, in general, required only minor modifications.

Finally, we investigated the reliability or internal consistency and the stability of each new measure. To compute internal consistency estimates we used a split half procedure. This included administering each test as two separately timed halves and computing the correlation between part one and part two for each test. The Spearman-Brown correction procedure was then used to estimate the reliability for the test as a whole. We estimated the stability of each test by collecting test-retest data on a sample of about 100 soldiers at Fort Knox. A period of two weeks separated the two test sessions.

Internal consistency and test-retest estimates for each test appear in Table 1. Results from the Fort Lewis pilot-test indicate that the split half internal consistency estimates range from the high 70's to the low 90's for all tests with the exception of Reasoning 2. Test-retest estimates are lower than the internal consistency estimates but are at acceptable levels ranging from .57 to .84. The Reasoning 2 test once again yields the lowest value of all.

Note that in Table 1, we have also included internal consistency estimates for the Fort Knox sample computed using the Hoyt formula. and may represent overestimates for some of the more highly speeded tests. With the exception of Reasoning 2, these values range from the low 80's to high 90's.

TABLE 1

Reliability and Uniqueness Estimates for the Ten Paper-and-Pencil Tests Included in the Pilot Trial Battery

| | | | Reliability | | | Uniqueness | |
|---|---|---|---|---|---|---|---|
| | | | Ft Levels | Ft Knox | | | |
| Test | No Items | Time Allotted (in minutes | $r_{xx}$ Split Half N = 119 | Alpha N = 290 | $r_{xx}$ Test retest (N = 97 to 126) | ASVAB $R^2$ Using Split Half | ASVAB $R^2$ Using Split Half |
| Assembling Objects | 40 | 11 | 79 | 92 | .74 | 40 | .39 |
| Object Rotation | 90 | 7 5 | 66 | .97 | 75 | 19 | .67 |
| Mazes | 24 | 5 5 | 78 | 89 | 71 | 25 | 53 |
| Path | 64 | 8 | .82 | 92 | 69 | 29 | 53 |
| Shapes | 54 | 16 | 82 | 92 | .70 | .19 | .63 |
| Reasoning 1 | 30 | 12 | .78 | 83 | 64 | 29 | 49 |
| Reasoning 2 | 32 | 10 | 63 | 65 | 57 | .26 | .37 |
| Orientation 1 | 150 | 20 | 92 | 98 | .67 | 36 | .56 |
| Orientation 2 | 24 | 10 | 89 | .88 | 80 | 30 | 59 |
| Orientation 3 | 20 | 10 | 88 | .90 | 84 | 54 | 34 |

## Overlap Between the New Measures and the ASVAB

As we above, the major focus of this research involves identifying and developing measures of constructs not currently assessed in the ASVAB. One way to estimate the amount of overlap between each new measure and the measures contained in the ASVAB is to conduct uniqueness analyses. This procedure involves computing the squared multiple correlation between each new test and the ten ASVAB subtests. The resulting value is then subtracted from the reliable variance in that new measure (in this case we used the reliability estimate computed using the split half procedure). This value represents an index of the unique variance or variance that is uncorrelated with scores obtained on the ASVAB. Results from this analysis are reported in the final two columns in Table 1.

Across the ten new tests, the squared multiple correlations range from .54 to .19. It is clear that some of these tests are measuring abilities tapped by ASVAB subtests. On the other hand, the uniqueness estimates which range from .67 to .34, indicate that the new tests tap abilities independent from those assessed by the ASVAB subtests.

In sum, results from the uniqueness analysis are essentially what we would expect in assessing the amount of overlap between groups of tests that measure cognitive/perceptual abilities. The data are encouraging because they indicate that we are measuring ability constructs not currently assessed by the ASVAB.

## REFERENCES


Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1982). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10: 1981 Army applicant sample. (Technical Report 581). Alexandria, VA: U.S. Army Research Institute.

McHenry, J. J., & Toquam, J. L. (1985). Computerized assessment of perceptual and psychomotor abilities. Paper presented at the Military Testing Association Conference, 24 October, San Diego, CA.

Peterson, N. G. (1985). Mapping predictors to criterion space: Overview. Paper presented at the Military Testing Association Conference, 24 October, San Diego.

# Measuring Personal Attributes: Temperament, Biodata, and Interests

Leaetta M. Hough, Matt K. McGue, John D. Kamp,
Janis S. Houston, and Bruce N. Barge

Personnel Decisions Research Institute

Overview. I'm going to describe the development and evaluation of temperament, biographical, and interest measures - what we call non-cognitive measures - included in the Project A predictor battery. Non-cognitive measures were included in the predictor battery because of their potential for predicting important on-the-job criteria, criteria such as Effort, Initiative, Following Regulations and Orders. Adjustment, Leadership, and Self-Control.

The information I will present today suggests: 1) that non-cognitive predictors are likely to predict such criteria; in fact, more likely to predict such criteria than are other types of predictors; 2) that non-cognitive measures contribute unique variance to the predictor battery and are, therefore, likely to contribute incremental validity; 3) that the non cognitive measures we developed have good psychometric characteristics, they are internally consistent and show high test-retest reliability; and 4) that faking on personality inventories is not the problem it is often assumed to be. Our overall strategy was: to review the literature on temperament, biodata, and interest to identify constructs that were likely to be criterion valid; to obtain expert judgments about expected true validity of those constructs; to develop measures of those constructs; to remove or revise sensitive or objectionable items; and to evaluate and revise measures based on their internal consistency, overlap with other predictors, and their stability across time and different motivational conditions.

Literature Review Results. Our review and summary of the literature indicated that the validity of interest measures for important Army criteria were in the high .20s. The validities of biographical inventories for such criteria were in the .20s and .30s. These results were not too different from previous literature reviews. Our conclusions for the personality literature, however, differ from some of the other reviews, and I'd like to describe these results more thoroughly.

The criterion-related validities reported in the literature for temperament constructs are shown in Table 1. As you can see, the adjustment criterion, which includes such things as unfavorable discharge and drug abuse, is predicted very well by temperament measures. The predictor constructs Achievement and Locus of Control also predict Educational, Training, and Job Proficiency criteria. These results differ from those reported by Guion and Gottier in their 1965 Personnel Psychology article. Our results are, however, similar to those reported by Ghiselli in his 1973 Personnel Psychology article. We believe the results are explained by the approach we used.

Our approach was to develop a predictor taxonomy and to classify temperament scales into the taxon or construct with which they were most similar. We accomplished this classification by searching the literature

Table 1

Summary[a] of Criterion-Related Validities of Temperament Constructs

| Temperament Construct | Type of Criterion: Educational | Training | Job Proficiency | Job Involvement | Adjustment |
|---|---|---|---|---|---|
| Potency (Surgency) | .06 (42)[b] | .13 (36) | .07 (65) | .04 (13) | -.17 (31) |
| Adjustment | .14 (43) | .19 (28) | .11 (65) | .17 (16) | [-.33] (52) |
| Agreeableness (Likeability) | .03 ( 9) | .08 ( 5) | .03 (22) | -.02 ( 5) | -.03 ( 5) |
| Dependability | .13 (24) | .12 (20) | .11 (49) | .14 (15) | [-.43] (40) |
| Intellectance (Culture) | .17 ( 6) | .19 ( 5) | .01 (16) | -.09 ( 9) | .18 ( 3) |
| Affiliation | -.03 ( 5) | --- --- | -.02 ( 6) | .09 ( 4) | -.07 ( 4) |
| Achievement | [.30] ( 8) | [.33] ( 4) | [.24] ( 4) | --- --- | [-.33] ( 5) |
| Masculinity | -.16 ( 8) | .09 ( 3) | .10 (10) | .03 ( 4) | -.13 (11) |
| Locus of Control | [.32] ( 1) | [.29] ( 2) | [.25] ( 7) | --- --- | --- --- |
| Unclassified Military Scales | --- --- | .18 ( 8) | .18 (25) | --- --- | [-.22] (20) |

[a] Medians are reported as the summary index.
[b] The number in parentheses is the number of correlations on which the median is based.
NOTE: Median correlations greater than .20 are indicated by a box.

for reported correlations between temperament scales and then using these correlations to categorize the temperament scales into the five factors identified by Tupes and Christal (1961) in their peer rating research. We then added four constructs to the taxonomy to increase the homogeneity of the constructs. We also used a taxonomic system for the criteria. These consisted of Educational, Training, Job Proficiency, and Adjustment criteria.

We then summarized the criterion-related validities reported in the literature according to our predictor and criterion taxonomies. Guion and Gottier did not summarize the literature according to constructs; Ghiselli, however, reported results only for studies for which he felt the predictor was conceptually appropriate for the criterion. Our literature review, which summarized the reported validities according to a data-based classification of scales into constructs, supports Ghiselli's results and conclusions. We believe the construct approach highlighted the predictor-criterion relationships by reducing the "noise," if you will, and that the Guion and Gottier approach masked such relationships.

Expert Judgments of True Validity. Using the construct approach, we identified the temperament constructs that were likely to yield good criterion-related validities. We then asked experts to estimate the expected true criterion-related validities of predictor constructs for important Army criteria. These estimated validities also indicated that the non-cognitive predictors were likely to predict Army criteria - criteria such as Initiative/Effort, Following Regulations and Orders, Leading and Supporting, Self-Control, and others in the .20s, .30s, and even .40s. I might add that the cognitive and psychomotor measures were not expected to predict these criteria nearly as well.

Development of Construct Measures. Using the results of the literature review and expert judgments, we identified "good bets" for predicting important Army criteria. We developed scales to measure these constructs.

We wrote temperament and biodata items for the ABLE, which stands for Assessment of Background and Life Experiences, and we wrote interest and biodata items for the AVOICE, which stands for Army Vocational Interest Career Examination. We also developed four "response validity scales" which we called Social Desirability, Poor Impression, Self-Knowledge, and Non-Random Responses and included the items in these four response validity scales in the ABLE.

We next examined the ABLE and AVOICE items for sensitivity, or the extent to which people might object to the content of the questions. The Army and their scientific advisors also reviewed the items for sensitive content. We revised or removed the objectionable items and administered the ABLE and AVOICE to soldiers at Ft. Lewis, Ft. Campbell, and Ft. Knox. After each administration we examined the psychometric characteristics of the items and scales and revised them for each subsequent administration.

The last administration was at Ft. Knox where about 275 soldiers completed the ABLE and AVOICE. We evaluated the scales for internal consistency, test-retest reliability, and their unique contribution to the predictor battery. For the ABLE scales, the median internal consistency was .84, with a range of .70 to .87. For the AVOICE, the median was .86, with a range of .68 to .96. About 125 soldiers returned two weeks later to complete the ABLE and AVOICE a second time. The median test-retest coefficient for the ABLE was .79, with a range of .68 to .83. For the AVOICE, the median test-retest was .76, with a range of .56 to .86. Uniqueness analyses we conducted show that both the ABLE and AVOICE share very little variance with the ASVAB or with the cognitive and psychomotor tests included in the predictor battery. In short, the psychomotor characteristics of both the ABLE and AVOICE are very good; they are internally consistent, stable over time, and likely to contribute incremental validity to the predictor battery.

Faking Study. The next issue we addressed was faking. The concern was that self-report measures are susceptible to intentional distortion. We, therefore, conducted a faking study, the purpose of which was 1) to determine the extent to which soldiers can distort their responses to temperament and interest inventories when instructed to do so; 2) to determine the extent to which the ABLE response validity scales detect intentional distortion; 3) to determine the extent ABLE response validity scales can be used to adjust or correct scores for intentional distortion; and 4) to determine the extent to which distortion is a problem in an applicant setting.

We gathered data from 125 Army applicants people who wanted to be accepted into the Army and would have a motive for distorting their responses; we used the Ft. Knox data as an honest comparison sample; and we conducted an experiment in which soldiers were instructed to respond honestly or to distort their responses in a specified way.

The participants in the experimental group were 245 enlisted soldiers at Ft. Bragg. We created four faking conditions: fake good on the ABLE, fake bad on the ABLE, fake interest in combat activities on the AVOICE, and fake interest in non-combat activities on the AVOICE. We also created two honest conditions: honest on the ABLE, and honest on the AVOICE.

The design was a repeated measures with faking and honest conditions counter-balanced. Thus, approximately half the experimental group, or 124 soldiers, completed the inventories honestly in the morning and faked in the afternoon, while the other half (121 soldiers) completed the inventories honestly in the afternoon and faked in the morning. In summary then, we had a 2 x 2 x 2 fixed-factor, completely crossed experimental design.

We performed a multivariate analysis of variance on the ABLE and AVOICE scales separately. All the relevant fake x set interactions for the ABLE were significant at the .01 level, indicating that soldiers can distort their responses. The fake x set x order interactions, significant at the .05 level, indicate that the order in which the conditions occurred has a significant effect on scores. We performed a multivariate analysis of variance on the AVOICE scales and found similar results; people can distort their responses to an interest inventory.

Another research question was the extent to which the response validity scales detected intentional distortion. The results indicate that the Social Desirability scale detects faking good; the effect size of the difference between the means for the honest and fake good conditions is 1.02, or one standard deviation. The Poor Impression scale detects faking bad; the effect size of the difference between the means for the honest and fake bad conditions is 2.67, or just over two and one-half standard deviations.

We next examined the extent to which we could use the response validity scales, Social Desirability and Poor Impression, to adjust ABLE content scales and AVOICE occupational scales for faking. We regressed out Social Desirability from the fake good condition and Poor Impression from the fake bad condition. Table 2 shows the median effect sizes between the honest and faking conditions for the ABLE and AVOICE scales before and after regressing out Social Desirability and Poor Impression. The median difference in ABLE scores between the honest and fake good condition before regressing out Social Desirability is .49 or half a standard deviation. That is, ABLE scale scores differ by about half a standard deviation in the fake good condition as compared to the honest condition. After regressing out Social Desirability from the fake good condition, the ABLE content scales are only .14, or just over 1/10 of a standard deviation, different from the honest condition.

The median difference in ABLE scores between honest and fake bad before regressing out Poor Impression for is 2.10. That is, ABLE content scale scores in the fake bad condition differ by approximately two standard deviations from ABLE content scales in the honest condition. However, after regressing out Poor Impression from the scales, the difference is less than half a standard deviation. Clearly, the response validity scales Social Desirability and Poor Impression can be used to adjust scale scores for the ABLE for intentional distortion. We do not know, however, whether the adjustment formula will cross-validate and be as effective in another data set. Nor do we know whether adjusting the scale scores improves the criterion-related validity of the scales. It may be that the unadjusted scale scores are more criterion-valid than adjusted scores.

We performed the same computations for the AVOICE occupational scales and

Table 2

Effects of Regressing Out Response Validity Scales
(Social Desirability and Poor Impression)
in Faking Conditions for ABLE and AVOICE

| | Honest vs Fake Good/Combat Effect Size | | Honest vs Fake Bad/Non-Combat Effect Size | |
|---|---|---|---|---|
| | Before Adjustment | After Adjustment | Before Adjustment | After Adjustment |
| ABLE Content Scales | .49 | .14 | 2.10 | .45 |
| AVOICE Combat Scales | .43 | .33 | .97 | .86 |
| AVOICE Combat Support Scales | .53 | .39 | .49 | .34 |

Median values are reported.

found that the results are not nearly as impressive. The bottom two rows show the median effect size of the differences between the honest and faking conditions before and after regressing out the appropriate response validity scale for the AVOICE.

These data demonstrate that: 1) people can distort their responses to temperament and interest scales, 2) response validity scales detect such distortion, and 3) the response validity scales can be used to adjust temperament scale scores for distortion. However, the question remains: To what extent do applicants distort their responses? To answer this question we compared scale scores from the Ft. Bragg experimental honest condition and the Ft. Knox honest condition with the scale scores of approximately 120 Army applicants. These comparisons suggest that applicants do not appear to distort their responses. As shown in Table 3, the applicant means on the temperament scales (ABLE content scales) are lower than one or both of the honest means nine out of eleven times. The results for the AVOICE are similar. In short, applicants do not tent to distort their responses.

Summary. To briefly summarize our approach and results: we identified constructs and developed measures of constructs that had demonstrated criterion-related validity in the past and were judged by expects as likely to be criterion-valid for important Army criteria. The measures we developed contributed unique variance to the predictor battery, were internally consistent or homogeneous, and yielded reliable and stable scale scores across time and motivational conditions.

Our next step is to criterion-validate these measures with Army criteria. Data gathering for that is currently underway.

Table 3

Comparison of Ft. Bragg Honest*, Ft. Knox, and MEPS (Applicants) ABLE Scales

| ABLE Scale | Ft. Bragg (Honest)* N | Ft. Bragg (Honest)* Mean | MEPS (Applicants) N | MEPS (Applicants) Mean | Ft. Knox N | Ft. Knox Mean | Total S.D. |
|---|---|---|---|---|---|---|---|
| Response Validity Scales | | | | | | | |
| Social Desirability (Unlikely Virtues) | 116 | 15.91 | 121 | 16.63 | 276 | 16.60 | 3.21 |
| Self-Knowledge | 116 | 29.54 | 121 | 28.03 | 276 | 29.64 | 3.63 |
| Non-Random Response | 116 | 7.58 | 121 | 7.79 | 276 | 7.75 | .64 |
| Poor Impression | 116 | 1.50 | 121 | 1.05 | 276 | 1.54 | 1.84 |
| Content Scales | | | | | | | |
| Emotional Stability | 112 | 66.22 | 118 | 66.03 | 272 | 65.05 | 7.86 |
| Self-Esteem | 112 | 34.77 | 118 | 34.04 | 272 | 35.12 | 5.00 |
| Cooperativeness | 112 | 53.73 | 118 | 54.60 | 272 | 54.19 | 6.05 |
| Conscientiousness | 112 | 46.37 | 118 | 46.49 | 272 | 48.97 | 5.86 |
| Non-Delinquency | 112 | 53.24 | 118 | 54.36 | 272 | 55.49 | 6.91 |
| Traditional Values | 112 | 36.67 | 118 | 36.97 | 272 | 37.28 | 4.50 |
| Work Orientation | 112 | 59.71 | 118 | 58.37 | 272 | 61.40 | 7.73 |
| Internal Control | 112 | 49.48 | 118 | 51.90 | 272 | 50.37 | 6.13 |
| Energy Level | 112 | 57.56 | 118 | 56.67 | 272 | 57.19 | 6.95 |
| Dominance (Leadership) | 112 | 35.54 | 118 | 32.84 | 272 | 35.41 | 6.05 |
| Physical Condition | 112 | 32.96 | 118 | 28.27 | 272 | 31.08 | 7.49 |

*Scores are based on persons who responded to the honest condition first.

## References


Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Guion, R. M., & Gottier. (1965). Validity of personality measures in personnel section. Personnel Psychology, 18, 135-164.

Tupes, E. C., & Christal, R. E. (May, 1961). Recurrent personality factors based on trait ratings (ASD-TR-61-97). Lackland Air Force Base, TX: Aeronautical Systems Division, Personnel Laboratory.

## SYMPOSIUM OVERVIEW

# THE TRAINING AND SELECTION OF ARMY MANAGERS: QUANTITATIVE/QUALITATIVE APPROACHES

Gerald P. Fisher, Human Resources Research Organization (HumRRO)

Richard Lilienthal, Army Civilian Personnel Center (CIVPERCEN)

The U.S. Army's civilian workforce currently exceeds 450,000 employees. As this widely dispersed workforce becomes increasingly dependent on highly skilled employees capable of keeping pace with rapidly advancing technologies, major personnel management challenges are presented. Among those challenges are the needs of a personnel system to select, train, assign, promote and retrain employees and managers within the Army. The presentations this morning include the following: Selecting and Training Logistics Managers; Designing an Executive Development Program; and Managerial Competencies for LOGAMP.

This Military Testing Association symposium concerns three projects related to training and selecting Army civilian professionals, managers and/or executives. In order to promote and develop a diverse and talented workforce, emphasis needs to be placed on recruiting, selecting, promoting and training generalists who can meet the challenge of leadership and change. Each of the three presentations focuses on some aspect of how training requirements can be established and sequenced to effectively train specialist supervisors in the direction of generalist managers. Through promotion, cross-training, management development programs, and other methods, it is presumed that the goal of selecting and producing a well trained generalist can be accomplished.

At the same time, the technical sophistication of the equipment and technology that the new Army must deal with today and tomorrow demands that civilian professionals and managers have state-of-the-art training, the capability of handling increasingly complex concepts and the skills of a specialist. The generalist/specialist dichotomy runs through each of our presentations. Which skills are essential and what knowledge(s) must be possessed depend on using reliable and defensible data gathering and decision making methods. Thus each of these three presentations relies on both in-person interviews as well as questionnaires and formal data gathering instruments. While sample size varied from 10,000 incumbents in the logistics study to as few as 10 executives in the finance center effort, each of the studies does rely on more than one data gathering method.

All presenters this morning are outside the field for which they are recommending training or selection criteria. Each of the papers will hopefully shed light on two related issues. The first is the question of the qualitative role of the analyst at key stages of each of the studies. Each of us in the symposium recognizes that when we ask a question of a group of job incumbents that we in some way influence the response that is received.

When we analyze a data set we also bring to bear our experiences and perspectives. One of the issues in this panel is the role of the analyst and how his or her experience and perspectives influence eventual policy recommendations and training or selection decisions. A second issue of concern in these presentations is that of sample size and the value and limits in using either or both large and small sample data gathering methods. Each of the speakers will discuss this variable in our presentations.

One of the key issues addressed in the logistics study (Lilienthal, Fisher, Hough) is whether a single job analysis can be used effectively for multiple purposes. In our paper we describe a job analysis project in which both training and selection uses evolve from the same task data base. Both management and employees complain about the presence of separate job analyses by hiring, training, and classification personnel. A single procedure that would serve all of management's job analytical needs has been a goal of job analysts for years. This project has not discovered that procedure. It does, however, employ a true multi-purpose procedure in that the products will be used in both selection and training systems.

The three major uses of job analysis, i.e., selection, classification, and training, have different goals and needs. For selection purposes, the analyst typically looks for similarities among jobs. This is because the more that jobs are found to have in common, the more they can share common selection procedures. General aptitude batteries such as DoD's Armed Forces Vocational Aptitude Battery (ASVAB) and Office of Personnel Management's former Professional and Administrative Career Examination (PACE) could not exist without commonality across a number of jobs. Tasks need not be written at a detailed level of specificity if the resulting knowledges, skills, and abilities (KSAs) are general dimensions such as deductive reasoning, inductive reasoning, spatial, numerical, oral, and writing abilities. At the extreme of this viewpoint are supporters of the validity generalization concept who suggest that a selection-oriented job analysis can be quite simple, needing only to demonstrate that the job under study belongs to a general class of jobs for which selection validation data exist.

For classification and pay purposes, the analyst typically looks for differences among jobs. This is because the analyst must develop or support a system which assumes differences among jobs (e.g., the Federal civilian classification system which assumes that there are hundreds of series, each having multiple pay grades). To obtain these differences, the analyst writes task statements that are specific to series and pay grades.

For training purposes, the analyst needs more detailed tasks and KSAs than for selection purposes. One reason is that the task of the trainer is to bring novices to the full performance level. While the selection specialist can concentrate on only the tasks and KSAs that distinguish superior job performance, the training specialist must be concerned with teaching everything necessary to perform the job. Another reason is that learning objectives are more specific than dimensions such as oral ability or writing ability.

This project uses the same task inventory for selection and training purposes. The dual purpose does not become apparent until the SME panels. The level specificity of the task statements is that required for training.

Although that degree of detail is not required for selection purposes, there is no problem caused by having it. Job classification is not a purpose of this project. If it were, many more tasks delineating level of supervision and responsibility would be added. For selection and training purposes such tasks are not needed because they do not add anything useful to the analysis. For example, the same KSAs are needed to write a report as to review and approve it.

The two purposes manifest themselves in the SME panels, not just in the rating scales used, but in the very nature of the KSAs. There are typically more KSAs identified for training, i.e., they are written at a more detailed level of specificity. Some KSAs appear as products for both purposes but in different forms. For example, in the Army Civilian Personnel Administration career program, Army pay policy might be identified for training purposes while Federal pay policy would be identified for selection purposes. Although we would teach Army policy we would not want to use an Army-specific KSA in the selection system. Agency-specific selection KSAs tend to restrict competition and result in organizational "in breeding." Other KSAs appear as products for only one purposes in any form. For example, we offer courses on stress management but we do not rate and rank applicants on stress management. Because of regulations and litigation our criteria for KSAs are more strict for selection than training purposes.

In summary, each of these projects found that the former methods of a few decision makers and staff members getting together to decide the content of promotion elements or training requirements has had to be replaced by a more rigorous and legally defensible multi-method approach. We look forward to sharing details of these projects with the readers of the papers and with the audience present at MTA.

# SELECTING AND TRAINING LOGISTICS PROFESSIONALS AND MANAGERS: QUALITATIVE/QUANTITATIVE APPROACHES

Gerald Fisher
Human Resources Research Organization

Richard Lilienthal
Army Civilian Personnel Center

Leaetta Hough
Personnel Decisions Research Institute

This paper focuses on methodological issues involved in one of the three studies in this symposium--a Comprehensive Occupational and Data Analysis Programs (CODAP) job analysis producing selection criteria and training requirements for GS/GM 11-15 Army civilians in three career programs--Supply Management, Materiel Maintenance Management, and Transportation Management. To understand the need for this job analysis, it is necessary to outline several recently developed Army managerial training, career development and selection programs.

The Army is in the process of developing two systems for selecting and training civilian managers. The Army Civilian Career Evaluation System (or ACCES) is the result of a joint effort between the U.S. Army Civilian Personnel Center (CIVPERCEN) and the U.S. Office of Personnel Management to improve the current civilian promotion and referral system, known as the SKAP system. ACCES requires a rigorous task analysis effort so that the job tasks required can be quantitatively specified. Once the task clusters are identified, appropriate knowledges, skills, and abilities (KSAs) can be specified for promotion and referral purposes throughout the Army. ACCES will be the Army's future centralized evaluation and referral system and has thus far been implemented in two civilian career programs--Manpower and Force Management and Civilian Personnel Administration. This particular study looks at the problems of developing selection criteria and training requirements in three other career programs within a single effort.

The second purpose for which this job analysis was conducted is the newly developed Army Civilian Training, Education and Development System (ACTEDS). ACTEDS is targeted toward improving the development of the Army's civilian work force through systematic technical, professional, and managerial training and development. The Army recognizes that ACTEDS is needed because:

> As presently designed, the civilian training and development system does not fully support the progressive development of Army's future top civilian managers. Contrary to the desired orderly, systematic approach to technical, professional, and managerial skills training, civilian employees typically participate in programs on a self-initiated rather than management planned basis. In most

instances, the training and assignments they receive are not sequentially interrelated to contribute to progressively increasing and strengthening the experience and knowledge base over their entire career.

One of the first civilian career programs in which ACTEDS is being implemented is the Logistics and Acquisition Management Program (LOGAMP) consisting of all GS/GM 11-15 managers in six selected Army career programs: Contracting and Acquisition; Quality and Reliability Assurance Engineers and Scientists, Material Maintenance Management; Supply Management; and Transportation Management. (The last three career programs noted are being analyzed in this study.) The method for developing ACTEDS training requirements, as well as selection criteria for promotion within the ACCES program, is through a CODAP-based job task analysis inventory followed by subject matter expert (SME) workshops wherein KSAs are developed both for training requirements and for setting selection criteria.

## Method

Since September 1984, the Personnel Decisions Research Institute (PDRI), joined by the Human Resources Research Organization (HumRRO), has been conducting a CODAP-based job analysis for the Army Civilian Personnel Center (CIVPERCEN). Individual job task and KSA lists for the 20 job series within the three career programs were initially developed. The lists were based on a review of 2,000 position descriptions (out of 10,000 job incumbents). Using the review of current classification and qualification standards for each series as well as the initial inventories as a starting point, a sample of nearly 400 incumbents was interviewed in small group meetings to add, modify, or eliminate task statements. The analysts merged the individual inventories, resulting in a single job description inventory of more than 300 task statements covering all three career programs. The single LOGAMP task inventory was then distributed to all 10,000-plus job incumbents at Army installations throughout the world. Following receipt of completed inventories, CODAP analyses are to be conducted and pertinent task and duty clusters will be developed and validated by SMEs.

## Development of the CODAP Inventory

CODAP was originally developed by the Air Force for occupational analysis purposes in the mid to late 1950s. CODAP includes a set of interrelated statistical programs used for the purpose of job analysis.

The primary focus of CODAP is the analysis of jobs, jobs being defined as a set of tasks selected from a standard inventory. The selection of tasks is performed by the job incumbent to reflect those tasks actually performed by him or her. In addition to identifying the tasks, the job incumbents are asked to categorize the relative time spent performing each task. When these "time spent" categories are assigned numerical weights and converted to percentages, individual job descriptions are produced. In this inventory we asked incumbents to rank both relative time spent (to other tasks) and importance to job on each task performed.

In our current CIVPERCEN project, we plan to analyze the data on the basis of empirical task/duty clusters, rather than on previously developed job series within the various career programs that are being reviewed. Furthermore, we have attempted to go beyond the definition of jobs defined merely by a time spent analysis and have included an importance to the job scale with each task statement. Additionally, because the types of equipment and commodities are thought by incumbents to be such a critical variable (in many cases) for selection and training of high level professionals and managers, we have added a separate section to the job description inventory to identify the equipment/commodites that differentiate various jobs at various levels of the organization. This additional section also allows us to greatly shorten the length of the inventory. For instance, rather than a separate task statement such as "develops long range budget for fixed wing aircraft," as well as a similar statement for each type of equipment covered in the career program(s), we merely listed each commodity and equipment type once and shortened the task statement to read "develops long range budget." As can be envisioned, this shortens the inventory from several thousand items to several hundred (349 task statements along with 88 commodities and equipment categories.)

Once the statistical clusters of tasks and duties are established through the various CODAP computer programs, selected SMEs will review the data and establish the task/duty clusters. Following the specification of task and duty clusters, SME panels will be formed for each job cluster to develop KSAs needed for training requirements and selection criteria.

## DISCUSSION

The methodology employed in our study is a multi-method multi-purpose job analysis. The two methods are 1) a task inventory with CODAP analysis and 2) panels of subject-matter-experts (SME). The two purposes are to identify jobs and KSAs for both promotion and training. The task inventory is used to identify and define jobs; the SME panels are used to identify the KSAs for promotion and training.

### Multiple Method

The importance of this project required the best possible job-analysis procedure. A multi-method approach was considered desirable despite its higher cost in terms of time and money. The two procedures were selected because they were felt to represent the best of all those available and because they complement each other. That is, each overcomes a potential weakness in the other.

The task inventory/CODAP method could be used alone. The function of the SME panels, i.e., identification of KSAs, could be accomplished by adding a KSA section to the task inventory booklet. Respondents could rate KSAs on individual 7-point scales relating to promotion and training. This procedure was not considered desirable for three reasons. First, the addition of a KSA section would have increased the inventory length to a point where the completion time for the inventory would be prohibitively long. Civilians cannot be required to complete the task inventory and every effort must be made to

secure their voluntary participation. As the inventory booklet was already quite long (31 pages), any additional length would have a negative effect on the return rate. Second, the identification of KSAs for promotion and training requires more information than can be obtained by a single rating scale on each topic. As is described later in this paper, three rating scales are used just to identify KSAs for promotion purposes. CODAP-type task inventories generally do not employ this number of rating scales. Third, a job analysis consisting solely of task-inventory data is not considered as defensible from a content validity standpoint as is one containing data from multiple methods. Items in a task inventory, like any other questionnaire, are subject to varying interpretations by respondents. No matter how carefully task statements are written, they will be perceived differently (i.e., erroneously) by some respondents. The problem of differential interpretation is especially acute in this project because one task inventory is being administered to employees in three different career fields. The potential for this problem may be greater for rating KSAs than for rating tasks. This is because employees are familiar with task statements in job descriptions and performance standards. They are not equally prepared to deal with KSAs. In general, the content validity of the job analysis will be strongest when the results of the task inventory are corroborated by another method.

The SME panel method also could be used alone. The function of the task inventory/CODAP analysis, i.e., identification and description of jobs, could be accomplished by having the panel members write task statements and rate them on the same scale(s) that would otherwise be used in the task inventory. In essence, this method would be a task inventory without a mailout. The same employees would provide task statements, rate them, and assign them to jobs. This procedure was not considered desirable for three reasons. First, it could have the appearance of a closed-door, "old boy" system to both employees and the courts. Even though it may provide identical results, the large-scale survey method appears more objective than the SME panel method. Second, it does not publicize the job analysis the way a large-scale survey does. The publicity function is important in this project because the promotion and training systems which depend on the job analysis results need to be accepted by employees in order to operate as desired. Employees may not pay much attention to newsletters, memorandums and other information-sharing attempts, but they must attend to a task inventory even if they choose not to complete one. A 100 percent sample is used in the task inventory mailout in part for its public relations value. Although published studies show that reliable results can be obtained with surprisingly small samples, a 100 percent mailout is seen as worth the relatively small extra cost because of the way it involves employees in the job analysis process. Also, the smaller the sample, the more the procedure resembles that of the SME panel methodology and the project desired two quite different procedures. Third and finally, an SME panel approach was not used by itself because SMEs were not comfortable with the task of defining all the jobs in the career programs. Management wanted the job analysis to determine "what is out there," or what is objectively found by the analyst to be present in the field. Management officials and good psychometric practice dictate that a few SMEs are poor candidates to attempt to specify the job tasks of several thousand employees throughout the world.

Qualitative-Quantitative Distinction

The task inventory and panel methods are considered by some to be at opposite ends of a quantitative-qualitative continuum. Our experience with these methods shows this not to be the case. We consider panels to be every bit as quantitative as the task inventory method (conversely, task inventories to be as qualitative as panels). A good deal of statistical analysis follows our panels. The main difference appears to be in terms of sample size. Perhaps the qualitative reputation of panels stems from the use of panels by some job analysts merely to provide narrative information on tasks or KSAs. In our method, panels are used to provide numerical ratings on a number of dimensions. To illustrate this, consider the steps the panel goes through in identifying KSAs for just one purpose, that of promotion:

1. Editing of important tasks. Panel members are provided with a listing of the tasks identified by the CODAP analysis as being most important for their job. ("Important" tasks are separated from all the tasks in the CODAP-produced job description by the use of some heuristic rule. One such rule typically used is that an important task is one that is performed by at least 50 percent of the respondents and is in the upper 50 percent in the cumulative group rating column of the job description.) The panel determines whether performance on those important tasks differentiates superior from average employees. The panel members do this by individually rating the extent to which superior and average employees differ on performance of each task. A three-point scale ranging from "not at all/slightly" to "significantly" is employed. Tasks on which there is little variation in job performance are dropped from further consideration. This step occurs because there is little predictive value in identifying KSAs for tasks that all employees perform equally well. Our ultimate goal is to identify KSAs that result in superior job performance.

2. Identification of KSAs and linkup with individual tasks. Some job analyses stop with the identification of job-related KSAs. Our procedure follows the identification of job-related KSAs with a rating of the KSAs differentiation ability. Panel members individually rate each KSA on the extent to which possession of it is important in distinguishing superior from average employees in performance of each differentiating (see step #1) task. A three-point scale ranging from "not at all/slightly important" to "highly important/critical" is employed. Each differentiating task is rated in turn. Identifying the KSAs that show the largest difference between superior and average employees is done because the most differentiating KSAs should be the most valid ones for selection purposes. For any job, there are many KSAs that are job-related. However, many job-related KSAs show only minor differences between superior and average employees and, therefore, are not very useful for selection purposes. Since there are practical limits on the number of KSAs upon which an applicant can be rated, the use of the most differentiating ones maximizes the validity of the rating process.

3. Linkup of KSAs with overall job. In this step, panel members identify the KSAs required for the job as a whole and determine how well those KSAs differentiate superior from average employees. This step is similar to the last one, the main difference being that the job-related and differentiating KSAs are identified for the job as a whole instead of for individual tasks. The overall-job linkup is used in addition to the task linkup because there is no accepted best way to combine KSAs ratings on individual tasks. Also, there may be KSAs required for superior overall job performance that are not related to individual tasks (e.g., the ability to handle numerous activities concurrently).

Perhaps the quantitative reputation of the task inventory method stems from the amount of computer time required and the quantity of printout produced by CODAP. There is no denying that the task inventory method requires a good deal of statistical analysis. However, there are many qualitative aspects of the task inventory method which are often overlooked or glossed over. Two general areas will be used to demonstrate the qualitative nature of the task inventory method. First, consider the process of determining the sampling plan, i.e., determining which employees to interview when writing task statements. This is the "Catch 22" of the task inventory method because the analyst has to have a pretty good idea of what the results will be before the job analysis is conducted. That is, one needs to have an idea of what is done at each location (major command, office, division, etc.) in order to interview the full range of employees and develop a comprehensive inventory. The interview plan is based primarily on qualitative information.

A second example of the qualitative nature of the task inventory method is the number of decisions necessary in the writing of task statements. The job analyst can alter the results of the CODAP analysis by altering the manner in which task statements are written. By writing more general statements, the analyst can "hide" differences between groups. Conversely, by writing very specific task statements, the analyst can make jobs appear more different. By judicious use of verbs that relate to level of responsibility (e.g., draft vs. write, propose vs. approve), the analyst can control whether grade and staff-operating differences appear in the CODAP analysis.

Thus, both job analysis methods have qualitative and quantitative aspects. The qualitative side of the task inventory method may be more apparent to those who have employed it repeatedly. The quantitative nature of the SME panel method may be hidden by the few analysts who continue to use it to produce only narrative information.

## CONCLUSION

In order to meet current legal and federal guidelines in the implementation of Army programs, a rigorous job analysis is necessary. In the case of this logistics job analysis, we found that an integrated (multi-method, multi-purpose) approach was called for. Our experience is that using this multiple approach will provide the most useful product available.

# Managerial Competencies Assessment for Army Civilians

Dr. Grenville C. King
U.S. Army Management Engineering
Training Activity

In the autumn of 1982, the U.S. Army Civilian Personnel Center tasked the Army Management Engineering Training Activity to develop a competency based system to support manager and executive development (Army Civilian Executive and Manager Development System, ACE&MDS).

Research on the project was initiated in January 1983 with the review and analysis of prior research studies. However, frequently during the course of the project, several major activities were initiated and researched concurrently. The multiple concurrent chronologies, involving the major design, development, and test activities of the ACE&MDS project are grouped as follows by functional category:

1. Isolating the specific managerial competencies which would be used in ACE&MDS.
2. Testing a competency assessment instrument for individual managers.
3. Identifying and evaluating training and development opportunities.
4. Developing an automated data base and processing system.
5. Designing and testing a complete managerial competency assessment program for Army management personnel at all levels and in all career fields.

## The Managerial Competency Concept

The initial tasks that faced the ACE&MDS project team involved isolating the specific managerial competencies that would be used as the basic foundation for the system. An extensive literature search was conducted to review the development of the competency concept and to attempt to identify potentially useful assessment systems. Programs that had been developed within the Federal Government were given particularly careful review since their competencies were potentially more likely to be representative of the Army civilian population than studies of private-sector management. The most significant public-sector competency programs that were reviewed included: U.S. Army Training and Doctrine Command Soft Skills study, U.S. Air Force Training Command, U.S. Army Organizational Effectiveness Center and School, U.S. Office of Personnel Management(OPM) Management Training Needs Profile as modified for U.S. Army Tank Automotive Command (1976), and the OPM proto-type Management Excellence Inventory (1983).

An examination of all the foregoing studies, and numerous others of less significance indicated that there was continued interest in many circles in defining those competencies required to manage successfully. The various studies, however, approached the question of managerial competency

in different ways. Some were elaborations on the studies undertaken by Shartle at Ohio State University in the 1950's and 1960's; while others focused on the Roles of Management (Mintzberg, 1973) or investigated functional groupings of managerial work (Tornrow & Pinto, 1976). The definition of managerial competency was also quite diverse.

Definition of Competency

Early in the ACE&MDS project, the research team developed a standard definition of competency and adopted criteria that had to be met by a competency system and its assessment process for it to be of value supporting DA requirements. A Competency was defined for the purpose of ACE&MDS as, "Aggregates of behaviors and activities expressed in terms of knowledges, skills, and abilities, necessary to perform a task in an acceptable manner." A managerial competency was further defined as, "Aggregates of behaviors and activities, expressed as knowledges, skills, and abilities necessary to perform in a competent manner in the manager role."

A significant criterion of competencies for use in ACE&MDS was that the competencies must be based on observable, teachable, and trainable behaviors. Non-teachable aspects such as physical traits and characteristics were not used.

In addition to the teachable/trainable criterion, it was decided that a competency assessment system for the Army would have to have the following characteristics:

- o Accuracy; identifies true needs.
- o Objectivity; factual and unbiased.
- o "User-friendly"; easy to use; little time required to administer; timely; relevant feedback.
- o Relevancy; results would be job related and also allow information for career growth.
- o Hierarchical; data had to be in a form capable of easy consolidation for policy decision.
- o Cost effective; per capita processing costs had to be low enough to allow large-scale use.

Examination of Existing Managerial Competency Assessment System

When all the factors listed above were taken into consideration, three existing managerial competency assessment systems, two from the Office of Personnel Management, met enough of the criteria to be evaluated as possible systems for the Department of the Army.

After careful evaluation, it was determined that it would involve as much work to modify the existing inventory and test it as it would to build a new inventory. Further work with the OPM inventory was discontinued and the ACE&MDS project team started formulating a plan for a managerial competency assessment process tailored to Army needs.

The Army-specific competency problem which initially appeared as a major obstacle, was solved through the use of the competencies identified for the Automated Program for Executive Development(APED). This earlier AMETA project was performed for the SES Office, HQ, DA, in 1981.

In order to be beneficial for developmental purpose, the specific activities performed by Army executives had to be identified. Since the SES personnel came primarily from a broader management pool, the identification of managerial activities for APED was cascaded through all levels of management and expanded across the Department of Defense. Thus, the APED list not only applies for Army management, but also for Air Force and Navy management. The basis for APED's usefulness lay in its identification of 189 specific activities actually performed by Army (and DOD) managers as presented in the 1981 study, the APED activities were cast as teaching objectives grouped qualitively into 28 categories which were likewise grouped into nine major areas. After carefully examining the APED study and reviewing the study methodology with its authors, it was determined that with some refinement and testing, the study could be used as the initial basis for development of an Army-wide competency system.

Designing a Managerial Competency Assessment Inventory

Initial ideas on the ACE&MDS inventory format and content coalesced rather quickly. It was decided to base the inventory on the 28 competencies and nine Macro-Competency Areas instead of all 189 specific activities (which existed at that time) to avoid problems with length and time that had been observed with the OPM instrument. To offset potential problems with this approach, test subjects (i.e., individuals used to test the inventory instrument) were furnished with a list of all 189 activities to allow them to refer to specific activities when necessary.

Once the initial decisions were made on the number of competencies that would be used, the project team began development of the demographic portion of the inventory. An initial requirement established for ACE&MDS was that competency data would be able to be aggregated in various combinations for different groups of managers. The demographics would provide a mechanism that would allow the research team to not only aggregate data by various groups but also to directly address specific characteristics of the population. As a result, data now exists to address questions of population mobility versus promotion, education, and experience factors in relation to grade, career field, etc. Since these questions can be addressed in multiples for different career fields, installations, level of management, and so forth, the demographics represent a valuable planning tool for personnel management.

While the demographic portion of the inventory was being designed, developmental effort was underway with the competency assessment portion of ACE&MDS. Since actual measurement of competency is so time consuming, it was determined the format of the ACE&MDS inventory would be a self-report based upon the concept that one person who best knows the strengths and weaknesses of that individual is that individual himself(herself). After using this self-assessment concept in tests with approximately 500 managers, it was found to be desirable to involve the immediate supervisor in the process.

To avoid similar problems observed with other inventories, the competency assessment part of the ACE&MDS inventory was segmented into three portions which were rated separately. In Part I, the respondent was asked

to rate the importance of each of the nine macro-competency areas to their job. This was a forced choice rank-ordering process where the person taking the inventory uses whole numbers, 1 to 9, without using the same number twice. In Part II, the respondent indicates for each of the 28 competencies on two factors, a. "Do you need training?" - YES/NO; and b. "How often do you use this competency?" - Never/Rarely/Occasionally/Frequently. Part III asks several questions that were generated from research with managers who were rated exceptional on their last performance appraisal. Each aspect of the competency assessment is independent thus avoiding problems of mixed quesion/answer results. The independence of these elements is maintained throughout the data summaries; however, they can also be considered together if the need arises. For example, the elements on an inventory can be combined to develop an individuals training needs priority in which case they are weighted and multiplied.

Tests with the Proto-Type Inventory

The ACE&MDS inventory underwent eight iterations, each of which were refinements of the previous version. All versions of the inventory shared the basic 3-part characteristic; therefore, the following discussion of tests with the inventory refers to the proto-type inventory as if it were one version.

The inventory tests were initiated in late March of 1983 with ten AMETA management students. Based on interviews with these students, the inventory was revised and tested with subsequent management classes. This process was repeated until, by late June 1983, 373 students had taken the inventory and been interviewed. The interviewing was done both individually and in groups and resulted in improved instructions and better formatting. Additional studies were also conducted, such as one performed with exceptional managers where inventory results were compared with the results of detailed critical incident interviews. These results were calculated manually. The reaction from the test participants was very positive, both to the inventory and to ACE&MDS as a system. A final test of the inventory involved test/retest reliability. This action was conducted to determine if the inventory gave stable results over time. The analysis of test/retest data indicated that the ACE&MDS inventory was highly reliable. With this information, it was decided to test the ACE&MDS inventory across a representative population of Army managers.

The Low-Scale Inventory Test Phase

The low-scale test of the inventory took place between July and September 1983. To insure a representative sample of participants, samples from installations in different geographic areas, from every Major Command(MACOM) and from all organizational levels within the commands, were selected. The result was a test population of 2,400 managers from 42 installations ranging in size from Tooele Army Depot, Tooele, Utah, to Headquarters Training and Doctrine Command(TRADOC), Fort Monroe, Virginia.

By this time, a rudimentary automated processing system was in place and participants were given the option to have training needs printouts returned to them; over 1,400 requested the training needs printouts. A

structured observation checklist was used by researchers to record the actions, questions, or comments of the participants as they participated in that test. As with the proto-type test, interviews were conducted of groups and individuals and information collected to improve the system. A long form inventory was also given at selected sites as a cross-check during this phase.

The outcome of the low-scale inventory test was that the ACE&MDS inventory worked. Modifications were made based on the information received from test participants but these concerned "user-friendliness" issues.

The Three-Step Inventory Process

As mentioned previously, the proto-type inventory relied totally on the self-report of the position incumbent. Analysis of the results of the proto-type test indicated that a small number of people were trying to manipulate results; for example, some reported training needs in everything, others reported no training needs at all. While the position incumbent is usually the best informed about the requirements of that position, a check was required to minimize this tendency of some managers to manipulate the inventory results. In addition to the personal manipulation problem, the study team was concerned about how the supervisor's perceptions could be considered in the process.

Even though the incumbent's perception of what his job demands are is likely to be very accurate, it is the supervisor who established the job requirements, evaluates performance, and initiates training requests to satisfy needs. The supervisor does this from his knowledge and perception of the incumbent's position. Hopefully, the two perceptions would converge; if not, a mechanism to reconcile the two views of job requirements and training need was needed. This mechanism evolved into the AMETA Three-Step Assessment Process.

In the Three-Step Assessment Process, Step 1 required that the subordinate fill out the ACE&MDS inventory based on how he views his job requirements and how he sees his personal need for training and development. In Step 2, the immediate supervisor fills out the inventory based on this assessment of the subordinate's job and on his assessment of the subordinate's need for training and development. Step 2 is completed without discussion between the two parties or to any reference as to what the subordinate wrote. In Step 3, the subordinate and supervisor meet and negotiate their perceptions of the job's requirements and the subordinate's need for training and development. The first two steps are important because they give both individuals the chance to record their views objectively without concern about the other person's perceptions.

Experiments with the Three-Step Process were conducted in August, 1983 at various installations. Matched pairs of actual subordinates and supervisors role-played the 3-step process. Observers watched the process and conducted post-test interviews to determine facility, accuracy, and reliability of the 3-step process. The superior-subordinate interaction that takes place during the Three-Steps also has important fringe benefits. It forces participants to compare perceptions using a standardized list of observable performance-based competencies; this puts that the interaction

on a more objective footing. Respondents in interviews reported that their interactions, while not always comfortable, proved enlightening since participants had previously relied heavily on assumptions regarding each others perceptions.

The Three-Step Process was incorporated into the ACE&MDS inventory. The inventory and 3-Step Process was now ready for a full system operation test.

To insure as accurate of a system test as possible, the system was tested throughout the management population of a contingucus segment of a Major Command(MACOM). Everyone from team leader through executive; from camp/post/station through MACOM, HQ, would participate. Participants received individual training needs printouts. Summary data on organization populations were provided to local T&D officers after the test. Post test interviews were conducted with a sample of participants and will all T&D officers.

LOGAMP

The full ACE&MDS system was also used to support the Logistic and Acquisition Management Program(LOGAMP). The LOGAMP is a proto-type program to improve the managerial and technical competency development of high potential GS-13-15 personnel in a family of six Logistics-related career fields. While the ACE&MDS MACOM test was underway, the LOGAMP program was finalized. LOGAMP selectees were administered the ACE&MDS inventory as a means of developing IDP's. AMETA personnel provided follow-on assistance after the Three-Step Process by assisting the LOGAMP participants and advisors at a special IDP seminar.

The experience of the system operational test in the U.S. Army Corps of Engineers and the use of ACE&MDS to support LOGAMP clearly indicated all components of ACE&MDS were functional as designed. The total integrated system provides timely and accurate managerial competency needs identification for individuals and groups and is an effective mechanism to meet the developmental needs surfaced by the Three-Step Process.

# TYPES AND QUALITY OF NATIONAL TRAINING CENTER DATA

Patrick J. Whitmarsh
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

The National Training Center (NTC) at Fort Irwin, CA provides U.S. Army heavy battalion task forces (TFs), controlling brigade headquarters and supporting units training with Multiple Integrated Laser Engagement System (MILES) and Live Fire under realistic combat conditions not found at home station. The NTC is in a position to provide Army decision makers information on battalion task force training, doctrine, and readiness in the context of the Army Training and Evaluation Program (ARTEP). One objective of the Army Research Institute Presidio of Monterey Field Unit (ARI-POM) NTC Training Research Program is to establish an operational, computer-supported NTC Data Management and Analysis System to support research on these complex issues.

The purpose of this paper is to present the types of NTC information currently available; research problems with respect to collection and merging of information; data quality issues; and a discussion of standardized procedures.

## TYPES OF INFORMATION AVAILABLE

Operations Plan. The Operations Plan is a planning document provided by the NTC to the TF controlling brigade to be executed by the battalion TFs at the NTC. Information typical of the Operations Plan include the MISSIONS which will be executed, and the general Astronomical data across the time period included; a SCENARIO PLAN; a TIME/EVENT SCHEDULE for Mechanized and Armor TFs, and OPFOR; a GENERAL SITUATION narrative describing a hostile international condition with potential for war between the two superpowers' OPFOR and TF with GENERAL SITUATION UPDATE(S); an ANALYSIS OF AREA OPERATIONS narrative describing a General Description of The Area, Military Aspects of The Area, Effects of Characteristics of The Area; a DEFENSE INTELLIGENCE REPORT on the Handbook Krasnovian Army (OPFOR): and a series of OPERATION ORDERS (OPORD), WARNING ORDERS and accompanying overlay maps.

Take Home Package (THP). The THP is prepared for each TF Commander by the NTC Training Analysis and Feedback (TAF) Division. The document contains narrative and numerical descriptions of unit performance at the NTC and recommends additional training for home station training. THP information includes a General overview of the document in terms of Purpose, Scope and Organization; the Missions Conducted; a 14-Day AAR Briefing describing the Offense (missions) and Defense (missions) by Engagement Simulation in terms of Equipment Loss and Equipment Loss Ratio and by Live Fire in terms of Rounds Fired, Target Hits, Rounds Per Target Hits, Targets Hit, Rounds Per Targets Hit, Kills and Rounds Per Kill, and Trends describing TF Performance and Battalion Equivalent Loss Comparison: the Daily AAR Presentations describing in narrative each mission in the context of Operating System, Reported Event, Effect, Reason, Cost and Doctrine and describing numerically the Equipment Losses by Company/Team across Tanks, APC and TOW and the Radio Transmissions by Task Organization across Number of Transmissions, Average Length of Transmissions, Number of Transmissions Greater-Than-25-Seconds-Less-Than-50-

Seconds and Number of Transmissions Equal-To-and-Greater-Than-50-Seconds; and a List Of Audio/Visual Materials.

Company/Team Take Home Package. The Company/Team Take Home Package is prepared by the TAF Division for each TF company/team. The document describes in narrative form, the Plan, Prepare and Execute phases for each NTC mission for each Company/Team.

Video Tape After Action Review. The video tape After Action Review (AAR) is prepared by the TAF Division on each TF mission. It is intended to supplement the THP in establishing future home station training. The AAR, considered the key to training effectiveness, is a tactical discussion among all soldiers conducted by the senior trainer immediately following each mission training exercise.

Unit After Action Report. The Unit After Action Report is a detailed narrrative and numerical description provided by the brigade commander to the division commander on the NTC training period. The Unit After Action Report narrative includes a Narrative Overview, each TF Commander's Comments, Task Organization, Specific Comments and Recommendations on Command and Control, Maneuver, Fire Support, Intelligence, Air Defense, Mobility and Countermobility, Combat Service Support, Survivability, and Recommendation for Improvement of the NTC Experience. The numerical contents include Maintenance Production and Major Assembly Usage, Vehicle Requirements, and Budget.

National Training Center Instrumentation System (NTC-IS) Data Tape. The NTC-IS is divided into the Range Data Measurement Subsystem (RDMS), Range Monitoring and Control Subsystem (RMCS), and Core Instrumentation Subsystem (CIS). The RDMS provides real-time position/location and engagement event data on all instrumented players during the TES. The RMCS monitors and controls all activities on the engagement simulation and live fire ranges. The CIS provides all real-time data processing and interactive play capabilities needed to monitor, command and control all engagement simulations. The NTC-IS Data Tape contains information replay of the battle for display on "NTC workstation" graphics terminals and also for input into the ARI-POM NTC Research and Training Data Base. The tape data are formatted for the INGRES relational data base management system (RDBMS) on the DEC-VAX 11/780 computer system. INGRES was selected based on the research data base (RDB) criteria: maximize data element relationality, user friendly, retrieval efficient, modularly expandable, allow researchers continuing study, reliably supported and meet computer system specifications. INGRES facilitates research by manipulating the existing 61 data tables. Continuing research anticipates restructuring, adding and deleting variables with INGRES.

RESEARCH PROBLEMS

Collection. Data collected from the battlefield must be interpreted from a given context. To get a complete picture for off-line analysis we need plans, scenarios, orders, communications, and accurate position location and firing data. In addition, we would like to have data on transient events such as smoke and on factors which are basically not instrumentable (e.g., what did the commander know?). This is an ideal not readily obtainable. Our goal is to work with NTC personnel to obtain as much as we can and understand what we are missing. This will permit us to better interpret performance.

Merging. Certainly one major issue is integration of information from multiple sources to permit more complete and effective analysis and interpretation. The ARI-POM NTC Research and Training Data Base is designed to accommodate text as well as numerical data thereby providing an efficient RDB.

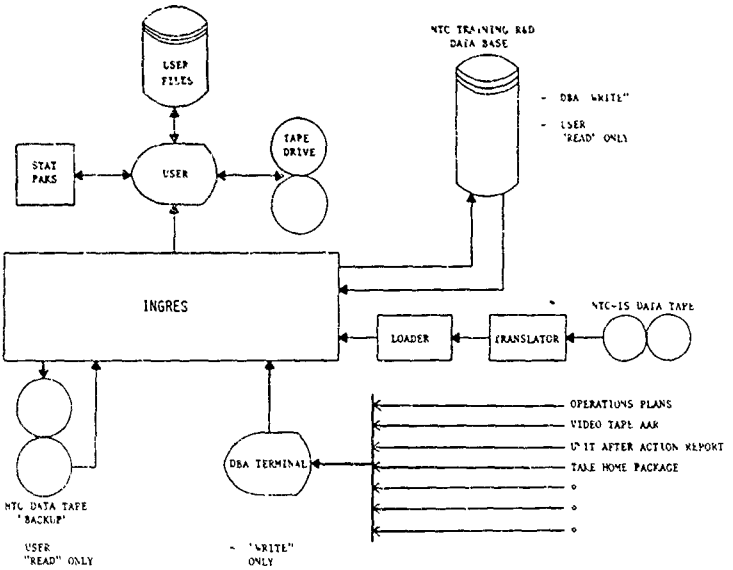



Figure 1
ARI-POM NTC Research and Training Data Base

Figure 1 indicates Operations Plan, THP, Company/Team Take Home Package, Video Tape After Action Review (convert audio to text) and Unit After Action Report are entered to the NTC Training and Research Data Base with INGRES by Data Base Administrator (DBA) from a terminal with "write only" privilege. The NTC-IS Data Tape is entered by tape drive thru NTC Translator and Loader software and INGRES by computer system operator. To simplify subsequent NTC-IS Data Tape operations, a NTC Data Tape "backup" is performed whereby a user can access the data tape thru INGRES with "read only" privilege. The user has access to the tape drive, statistical packages and user file system, and the RDB thru INGRES with "read only" priviledge.

DATA QUALITY ISSUES

Accuracy. Previous investigations have attempted to utilize the NTC data sources without knowledge of the conditions under which the data were collected, the result of which has been to limit the research to one data

source. With multiple sources and types of data and efficient merging, comparisons will be possible to identify data reliability. As a first step we investigated the accuracy of data entry on vehicle loss and radio communication variables across THP, INGRES RDB, Graphics Displays and Summary Statistics derived from the NTC-IS Data Tape for seven selected TF missions. The results indicated perfect agreement, in all cases, between the Summary Statistics and the INGRES RDB only. Some of the problems are attributed to missing data entries by operator personnel and possible software errors.

## STANDARDIZED PROCEDURES

Operational Control Units. Data quality can be improved by use of standardized procedures, at the same time recognizing data quality can never be perfect because some data are collected during real time. However, procedures can be designed to support both training at NTC and improve data utility collected for off-line analysis. We, at this time, are preparing a procedural document to assist the NTC CIS Operations.

## CONCLUSION

The NTC represents a complex environment providing the Army, for the first time, an opportunity to evaluate training readiness of TF units under simulated combat conditions. This paper examined types of information currently available, identified some research problems, discussed data quality issues and standardized procedures for the purpose of supporting the NTC RDB. A fully supported NTC RDB is the key toward effectively analyzing unit performance, thereby providing for an efficient and effective fighting force.

## REFERENCES


Banks, J. H. (1985, May). Some issues and concepts for unit training and evaluation at the National Training Center and at home station. Paper on evaluation of collective training for technical panel UTP-2 of The Technical Cooperation Program (TTCP). Alexandria, VA: U. S. Army Research Institute.

Kroger, A. (1983, April). Considerations in the establishment and use of a research data base for analysis at Army training (Jet Propulsion Laboratory Working Paper). Pasadena, CA: Jet Propulsion Laboratory.

Nichols, J. J. (in preparation). The DeAnza Primer (BDM Research Product). Monterey, CA The BDM Corporation.

Science Applications Inc. (1983, June). EMC/TAF operating manual for the NTC Core Instrumentation Subsystem (CIS) (500 player system) (Subcontract No. SD-6200). La Jolla, CA. Science Applications Inc.

Science Applications Inc. (1984, June). INGRES table formats and descriptions for a prototype National Training Center (NTC) research data base system. La Jolla. CA: Science Applications Inc.

Whitmarsh, P. J., & hamza, A. N. (in preparation). National Training Center data reliability A comparison of multiple data sources (ARI Technical Report). Alexandria, VA: U. S. Army Research Institute.

# An Overview of ARI's Research Program on the National Training Center

James H. Banks
U. S. Army Research Institute Field Unit
Presidio of Monterey, California

The National Training Center (NTC) was established at Fort Irwin, CA to train battalion task forces under highly realistic and intense conditions which are not obtainable at home station. A second purpose was to provide information to help the Army evaluate its training, doctrine, organization, equipment, and readiness. Thus, the NTC provides capstone training and potentially permits measurement of the output of the Army unit training system. However, combined arms training exercises of the scale, complexity, and realism of those at the NTC have never before been conducted, nor has information of this richness been available for analysis and interpretation. ARI's NTC-based research program was established to help get the maximum benefits from the NTC and to increase the effectiveness of ARI's overall R&D program. I am going to provide an overview of the program, outline a few of the problems in measurement and interpretation of unit performance, and in diagnosing and correcting problems when they are detected, and generally describe what we hope to do. Other speakers will describe the NTC, the types of data available and potentially available, and some early analyses.

## Measurement and Interpretation of Task Force Performance

The NTC provides an unparalleled opportunity to objectively measure and analyze unit performance to detect strengths and weaknesses, typical performance, performance ceilings, and trends. Measurement and interpretation must take into account both the nature of the Task Force and the combat environment in which it must function.

The Battalion Task Force is a complex system containing maneuver, intelligence, fire support, air defense, mobility/countermobility, and combat service support elements or subsystems, all bound together by a command and control subsystem, as shown in Figure 1.

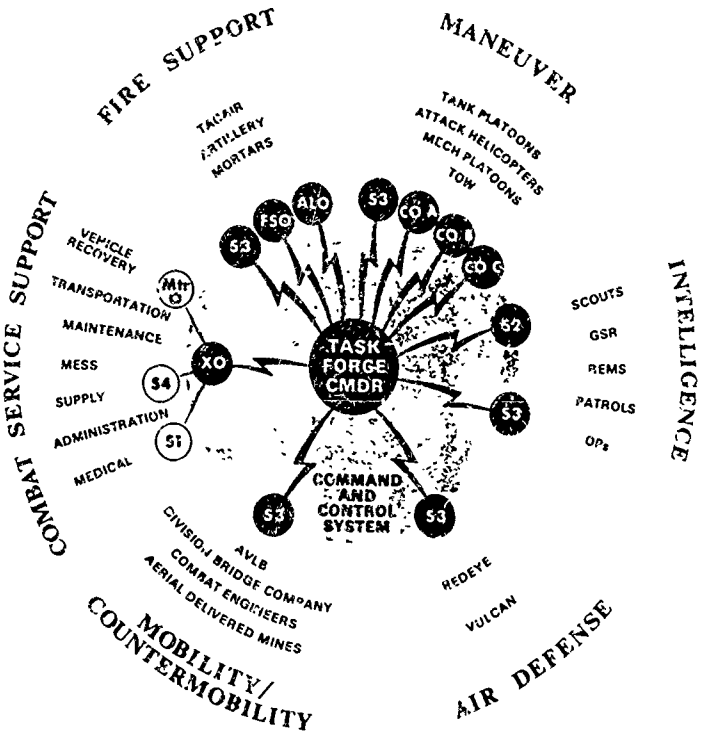



Figure 1
One Model of the Battalion Task Force as a System

The combat environment for the Task Force--or the NTC as a high-quality simulation--is complex, rapidly changing, and uncertain. "Correct performance" always involves trade-offs of capabilities, opportunities, and risks, and always in the context of the commander's orders and intent for a mission. Moreover, friendly actions are opposed and countered by an intelligent adversary. Thus the problems faced in combat, by the leader and by the soldier, are typically "fuzzy"--not clearly stated, where the needed information is not all available, where no set procedure can be used to reliably produce an answer, and where there may be no single answer that is demonstrably correct.

Given the complex nature of the Task Force and the combat environment and requirements, the problems in measuring and interpreting unit performance at the NTC are twofold. First, facts do not "speak for themselves": rather, they are meaningful only in relation to other facts and expectations. These relationships can be informal (e.g., "common sense") or quite elaborate, formal, and explicit, as in scientific theories. For the major relevant problem areas--how to fight, how to evaluate collective training and performance, how humans learn, how to instruct teams--the theories are not well developed or integrated. We,therefore, need descriptive models of the Task Force system and subsystems to guide observation and interpretation. Initially these can be derived from doctrine but they need expansion, synthesis, and empirical validation. Second, enormous practical problems exist in operationally defining concepts in terms of procedures and data elements, from the NTC or elsewhere, and actually conducting the desired analyses. While "quality control" of NTC instrumented data can certainly be improved, it will never be perfect because it is collected during real training exercises. In addition, many aspects of performance which might be desirable to observe are not readily collectable by automated methods, and capabilities for non-automated collection are limited and hard to reliably and validly implement. Under these conditions, analyses must be highly robust i.e., permit interpretation despite data gaps and complex interactions. Traditional data analysis methods are not well suited for such applications. However, human beings do, routinely, process and use information of this type. Therefore, we will be interested in developing "expert" approaches and models to supplement conventional methods.

## The NTC as Part of the Unit Performance System

The NTC is not a stand-alone activity but, rather, is part of the Army training system. Interpretation and correction of performance observed at the NTC requires not only measurement of performance but also understanding of the effects and interactions of inputs to unit performance.

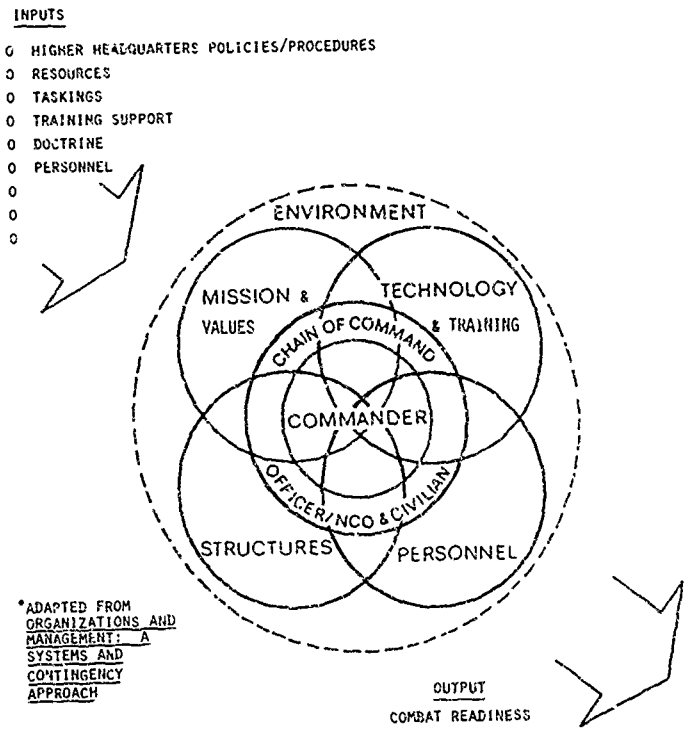



Figure 2
A Model of Unit Performance

This model views the unit as an open system composed of subsystems--missions and values, technology and training, personnel, structure, leadership. Thus, factors internal to the unit--training validity, timing, completeness, and mastery level; leadership priorities; individual and organizational values; morale; cohesiveness; etc.--all are in active and continual interaction with each other and with the environment in which the unit operates. Inputs include those found in Table 1.

Table 1
Inputs to Units

| | |
|---|---|
| Doctrine | Threat analysis, strategy, missions, tactics, TO&E. |
| Institutional Training | Validity, timing, completeness, mastery level, training evaluation, performance accountability, values, quality of exported training support products. |
| Unit Training | Training resources, training evaluation, performance accountability, quality of aids and literature, fill level/job assignments/MOS mismatch, force modernization/transition, housing and post services, management skill, values, leadership. |
| Personnel | Recruiting, selection, classification, job design, aptitude measurement, school assignment, unit assignment, promotion, retention, elimination, retirement, pay and benefits. |
| Equipment | Operability, maintainability, durability/reliability, trainability, performance capability, cost. |
| Logistics | Speed and range, quantity/lift, survivability, economy/consumption, cost. |
| Societal Values | Attitudes towards the military, national goals and pride, etc. |

To obtain maximum benefits from the NTC, it must be treated as a part of the unit performance system. Therefore, interpretation and improvement of performance must consider both home station factors and outside inputs to the unit. ARI's R&D concept takes this into account.

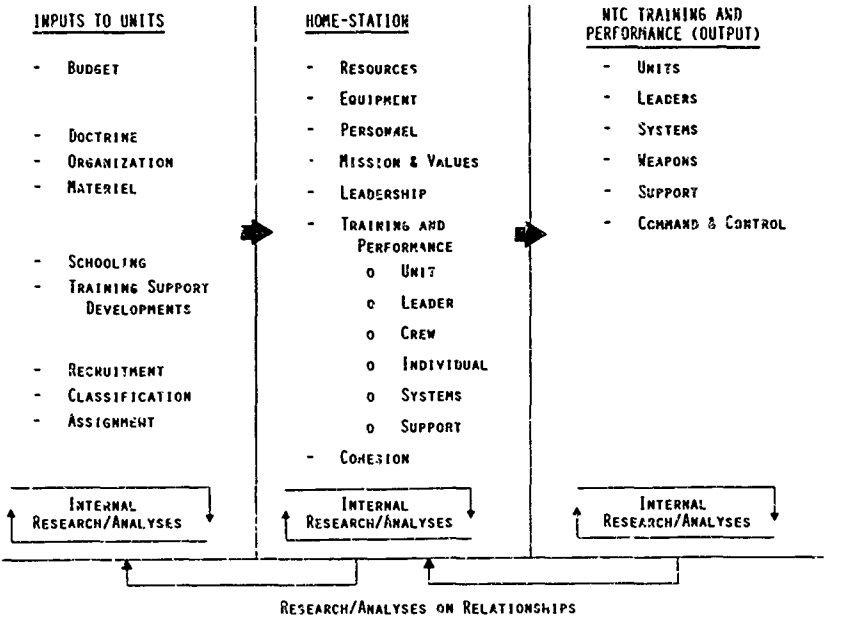



Figure 3
Research and Development Concept for Unit Performance System

In this concept, research on NTC training and performance, on home station factors, and on the various inputs to units will be conducted because of their importance in their own right to Army sponsors and clients of the ARI program. In addition, ARI has established the Unit Performance R&D Center at its Presidio of Monterey Field Unit, with scientific and Army training expertise and computer analytic and data base capabilities for handling data from NTC and other sources. As the data base is established over time, it will permit development of a scientific and practical approach to the measurement and interpretation of unit performance and, for the first time, effective research on the relationships and interactions of factors in the unit performance system. The primary sponsor for the NTC-based program is the Combined Arms Training Activity which is the Army executive agent for use of NTC information for assessment and improvement of Army training, training support, doctrine, organization, and materiel. It is expected that research at the Center will also potentiate the effectiveness and value of other ARI research and development.

# LEADER PERFORMANCE CRITERIA AT THE NATIONAL TRAINING CENTER (NTC)

Earl C. Pence
U.S. Army Research Institute

## Introduction

The presentations you have heard thus far have described the operation of the NTC and the types of data which are currently generated during the training exercises. For the most part, these data describe WHAT happened during the preparation and execution of a mission and are performance data at the unit level (battalion, company or platoon). The focus of the current presentation is on the development of a new data collection system which will provide information to aid in understanding WHY certain events or outcomes occur at the NTC. The new data collection system is being developed to aid the NTC Observer/Controllers (OCs) in observing and recording performance indicators of leaders at the battalion, company, and platoon levels.

## The Current System

The OCs at NTC have, of course, been observing the actions of leaders at NTC since the time the training center first opened. In fact, the After Action Review (AAR) process used to provide feedback after the end of missions at NTC is a leader-oriented process in that the information at the battalion and company levels is provided to the key leaders in the battalion. The information itself, however, is focused on unit performance with little emphasis on leader performance except for probing questions as to why leaders made certain decisions or took particular actions which the OCs judged as critical to the outcome of the mission.

In August of 1984, the Leadership and Management Technical Area of ARI, in conjunction with the Center for Army Leadership (CAL) at Ft. Leavenworth, began work on a leadership research effort at the NTC. The initial task was to assess existing leadership training and data collection processes at NTC and to determine the potential for enhancing the quantity and quality of leadership data collected at NTC as well as the means for improving the development of leaders participating in exercises at NTC. During the fall of 1984 a research team composed of a researcher from ARI and a Major in the research branch of CAL spent a two week rotation in the field at NTC. The team traveled with the OCs and focused their observations on the manner in which the OCs gathered information, recorded observations, and provided feedback to leaders training at NTC. The research team also recorded the conditions under which the OCs performed their work.

While I will describe the conclusions drawn from these observations verbally, I think a few slides illustrating the conditions under which the OCs perform their job will provide a better understanding than any verbal description I could give.

The first three slides illustrate the NTC OC's primary work site--the 1/4 ton jeep. The jeeps are loaded with 7 to 9 days' supplies and all the OC's tools of the trade, including a lot of pyrotechnics and a map strapped to the hood. The fourth slide illustrates an example of an OC observing the deliv-

---

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.

ery of a company operations order. The fifth through seventh slides give you an idea of what the OCs observe while chasing tanks and APCs (Armored Personnel Carriers) during a training mission. The eighth slide illustrates another part of the OC's job, crawling up on tanks to verify MILES kill codes. The ninth and tenth slides illustrate the platoon OCs meetings with the company OC at end-of-mission to compare notes and prepare the company level AAR and the delivery of the company AAR. The eleventh and twelfth slides illustrate the preparation process for the battalion AAR and the cite for the battalion AAR. The final two slides illustrate the typical sleeping conditions for the OCs when they have the opportunity to obtain more than an hour or two of sleep. If they have only a short time (which is typical) they will simply sleep sitting in the jeep.

Initial Research Findings

The key findings from the research team's initial visit to the NTC are listed on the first overhead slide.

---

SLIDE 1

INITIAL RESEARCH FINDINGS

The observer/controllers receive little or no training.

Considerable variation between OCs on decision rules guiding observations and feedback.

OCs made notes of their observations in a variety of means:

* 3 x 5 note cards
* Small memo pads
* In grease pencil on map cover on hood of jeep
* In grease pencil on top of ammo can

Notes used primarily for AARs and relatively little of the information was captured permanently.

---

A second round of observations in the field at NTC by the same research team in June of 1985, followed by interviews with approximately 30 past and current NTC OCs, confirmed all of the findings noted above. The only major new finding from the second data collection effort was that the company OCs are now required to produce a company level Take-Home Package which makes it even more critical for them to record their observations during or immediately after each mission.

The observations in the field and the interviews with the OCs indicated the need and opportunity for enhancing both leader performance data collection and leader development at NTC. The primary means for accomplishing both goals will be to develop a system which encourages more systematic, consistent, and accurate observation and recording of leader performance followed by feedback in the AAR on key leader events impacting on unit performance. The system must meet the criteria contained on the next overhead slide to be acceptable to the OCs who would use it.

SLIDE 2

CRITERIA FOR SYSTEM ACCEPTABILITY

1. The system must not increase OC work load.

2. The system must be convenient to use under extremely harsh environmental conditions.

3. The system must provide information which can be used almost immediately for preparing and delivering AARs.

The Design of the System

There are actually two distinct challenges to be met in designing the data collection system for leader performance criteria at NTC. The first is to develop the appropriate content for the data collection system (i.e., identify performance dimensions and operational measures of the performance dimensions). The second challenge will be to design the format or technology which provides a data collection process acceptable and useful to the NTC OCs.

Content

Development of the actual performance dimensions will be accomplished using a modified BARS (Behaviorally Anchored Rating Scale) development procedure. The key steps in this process are listed on the next slide.

SLIDE 3

STEPS IN DEVELOPMENT OF CONTENT FOR LEADER PERFORMANCE MEASURES

1. OC interviews to identify potential performance dimensions and examples of effective and ineffective leader performance.

2. Development of performance dimensions with definitions.

3. Work with OCs to develop behavioral anchors, performance indicators and decision rules.

4. Formative evaluation of anchors and performance indicators.

5. Development of OC training program.

6. Validation of the leader performance measures using unit performance measures as criteria.

The performance indicators and decision rules referred to in steps 3 and 4 will be guidelines to be used by the OCs in determining the appropriate rating to be given on any particular performance dimensions. For example, on

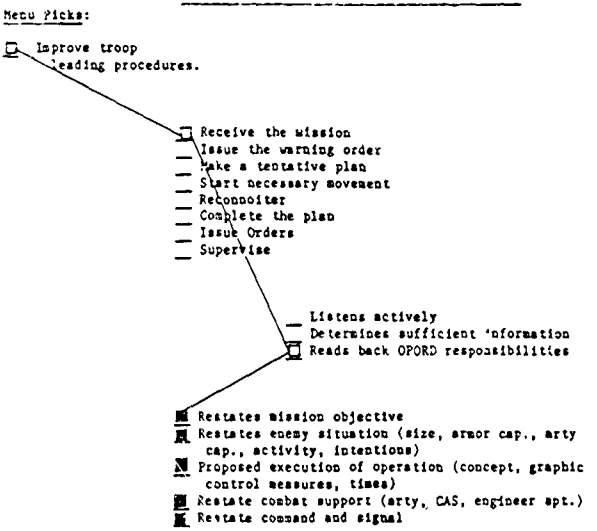
a performance dimension related to communication by the leader, the performance indicators and decision rules would probably include a series of questions which the OC could ask subordinate leaders and soldiers to determine the extent to which mission-related information has been communicated. The leader's rating on the communication dimensions would depend, in part, on the answers the OCs received when they asked the questions. The use of such performance indicators and decision rules combined with guidelines on recording specific leader behaviors would be part of the OC training program.

Format or Technology

The most difficult challenge for the data collection system is to package the system using a format or technology which meets the criteria regarding no increase in OC work load, convenience of use, and immediate information availability. The most promising concept for meeting these needs involves using an electronic clipboard under development by ARI's Training Research Laboratory. The clipboard is a small, hand-carried device which makes use of a touch sensitive screen for recording data. The format of the screen is programmable and seems ideal for a menu-driven data recording system. The menus operating on the screen could provide branching which would enable fairly detailed recording of observations or ratings with only two to four touches to the screen. The system would require no key punching and would be considerably more efficient and convenient to use than traditional check lists. The figure on the screen now illustrates an example of how such a menu-driven data collection system might work.

SLIDE 4

MENU-DRIVEN DATA RECORDING SYSTEM



The output of data or information from the clipboard could take any of a number of formats. The data could be presented on the clipboard screen, it could be printed, or it could be transferred directly to a computer. This variety of options would meet the need for immediate feedback of information in the field as well as eliminating the need for handling paper documents in the transfer of data for research purposes. Specific observations recorded on pre-printed, color-coded 3x5 note cards could be used to supplement the data collected on the automated system for illustrative purposes to explain ratings and provide feedback during the AAR.

## Progress on System Development

### Content

Interviews with the NTC OCs were conducted in June of 1985 and the initial content analysis of the interview transcripts has been completed. Examples of the types of leader performance information obtained in the interviews are illustrated on the next slide.

---

SLIDE 5

CHARACTERISTICS OF EFFECTIVE LEADERS AT NTC

PLANNING AND UTILIZATION OF TIME

Prepares detailed plans and gives specific guidance on what he wants to happen. (Communicates plans well, gives complete and clear operations orders.)

Conducts planning such that his subordinates have time to prepare for the next mission (manages time well).

Uses time effectively and sets work priorities.

COMMUNICATION OF INTENT

Communicates his intent as a commander clearly.

Disseminates information to the lowest level possible and keeps people informed.

Understands his commander's intent.

SUPERVISION AND STANDARDS

Knows what he wants to see on the ground (knows the standard).

Communicates what he wants with authority (communicates standards).

Enforces standards and holds individuals accountable for doing their jobs (makes rapid corrections).

Supervises tasks after giving order (i.e., checks progress, walks the line, inspects positions or ensures this is done).

---

This information will be combined with previous research findings to develop an initial set of approximately 12 leader performance dimensions which will be taken to NTC for modification and operationalization by the OCs.

Format

The initial development of the electronic clipboard is nearing completion and a number of the devices will be available for pilot-testing by ARI during the current fiscal year. The initial pilot testing of the clipboards will occur at a site other than NTC. Preliminary work on a format that may serve as the basis for designing the menu-driven system to be programmed into the clipboard is currently underway.

TITLE Battalion Performance on the Live Fire Range at the National Training Center (NTC)

AUTHORS Thomas K Forsythe and William J Doherty

CORPORATE AFFILIATION The BDM Corporation

## Introduction

The training benefits derived by units training at the National Training Center (NTC) have been of considerable interest to the U S Army The Army Research Institute (ARI) has developed and initiated a programmatic research effort to assess these benefits As a preliminary step in this effort, an exploratory data analysis of the performances of battalion task forces on the live-fire range at the NTC was conducted As stated by Tukey (1977) the purpose of exploratory data analysis is to be detective in nature not confirmatory. Thus, the investigation reflects incursions into the data designed to explicate the structure of the data rather than to confirm a particular model of the data Using the data provided in 54 Take Home Packages for the period of early 1982 to late 1984, it was possible to examine battalion performance for three live-fire scenarios Defend from a Battle Position (Day), Defend from a Battle Position (Night), and Movement to Contact (Day) The presentation of these data was organized around three primary issues

(1) Has battalion performance changed over time at the NTC?
(2) How do the performances of the Armor and Mechanized Infantry Task Forces differ?
(3) What factors seem to be related to performance at the NTC?

## Data Source

The live-fire data were extracted from fifty-four (54) Take Home Packages (THP) These data represented the live-fire performances of 96% of the battalion task forces that trained at the NTC from early 1982 to late 1984 Data from the remaining 4% (2 battalions) of the task forces were either not available, or erroneous due to target equipment malfunctions on the live-fire range It was felt that exclusion of this small amount of data from the investigation would not adversely affect its generalizability or statistical power

The live-fire results reported in the THP generally included the following data

- Number of Targets
- Percent of Targets Killed
- Tank Rounds Fired
- Tank Round Hits and Kills
- TOW/Dragon/Viper Laser Firings
- TOW/Dragon/Viper Laser Hits and Kills
- Number of Each Friendly Weapon System

These data presented ample opportunities for conducting preliminary research on battalion task force performances at the NTC. The performance data was first extracted from the Take Home Packages and a computer data base was established to assist in sorting and analyzing the data. In all cases, unit designations were omitted to preserve unit anonymity.

## Data Analysis

As indicated earlier, an exploratory data analysis (Tukey, 1977) approach was employed in this study. This resulted in a series of analyses being conducted and all aimed at understanding the structure of the data from the live-fire range at the NTC. Specifically, analyses were aimed at satisfying the first two objectives of the exploratory data analysis approach (Hartwig and Dearing, 1979).

(1) Understand each variable as a separate entity
(2) Understand pairs of variables as relationships

A number of data analytic techniques were applied to the live-fire data base. The initial efforts used univariate descriptive statistical techniques. The results of these were used to generate a picture of performance at the NTC. Generally, the results were transferred into graphic display. Somewhat more sophisticated techniques including T-tests and regression were used to examine bivariate relationships between different factors and performance at the NTC.

An important decision influencing the results contained in this report was the selection of a primary variable as measure of task force performance. Of the available variables, the one which seemed to most directly reflect overall unit performance was the PERCENTAGE OF ENEMY TARGETS KILLED.

It was also decided that because this report was designed to provide some early insights into unit performance changes at the NTC, only individual battalion performances would be considered and analyzed. That is, their performances would not be aggregated and analyzed at the brigade and division levels. However, an analysis of that order would be a logical follow-on to this analysis once a better understanding of the performance data at the battalion level is developed. For the same reason, only the effects of the tank ballistics are considered in this study although the performance of the TOWs/Dragon/Viper laser firings on the live-fire range might also be investigated for a subsequent study. (It must be noted, however, that the TOW/DRAGON/VIPER performance analysis would not be as meaningful as the tank performance analysis due to the narrow range in values for the laser firings as compared to the tank ballistics data.)

## Results

Using the "percentage of target kills" data as an indicator of meaningful unit performance on the live-fire range at the NTC and comparing the battalion performances in the first 18 months with those of the next 12 months of a 2-1/2 year period (early 1982 to late 1984), this study found that

(1) The percent of targets killed by the tanks of the Armor and Mechanized Infantry Task Forces increased between the first and second periods of this study

(2) The increase in percent of targets killed between the first and second periods was attributable to both one-time visitors to the NTC and to repeat visitors to the NTC. Therefore, the differences in performance was a function of some phenomena associated with time and not just some advantage factor acquired through repeat visits to the NTC as might be expected

(3) The change in performances on the live-fire ranges was probably not a function of the activities involved in the operation of the live-fire range by NTC cadre, i.e., the live-fire exercises were conducted in a uniformly consistent manner throughout the 2-1/2 year period of this study

(4) The change in percent of target-kills over time was not related to gunnery accuracy as this did not change over time. However, a significant increase in the volume of tank rounds fired by both the Armor and Mechanized Infantry Task Forces in the second period was likely related to the observed increase in target kills

(5) The increase in rounds fired was related to an increase in number of tanks assigned to the task forces (particularly to the mechanized units) in the second period and to an increase in the number of rounds fired per tank

(6) A positive and statistically significant relationship was found between the number of rounds fired per tank and the percent of enemy killed. This relationship was strongest for the day attack and day defend missions

## References

Hartwig, F. with Dearing, B. E. *Exploratory Data Analysis*. Beverly Hills, CA: Sage, Inc. 1979

Tukey, J. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, Inc. 1977

TITLE Analysis of NTC Force-on-Force Performance
AUTHOR Judith J Nichols and William J Doherty
AFFILIATION The BDM Corporation

## Introduction

U S Army battalions have been training at the National Training Center (NTC) at Ft Irwin, California since 1981 The objective of NTC training is to provide a facility where units can undergo essential combined arms training that cannot be accomplished at home stations due to physical limitations and the prohibitive cost of providing realistic training environment A secondary objective of NTC training is to gather information which can be used to contribute to the improvement of doctrine, tactics, training systems, equipment and procedures in the U S Army

In support of the NTC information-gathering objective, the Army Research Institute (ARI) has developed a research program which includes as a technical objective the collection, organization, and analysis of NTC data in order to assess those training benefits which may be accrued by the units training at the NTC This report is an investigation of the force-on-force data that are contained in the Take Home Packages (THPs) which are compiled and issued to the training units at the end of each NTC rotation The focus of this study was the relationship between task force performance on the NTC battlefield and the force modernization (i e , the changeover from M- to A-MTOE Series organization) which took place during the period under study

## Background

The NTC concept is geared to the training of the battalion task force (TF) Each battalion task force participates in approximately six force-on-force exercises during a two-week rotation period at the NTC These exercises usually are more or less evenly divided between offensive and defensive operations using laser-based engagement simulation instrumentation to provide real-time casualty assessment The simulator, the Multiple Integrated Laser Engagement System (MILES), is used on all principal weapons and casualties are assessed when a weapon fires and the MILES laser hits a target In addition to force-on-force training, units also perform three missions on the live fire range during their rotation The results of the investigation of live fire performance are presented in a separate paper for this symposium

The scenario dictates the force ratios of the combatants While terrain and scenario options are limited, no two scenarios are exactly the same When the TF conducts defensive missions

they are always attacked by an OPFOR that replicates a Motorized Rifle Regiment When task forces conduct offensive operations, they originally encountered a defending Motorized Rifle Company However, in the summer 1984, the force ratio was changed to deploy a defending Motorized Rifle Battalion (-) Analysis of the effect of this change resulted in no significant differences, so the data was pooled across time periods

## Scope

### Sample

The sixty-four (64) battalions which underwent training at the NTC during the period February 1982 through January 1985 are represented in this analysis The 64 battalions represent the rotations from six divisons and two separate brigades located in the Continental United States (CONUS) The 64 battalions included 32 armor and 32 mechanized infantry units which were cross reinforced to form 64 combined arms task forces (i e , 32 armor heavy and 32 mech heavy TFs) Two rotations by the opposing force (OPFOR) -- i e , two MTOE battalions permanently stationed at Ft Irwin -- were also included in the sample The OPFOR battalions did not undergo the standard rotation series of exercises but rather performed a mini-ARTEP However, the data extracted for the two missions performed by the OPFOR battalions (i e , movement to contact and defend in sector) were not substantially different from the data of the rest of the sample and so were included in the analysis

### Data Sources

As previously mentioned, Take Home Packages (THPs) are prepared and issued to the training units at the termination of each NTC rotation Separate packages are prepared for the respective armor and mechanized infantry task forces The THP is an overall description of unit performance during the rotation and includes statements of performance trends during the 14-day rotation period The THP is the final compilation of all after action review scripts and encompasses an assessment of all seven operating systems, live fire gunnery data, and TF and opposing force (OPFOR) aggregate losses

The force-on-force results are reported in the THPs in several formats and include the following data

- TF and OPFOR vehicle loss summaries for Offensive Engagement Simulation (OES) operations

- TF and OPFOR vehicle loss summaries for Defensive Engagement Simulation (DES) operations

- Numbers of TF vehicles started and killed for each mission

- Numbers of OPFOR vehicles killed for each mission

Offensive and defensive operations summary data were extracted from the 64 THPs. Mission-specific data for the six most commonly performed missions (i.e., movement to contact/meeting engagement, deliberate day attack, deliberate night attack, defend in sector, delay in sector, and defend from a battle position) were also extracted from the THPs and combined with the summary data to construct the data base for this analysis. Each individual task force was coded to preserve anonymity and the data were then subjected to a number of statistical operations.

## Data Analysis

The task force was designated as the unit of measurement and the data were divided into two groups in order to identify differences in task force performance at the NTC during the period that force modernization was taking place. The first group included all H-MTOE Series rotations occurring between February 1982 and January 1985 and the second group included all J-MTOE Series rotations occurring during the same time period. The two groups were further refined to identify differences between mechanized infantry and armor TFs within the H- and J-Series categories.

The limited scope of force-on-force data available in the THPs severely constrained the analysis in several areas. For example, information on the numbers of OPFOR tanks and APCs starting each mission was not available in the THP thus prohibiting the derivation of percent of enemy vehicles killed during an exercise. Although the force strengths for Motorized Rifle Regiment and Motorized Rifle Company are available in FM 72-1, The Tank and Mechanized Infantry Battalion Task Force, Appendix H, this does not allow for instances of vehicle breakdown in the field, numbers of vehicles actually available to the OPFOR during specific individual missions, and so forth. For the purposes of this report, it was decided that the assumption of consistent standards for OPFOR strengths would not be made.

Two measures of battalion task force performance were selected for study. The first, "%Lost", represents the percentage of TF vehicles lost on the NTC battlefield -- i.e., numbers of TF vehicles killed at each mission / numbers of TF vehicles started

at each mission This measure was available for each TF vehicle type killed (i e . APCs to include TOWs, and tanks) as well as an aggregate value ("Total Vehicles Lost") that combined all vehicle types Results of preliminary analyses revealed that TOW losses were not obviously different from APC losses and so TOW data were incorporated into the APC data to produce an APC/TOW category

The second dependent measure, "Casualty Exchange Ratio", consists of the ratio of vehicle casualties for the two forces -- i e . numbers of TF vehicles killed / number of OPFOR vehicles killed -- and was calculated for both specific vehicle types (i e . APC/TOW and tanks) and across vehicle types

## Results

Analysis of the percentage of task force lost indicated little difference between H-series and J-series The only statistically significant difference occured for the percentage of APC/TOWs lost on the Offensive missions The direction of this difference indicated that J-series task forces lost a lower percentage of APC/TOWs on the average than did H-series task forces (13% fewer APC/TOWs were lost by J-Series task forces -- significant at the p< 01 level) The difference was found for both Armor and Mechanized Infantry task forces, though the size of the difference was greater for the Mechanized Infantry (18% fewer APC/TOWs lost by J-Series Mechanized Infantry task forces compared to 14% fewer APC/TOWs lost by J-Series Armor task forces signi-ficant at the p< 01 level) No statistically significant differences occured in the contrasts between H-series and J-series on the Defensive mission summaries

The lack of a clear trend in performance differences between H and J-series units precludes any inferences being drawn concerning the force-multiplier effects of reorganization under the J-series MTOE

Analysis of the casualty exchange ratios again produced significant differences only for the Offensive missions The contrast between H-series and J-series showed no statistically significant differences for either Armor or Mechanized Infantry task forces This was true regardless of mission type However, there was a statistically significant difference between the two types of task forces in the exchange ratio for APC/TOWs on Offensive missions The mechanized infantry task forces under H-series had consistantly greater casualty exchange ratios than the Armor task forces (H-Series mechanized infantry task forces averaged two (2) more combat vehicle losses per kill in offensive operations than did H-Series armor units -- significant at p< 05) Analysis of the casualty exchange ratios also showed considerable

variability in performance across the six mission types examined in the study, with the greatest fluctuations found in the Movement to Contact mission

Based on the casualty exchange ratios, the OPFOR out performs the battalion task forces by achieving a much higher kill ratio of tanks and APC/TOWs when on both the offense and defense as compared to the TFs when they are on the offense or defense

## Discussion

The results of the analyses indicate that the determinants of force-on-force performance are more complex than the current information from Take Home Packages allow. Considerable variation in battalion task force performance was found in the results reported above Yet, only a small portion of that variation seemed to be related to diffrences in the organization of task forces (H-Series vs J-Series) or to the type of task force (armor vs mechanized infantry) Further, the differences in performance along these dimensions were not consistent either in terms of type of mission (offensive or defensive) or type of performance measure (percentage lost or casualty exchange ratio) This lack of systematic results and the relatively small size of the relationships provide considerable evidence that force-on-force performance can only be understood in the full context in which that performance took place To accomplish that end, it will be necessary to employ the full range of NTC digital data and possibly supporting data such as the commo tapes

Despite the limitations of the results, several findings are of note There was considerable variation in performance of the task forces measured either in terms of percentage of vehicles lost or by the casualty exchange ratio Variation in performance was greater for offensive missions than for defensive missions This suggests that either the causal mechanisms for offensive performance are more complex than for their defensive counterparts or that training in the elements of defense are more universally applied and incorporated.

Another finding of note is that performance variation was least visible in the loss of tanks. Differences between type of task force and task force organization, when tested on the percentage of tanks lost or the causlaty exchange ratio for tanks, failed to produce any significant differences. As tanks constitute the primary weapon system on the NTC battlefield, performance characterisitics in this area are particularly important. The high level of loss in this area and the lack of

relationship to the analysis variables studied here indicate a more generic cause for the troubled performance The nature of this cause needs to be studied

The last significant finding from these results concerns the performance of the reorganized task forces (J-Series) Since performance for these task forces was not found to be consistently better than their earlier counterpart (H-Series) it is important to determine why this should be the case One possibility is that the addition of new assets independent of sufficient training in the deployment of those assets may result in inefficient and ineffective use That is, the new assets may in fact be consumed at a higher rate than the previous assets (under the prior organization) because the task force commander is unable to incorporate them into the scheme of the maneuver Whether this hypothesis is true or whether some other causes are at work can only be determined by the more comprehensive analyses to be performed in the future

As indicated in the Data Analysis Section, the analysis of force-on-force performance data was necessarily limited by the type and level of information available for analysis While recognizing those limitations, it was felt that there was much to be gained by exploring performance variations to determine the extent of their existence and whether they were related to the macro-level variables available in the Take Home Packages The results of this analysis have accomplished that objective They have demonstrated that variation does exist and that the macro-level variables while related do not account for all of the observed differences in performance Further, the analyses have pointed to several areas to be pursued in the more comprehensive analyses to be conducted with the NTC digital data Thus, the results reported here represent the jumping off point for this new research effort

TITLE. Use of Instrumentation to Improve Combat Readiness
AUTHOR William L. Shackelford
AFFILIATION: The BDM Corporation

1 Introduction (Slide 1) Soldiers sit hunched in their fighting positions intently peering into the still desert darkness

The task force intelligence officer is frantically attempting to contact his scouts by radio who are twenty kilometers to the front looking for the enemy. The commander is being briefed on the status of the defensive obstacle preparation within the task force sector The brigade operations and intelligence net is crackling with sightings of a large enemy force on the move. Suddenly, the stillness is shattered by the deafening impact of steel from massed enemy artillery. The S-2's scream is barely audible that the scouts are in contact with a force of 150 to 190 armored vehicles. (Slide 2)

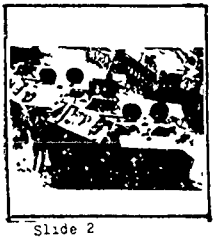
The enemy motorized rifle regiment is on the move and attacking into the guts of the task force defense

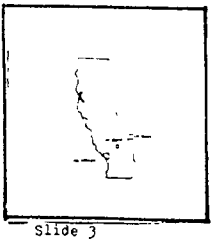
This action is repeated over and over again at the Army's National Training Center. It is but one of the battles that U.S. Soldiers will fight during their two weeks of pre-combat hell that prepares them to meet the Warsaw Pact in the defense of Europe.

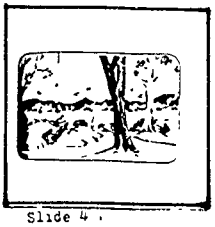
(Slide 3) The National Training Center was established in 1981 at Fort Irwin, California. Fort Irwin is located in the Mojave Desert at the point half way between Los Angeles and Las Vegas. The closest civilization is 40 miles away at the town of Barstow.

(Slide 4) The NTC provides a facility where units can conduct realistic training to develop individual, leader, and collective proficiency in combined arms operations. The environment replicates combat conditions more realistically than can be accomplished at home station due to physical limitations or resource constraints.

Slide 1

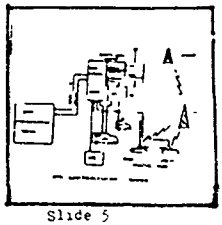

Slide 2


Slide 3

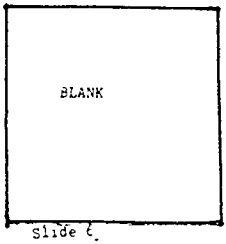

Slide 4

The NTC provides the Army with the ability to train and evaluate soldiers, leaders, and units in the execution of Army Airland Battle Doctrine (Slide 5)

Unique to the NTC is an unobtrusive instrumentation system which captures the actions of the task force in realtime. Instrumentation is used as training feedback and as an instrument to provide insights and solutions to Army-wide issues of major significance. Battle engagements are recorded by the use of an integrated position location system and multiple integrated Laser Engagement System, called MILES, affixed to each combat vehicle. Opposing U.S and enemy forces engage and cause casualties to each other through the use of eye-safe lasers which are attached to direct fire weapon system. Mobile field video units capture significant unit activities on tape Command and control communications are recorded for later feedback to the unit. Exercise control and observer controller assessments are supported by the communications system. All field events are transmitted from the simulated battlefield through a remote range station to a central operations center. It is in the Core Instrumentation System that all information is fed into central computers and training analysis personnel assess the performance of the unit (Slide 6)

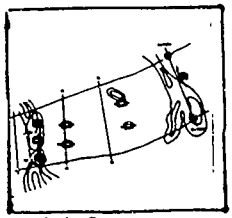


Slide 5

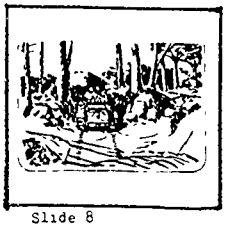


Slide 6

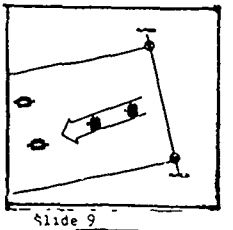
II. A Defense Scenario The NTC instrumentation system is the most powerful training system available to any army The instrumentation system allows the NTC trainers to gain insights into unit tactical performance not available through any other means. The results achieved by the instrumentation are fed back to the unit commander so that he may take the appropriate measures to correct his deficiencies. The following tactical scenario example will demonstrate the forceful utility of using instrumentation to improve the Army's fighting capability. (Slide 7)


Slide 7

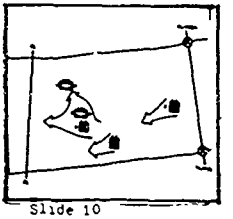
At 0500 hours U.S defending task force is deployed as follows: Scouts are to the front A company has two platoons defending forward with one platoon and the company headquarters tanks to the rear. Behind them are B and C Companies and the task force headquarters tanks. (Slide 8)


Slide 8

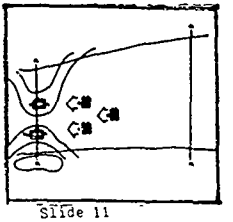
(Slide 9) At 0550 the motorized rifle regiment attacks into the task force defensive sector.

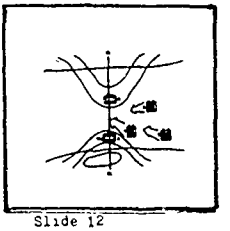

Slide 9

(Slide 10) At 0620 the lead motorized rifle battalion attacks the two forward defending platoons. With little loss of momentum the regiment reorganizes on the move after destroying the two forward platoons and drives deeper into the defenders position.

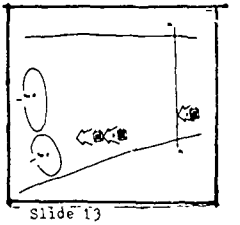

Slide 10

(Slide 11) At 0651 the regiment deploys to attack the remainder of A Company. At 0705 the regiment destroys the remaining A Company defenders who were deployed at the choke point in the terrain. (Slide 12)

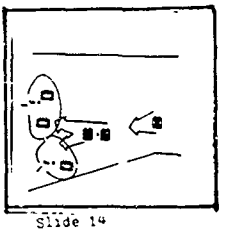

Slide 11


Slide 12

(Slide 13) At 0720 the regiment is deep into the defensive sectors 2000 meters from the defending B and C Companies.


Slide 13

At 0731 the regiment attacks into the final defenses of the task force and at 0759 defense is broken with elements of the regiment driving toward their assigned objective of the day. (Slide 14)


Slide 14

III. Analysis of Task Force Fire Control and Distribution of Tank Fires What went wrong? The task force had 36 hours to prepare the defense A lot of work has gone into preparing obstacles to slow the enemy The defense plan was well understood by all the leaders The commander knew this would be a tough fight but he never expected to be such an easy foe for the NTC opposing force regiment. These are some of the things that were going through the mind of the commander as he reflected on the battle The NTC instrumentation system gives the NTC trainers the capabilities to point out serious problems to the commander so he can take the appropriate training corrective actions. The following example will demonstrate how a major problem was revealed in the fundamentals of tank fire control and distribution which led to the defeat of the task force by the regiment

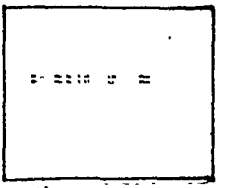
(Slide 15) The location and movement of the motorized rifle regiment in relation to the distance from the defenders was recorded using the color graphics position location capability of the system.

Slide 15

(Slide 16) The location of the opposing force is plotted on a graph by time and distance from the defenders

Slide 16

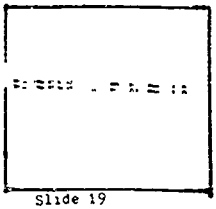
(Slide 17) Tank main gun firing events are captured from the system data base to reflect the volume of fire by the defenders.


Slide 17

(Slide 18) These tank firing events are plotted on a graph reflecting the number of rounds fired by the units at selected time intervals.

Slide 18

(Slide 19) The final element of information needed is to plot the effectiveness of the tank firings against the attacking regiment. Hits, kills, and near misses are recovered from the data base and placed upon a graph by event, range, and time the hit, kill, or near miss occured. (Slide 20)

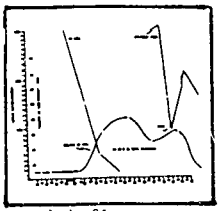


Slide 19

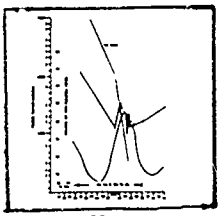
Slide 20

(Slide 21) By superimposing all three elements on a single graph a composite picture is then presented which depicts the effectiveness of the friendly unit's fire contol and distribution techniques. This composite pictures shows the initial fight of the forward platoons against the regiment The enemy attack was unempeded and friendly platoons were overrun.

(Slide 22) The volume of fire never increased until the enemy had overrun the platoon. The friendly fires were not effective and only resulted in killing four enemy vehicles and only after the platoons were overrun. This slide shows the results of the encounter between the remaining A Company platoon and the Company Command element Once again, the enemy is allowed to close in to a range of 1500 meters before the volume of fire and fire effectiveness causes eleven enemy vehicle kills. This is too little and too late because the motorized rifle battalion continues to press through the defenders.



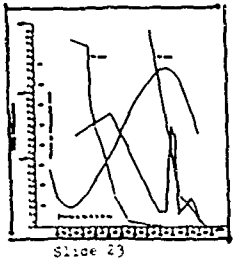
Slide 21



Slide 22

(Slide 23) The enemy regiment continues to press the attack into the last defenses of the task force. The 1st motorized rifle battalion is successful in closing on the defenders to be followed shortly by the 2nd motorized rifle battalion who completes the penetration. The friendly volume of fire is negligible against the 1st MRB and only increases when the defenders are being overrun by both MRB's

Fire effectiveness is again negligible with only a total of six opposing force combat vehicles being killed by the defender's tanks. (Slide 24)



Slide 23

BLANK

Slide 24

IV. Summary. The use of NTC instrumentation as you have seen during this presentation has a major impact on increasing the combat readiness of the U.S Forces. Armed with the clear picture of a serious problem he has within his task force, the commander can now take specific action to correct this deficiency The NTC instrumentation system gives the commander the tools to improve and adjust homestation training so that he is prepared to accomplish his mission in combat (Slide 25)

Slide 25

# Collecting and Utilizing Feedback Combined Arms and Services Staff School

Ralph W Ekwall, Ed. D
Ft. Leavenworth, Kansas

Since the Combined Arms and Services Staff School ($CAS^3$) is one of four schoolhouses within the Command and General Staff College (CGSC) one may ask why two different feedback systems exist within this college The answer is related to the substantial differences in the two programs. Since the programs differ greatly, their feedback requirements are also different. Without going into great detail, they differ in these respects length, rank of students, class size, curriculum design, evaluation system, and scope of instructor responsibility Some of these differences will be discussed in detail in the following section

## Background Information

Some background information is needed to fully understand the feedback program The Combined Arms and Services Staff School which we call CAS cubed is a fairly new school which was created to teach staff skills to Army captains The curriculum is divided into two phases. Phase I and Phase II Phase I is the non-resident portion which must be completed prior to attending the resident portion, it consists of 15 subcourses called modules The average student can complete Phase I in about 140 clock hours Phase II is the nine-week resident portion of the course. Phase II consists of six exercises varying in length from 4 to 15 days. Phase II exercises are subdivided into lessons Every student is assigned to a 12-person Staff Group This Staff Group is taught by an experienced lieutenant colonel called a Staff Leader who is responsible for the entire block of Phase II instruction. Since the same Staff Group stays with one Staff Leader for nine weeks, strong bonds of friendship are developed. The final course grade is a pass-fail grade, the emphasis is on providing immediate feedback on student performance for purposes of improvement We also emphasize the development of higher level mental skills such as analysis, synthesis and problemsolving, these skills are emphasized more than the mastery of information Hence, there are no exams in Phase II. This course represents the third mandatory educational experience for an Army captain. the first is the Officer Basic Course, the second is the Officer Advanced Course, and the third is CAS$^3$ This course may be the first course where the officer attends a school with officers of other branches The $CAS^3$ program makes Staff Group assignments to achieve maximal mixing of branches The $CAS^3$ course was extensively revised after its first pilot iteration in 1981. There was further revision after the second iteration. At present we are on a yearly rewrite schedule. Revision information is collected throughout the year and then used as a basis for a planned rewrite REvision information consists of information such as the following information extracted from doctrinal changes that must be integrated into the curriculum, command guidance, error correction and information collected from the $CAS^3$ Student Perception of Curriculum Relevance (SPCR) instruments All of these data sources are utilized in annual curriculum revisions, in this presentation the concentration is on creation, administration and utilization of SPCR The $CAS^3$ Coordinating Author's office has the responsibility of an after-action report to the Director after every nine-week class is completed The SPCR instrument is the basis for this report

## Purpose

The SPCR instrument was developed for the following reasons First of all, there was a need to collect student perceptions of the course so that curriculum writers and the Coordinating Author would know if the difficulty level was too high or too low They needed to know if the course was perceived as relevant, if the amount of time spent on various topics was appropriate and if students perceived the content as being current This is a good place to digress briefly on the topic of perception Student perceptions do not equate to student achievement Student perceptions tend to be soft data rather than hard data Nevertheless student perceptions are real, if students perceive a course as being irrelevant, too easy, too hard, or poorly presented, then the perception may be sufficiently real to cause the course to be poorly received or it may cause the course to fail Student perceptions are a fundamental basis for the success of any course To get back to the original topic -- we knew that student perceptions of $CAS^3$ were a critical element and we needed an instrument to collect these perceptions for purposes of curriculum revision A second reason for the development of the SPCR instrument was to provide feedback to Staff Leaders about the functions of their Staff Group and the morale of their Staff Group in comparison with other Staff Groups, and to do this on a timely basis so that corrective action could be taken Another aspect of the feedback process was providing information to the Director about morale and progress of the various Staff Groups

## Explanation of the System

The $CAS^3$ program is 42 days long The SPCR questionnaire is administered three times It is administered on day 9, on day 22, and day 37 The format is such that eight questions are asked about each lesson

TABLE 1

| | Responses - A--Strongly Agree, B--Agree, C--Neutral, D--Disagree, E--Strongly Disagree | |
|---|---|---|
| | Lesson 1, Staff Techniques Problemsolving | Lesson 2, Staff Techniques Military Writing |
| This lesson was valuable for an officer in my branch | 1. ____ | 9 ____ |
| Skills and information learned in this lesson will be useful to me in future assignments. | 2. ____ | 10 ____ |
| Feedback, critiques and after action reviews helped me to learn and improve my skills. | 3 ____ | 11 ____ |
| Phase I materials did a good job of preparing me for this lesson. | 4. ____ | 12 ____ |
| This lesson supported CAS$^3$ goals | 5 ____ | 13. ____ |
| The material in this lesson is current. | 6 ____ | 14 ____ |
| | Responses - A--Too High, B--High, C--About Right, D--Low, E--Too Low | |
| Judge the amount of time allocated to this lesson | 7 ____ | 15. ____ |
| Judge the difficulty level of this lesson | 8 ____ | 16 ____ |

We want to know if the lesson was relevant for an officer in that particular branch; we ask if knowledge and skills learned will be useful in future assignments, we ask about feedback, we ask about Phase I preparation, we ask if the lesson supports our four basic goals, we ask if lesson material is current, and we ask about difficulty and time allocations Note that questions 1-8 deal with Lesson 1 and questions 9-16 ask the same 8 questions about Lesson 2 and so on. Students are asked to give responses ranging from A=strongly agree to E=strongly disagree for each question. On questions 7 and 8, A=too high and E=too low Students write their responses on an IBM answer sheet Staff Leaders have the option of requiring students to first write their response on the questionnaire before putting responses on the IBM sheet The advantage of this procedure is that the Staff Leader receives feedback about his Staff Group's responses on an immediate basis Questions are asked about lessons and also about exercises

TABLE 2

| | Responses - A--Strongly Agree, B--Agree, C--Neutral, D--Disagree E--Strongly Disagree |
|---|---|
| | Staff Techniques Exercise |
| Morale in the Staff Group was good. | 89. _____ |
| Everyone in the Staff Group did their share of work | 90 _____ |
| | Responses - A--Very High, B--High, C--About Right, D--Low, E--Very Low |
| Judge the stress level | 91. _____ |
| Judge the workload | 92 _____ |

The last four questions refer to an entire exercise. They ask about morale, Staff Group functioning, stress level, and workload. The response forms are processed and the response data is placed in an Statistical Package for the Social Sciences (SPSS) file for analysis. This file contains information for the case data base on branch affiliation and Staff Group membership. The branch and Staff Group data provide us with a powerful tool for analysis of this data. Examples of analysis will be discussed in detail in the section following.

## Examples of Use

Army branches can be classified into four groups. (1) Combat Arms such as Infantry and Armor, (2) Combat Support such as Military Police, (3) Combat Service Support such as Finance and Adjutant General and (4) Professionals such as Army Nurse and Judge Advocate General. The professional branches usually receive professional training prior to entering the Army. They usually receive direct commissions and they usually enter the Army at a rank higher than second lieutenant. Their work in the Army resembles the work of their civilian counterparts more than it resembles the work of other officers in the Army. Our program wanted to know how these professionals perceived our program. We already know that the overall perception was positive, but we were not sure if this perception was shared by all branch groups. Nor did we know if different parts of our program were perceived equally by different branch groups. At this point a reminder is necessary. There are six exercises in the Phase II resident portion. The first exercise called Staff Techniques teaches students how to use several quantitative techniques. It teaches them how to prepare a variety of Army documents. The remainder of the exercises deal with Training Management, Budget, Mobilization, Preparation for Combat and fighting a war in Europe. First we looked at the professionals and the other branch groups in reaction to question 9 on the first administration.

TABLE 3

Lesson 2, F121, Military Writing
(First Administration)

| | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| Combat Arms | 79.8% | 14.9% | 5.3% | -- | -- |
| Combat Support | 83.7 | 11.6 | 4.7% | -- | -- |
| Combat Service Support | 75.0% | 15.6% | 9.4 | -- | -- |
| [illegible] | | | [illegible] | [illegible] | [illegible] |

"This lesson was valuable to an officer in my branch." This lesson topic was Military Writing One thing is remarkable there are no disagrees or strongly disagrees. Note that there is less enthusiasm among the professionals, but they apparently feel the lesson is valuable even though they show less enthusiasm

TABLE 4

Lesson 5, F121, Military Briefing
(First Administration)

| | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| Combat Arms | 84 0% | 13 8% | 2 1% | -- | -- |
| Combat Support | 81 4% | 16 3% | 2.3% | -- | -- |
| Combat Service Support | 81 3% | 18 8% | 0% | -- | -- |
| Professional | 66 7 | 26 7 | 6 7 | -- | -- |

Next, lets look at question 73 which refers to Lesson 5, Military Briefing and asks if that lesson was valuable to an officer in that branch. The response here is similar - just less enthusiasm Next, lets look at question 97, third administration of the SPCR.

TABLE 5

Lesson 12, F626, Division Operations Plan
(Third Administration)

| | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| Combat Arms | 42 9% | 45 7% | 9 5% | 1 9% | 0% |
| Combat Support | 27 7% | 66 0% | 6 4% | 0% | 0% |
| Combat Service Support | 25 7% | 51 4% | 20.0% | 9% | 0% |
| Professional | 26 3 | 42 1 | 26 3 | 0 | 5 3 |

The item asks how valuable the lesson on the Division Operation Plan is to an officer in that branch This data shows that the perception of professional branch officers is similar to Combat Service Support and Combat Support branch groups Our conclusion was that the professional branch members perceived some lessons less positively, but no action was required on this matter. We will now consider another example of use of the SPCR Remember that students are required to complete a set of 15 modules by correspondence prior to attending the resident portion We wanted to find out if students felt well prepared for the Phase II lessons We looked at responses to the question, "Phase I material did a good job of preparing me for this lesson "

TABLE 6

Phase II Lessons for Which Students
Felt Poorly Prepared

| Lesson | Percentage Who Agree and Strongly Agree |
|---|---|
| Lesson 6, F121, Meeting Management | 36% |
| Lesson 4, F121, Time Management | 37% |
| Lesson 3, F121, Quantitative Skills | 64% |

We found two lessons where students felt poorly prepared. The lessons were Lesson 2, Time Management and Lesson 4, Meeting Management both of which were in the Staff Techniques exercise. The responses to Lesson 3, Quantitative Skills are included for purposes of comparison It is the intent of our program to use Phase I to prepare students for Phase II, but in this instance we identified an oversight. This oversight is being considered by the curriculum committee.

The SPCR was used to analyze student perceptions of the Staff Ride. The Staff Ride is used to teach students how to analyze terrain. Students are taken by bus to various sites to practice terrain analysis uder the direction of Staff Leaders. There were some complaints from students so we decided to do an analysis of this lesson.

TABLE 7

Lesson 4, F626, The Staff Ride
(Third Administration)

| | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| Combat Arms | 31.4% | 41.0% | 18.1% | 7 6% | 1.9% |
| Combat Support | 36.2% | 38.3% | 14 9% | 10.6% | 0% |
| Combat Service Support | 20.0% | 34.3% | 31.4% | 8 6% | 5.7% |
| Professional | 36.8% | 36.8% | 21.1% | 5.3% | 0% |
| Total | 31.0% | 38.6% | 20 0% | 8.1% | 2 4% |

When we looked at it, this is what we saw. One analysis technique we often use is to combine agree and strongly agree responses. Here, the combined figures add up to 69% which is a more than 2/3 approval

TABLE 8

Lesson 7, F626, The Personnel Estimate
(Third Administration)

| | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| Combat Arms | 28 6% | 41.9% | 25.7% | 2.9% | |
| Combat Support | 14.9% | 48.9% | 29.8% | 4.3% | |
| Combat Service Support | 34 3% | 54.3% | 8.6% | 2.9% | |
| Professional | 26.3% | 57.9% | 10 5% | 0% | |
| Total | 26 2 | [illegible] | 21 9 | 2 9 | |

When we compared it to another, more or less representative lesson on the Personnel Estimate, the approval ratings were similar In this case we decided to do a little more analysis and to look at responses according to Staff Leaders. We looked at responses to the question, "Skills and information learned in this lesson will be useful to me in future assignments."

TABLE 9

Responses Analyzed According to Staff Leaders

Lesson 4, F626, The Staff Ride

| Staff Leader Identifier | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| D | 3 | 4 | 3 | -- | -- |
| G | 1 | 8 | 2 | 1 | -- |
| B | - | 6 | 3 | 1 | 3 |
| F | 3 | 2 | 3 | 3 | 2 |

(Analysis by Numbers)

Here are two examples of staff leaders who apparently convinced their students that the Staff Ride was critical to their future success and two other Staff Leaders who were not so successful. We decided, as a result of this analysis, to make sure that incoming Staff Leaders received training on how to conduct a Staff Ride This training has been implemented and we now hope to see an improvement in student perception of the Staff Ride.

The SPCR instrument is used in an analytical mode, but it is also used to provide feedback to instructors. You may remember that questions were asked about exercises in addition to the 8 questions asked bout each lesson Students are asked about their perception of morale in the Staff Group; they are also asked if everyone in the Staff Group is doing their share of work. Here is an example of the kinds of feedback provided to the Staff Leader. This is a response to the question, "Morale in the Staff Group is good."

TABLE 10

Example of Feedback Provided to Staff Leaders

| Staff Leader Identifier | Strongly Agree | Agree | Neutral | Disagree | Strongly Disagree |
|---|---|---|---|---|---|
| A | 4 | 6 | 0 | 1 | 0 |
| H | 0 | 1 | 9 | 0 | 0 |
| P | 6 | 6 | 0 | 0 | 0 |
| O | 2 | 3 | 3 | 3 | 0 |

(Analysis by Numbers)

Here is an example. Each Staff Leader receives information on all of the Staff Groups; he will know that his code letter is F or X or whatever, but he will not know the code letter of other Staff Leaders This gives a Staff Leader a chance to compare morale in his Staff Group with other Staff Groups. Since there are three administrations of this instrument, the Staff Leader can look at the status of morale at three different time periods. The Staff Leader can also compare morale levels with previous Staff Groups he has had. Note that in this instance Staff Leader H has nine people who are neutral about morale Morale in group P appears to be good. The same kind of information is provided in regard to student perception of other Staff Group members doing their share of work This information may cause that particular staff leader to re-look at a problem, or it may cause him to take corrective action.

## SUMMARY

This system which asks the same 8 questions about each lesson and then analyzes the responses by Staff Group and branch membership has proved to be quite successful.

# 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 and 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

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.

3 Each agency shall have no more than two (2) 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 issue 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 consultation 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 - Harry H Greer Award

A Selection Procedures

1 Recipients of the Harry H Greer Award will be selected by a committee drawn from the agencies represented on the MTA Steering Committee The CO of each agency will designate one person from that agency to serve on the Awards Committee Each committee member will have attended at least three previous MTA meetings The member from the coordinating agency will serve as chairman of the committee

2 Nominations for the award in a given year will be submitted in writing to the Awards Committee Chairman by 1 January of that year

3 The Chairman of the committee is responsible for canvassing the other committee members to arrive at consensus on the selection of a recipient of the award

4 No more than one person is to receive the award each year, but the award need not be made each year The Awards Committee may decide not to select a recipient in any given year

5 The annual selection of the person to receive the award or the decision not to make an award that year is to be made at least six weeks prior to the date of the annual MTA Conference

B Selection Criteria

1 The recipients of the Harry H Greer Award are to be selected on the basis of outstanding work contributing significantly to the MTA

C The Award

1 The Harry H Greer Award is to be a certificate normally presented to the recipient during the Annual MTA Conference The awards committee is responsible for preparing the text of the certificate The coordinating agency is responsible for printing and awarding the certificate

Article XII - 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 21 October 1985)

MTA STEERING COMMITTEE MEMBERS

US Army Research Institute

US Air Force Human Resources Laboratory

US Air Force 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 of Military Occupational Structures

Canadian Forces Personnel Applied Research Unit

Federal Republic of Germany Ministry of Defense

Royal Australian Air Force Evaluation Division

National Headquarters Selective Service System

US Navy Occupational Data Analysis Center

Defense Activity for Non-Traditional Education Support

# MINUTES

## MTA STEERING COMMITTEE MEETING
## 27TH ANNUAL CONFERENCE
## 21-25 OCTOBER 1985
## SAN DIEGO, CALIFORNIA

OPENING REMARKS

The meeting was opened by Dr Martin Wiskoff, the Conference Chairman for 1985

A list of the Steering Committee members present at this meeting is attached

The minutes of the 26th MTA Conference held in Munich, Germany were reviewed and unanimously approved by the Steering Committee

REQUEST FOR MEMBERSHIP

The Navy Occupational Data Analysis Center and the Defense Activity for Non-Traditional Education Support requested Primary Membership and were approved

HARRY GREER AWARD

A written nomination was submitted in December 1984 by COL Paul T Ringenbach, Commander, USAF Occupational Measurement Center for Mr Fred Hawrysh, of the Directorate of Military Occupational Structures, National Defense Headquarters, Ottawa, Canada for the 1985 Harry H Greer Award Following discussion, the Steering Committee voted in favor of the nomination.

BY-LAW CHANGES

Dr. John Ellis of the Navy Personnel Research and Development Center proposed that the by-laws be amended to include greater mention of training research and development After some discussion it was decided not to change the by-laws, but to provide greater emphasis to training in the annual call for papers and to expand the mailing list to include a broader range of training organizations

Two changes to the by-laws were approved. The intent of Article V, Section A of the by-laws was clarified to read "3. Each agency shall have no more than two representatives." A listing of current members of the MTA Steering Committee is to be included at the end of the by-laws. The revised by-laws appear in the 1985 proceedings

PROCEEDINGS GUIDELINES

There was discussion as to the desirability of establishing guidelines for submitting papers. A proposal will be made by the Federal Republic of Germany

MTA 1986

The U. S. Coast Guard Institute is scheduled to host the 1986 meeting, either in Oklahoma City, OK or in the New London, CT area. The 1986 MTA President will be CAPT Robert W. Davis and the Chairman will be Mr John Burt.

MTA 1987-1988

The Canadian Forces Directorate for Military Occupational Structures and the Canadian Forces Applied Research Unit will host the 1987 meeting in Ottawa and the Army Research Institute will host the 1988 meeting in an undetermined location.

ADJOURNMENT

There being no further business, the meeting was adjourned.

## STEERING COMMITTEE MEMBERS PRESENT

### 21 October 1985

| | |
|---|---|
| Carol V Ascherfeld | U S Navy Education and Training Program Development Center |
| Walter W Birdsall | U.S. Navy Education and Training Program Development Center |
| Arnold Bohrer | Belgian Armed Forces, Psychological Research Section |
| John A Burt | U.S Coast Guard Institute |
| John Ellis | U.S Navy Personnel Research and Development Center |
| SQNLDR Ken Given | Royal Australian Air Force Evaluation Division |
| Roger G. Goldberg | U S. Defense Activity for Non-Traditional Education Support |
| Wulf Gronwald | Federal Republic of Germany, Ministry of Defense |
| Robert F Holz | U S. Army Research Institute |
| CDR J E Olson | U S Navy Education and Training Program Development Center |
| LCOL Terry J Prociuk | Canadian Forces, Personnel Applied Research Unit |
| Martin L Rauch | Federal Republic of Germany, Ministry of Defense |
| COL P. T. Ringenbach | U.S Air Force Occupational Measurement Center |
| J S. Tartell | U S Air Force Occupational Measurement Center |
| Raymond O Waldkoetter | U.S Army Research Institute |
| Robert J Wilson | U S Navy Occupational Data Analysis Center |
| Martin Wiskoff | U.S Navy Personnel Research and Development Center |
| COL G. A Zypchen | Canadian Forces, Directorate of Military Occupational Structures |

# Military Testing Association
# 1985
# Harry H. Greer Award
# to
# FRED HAWRYSH

The Harry H. Greer Award is hereby presented to Mr. Fred Hawrysh of the Directorate of Military Occupational Structures, National Defence Headquarters, Ottawa, Canada, for outstanding work contributing significantly to the goals and purposes of the Military Testing Association.

You have been an active proponent of MTA for many years, first as an active-duty representative of the Canadian Forces and later as the senior civilian for your agency. We acknowledge your encouragement of MTA attendance and your leadership in stimulating quality presentations.

Your achievements in stimulating international exchange and cooperative work in the field of occupational analysis are outstanding. The continuing success of the Comprehensive Occupational Data Analysis Programs (CODAP) reflects positively on your dedication and enthusiasm.

The devotion and professionalism of your work are of great credit to yourself and the association. With gratitude, appreciation, and friendship of all those affiliated, the MTA recognizes Fred Hawrysh with this Harry H. Greer Award.

MTA
THE MILITARY TESTING ASSOCIATION

Howard S Eldredge
MTA President
Commanding Officer
Navy Personnel Research
and Development Center

# MTA 27TH ANNUAL CONFERENCE STAFF

MTA PRESIDENT

CAPTAIN HOWARD S. ELDREDGE

MTA CHAIRMAN

DR. MARTIN F WISKOFF

COMMITTEE CHAIRPERSONS/MEMBERS

PROGRAM COMMITTEE

DR. JOHN PASS
DR. JOHN ELLIS
DR NORM ABRAHAMS

OPERATIONS COMMITTEE

MR. ROB WELLS
MRS. MARGIE SANDS
MS. PAT MARSH
MR JIM JULIUS
MS JESSIE GRIER
ETI JIM LAW

FINANCE

MS. SUE STUMPF

PUBLICATIONS/GRAPHICS

MR. GENE STOUT
MS. DARLA LEITHISER
MR. PORTER WOOTEN
MR TED YELLEN
MS. CARMEN SCHEIFERS

HOSPITALITY

DR. JOHN ELLIS
DR JOHN PASS

"A"

*DR. NORM ABRAHAMS
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*HOMER W. ADKINS
US NAVY CHIEF OF NAVY EDUCATION
AND TRAINING (CNET N-131)
NAS PENSACOLA, BLDG #624
NAS PENSACOLA FL 32508

*DR. JEFFREY W. ANDERSON
12760 GAZEBO COURT
WOODBRIDGE, VA 22192

*WAYNE A. ANDERSON
HQ USMEPCOM MEPCT-M
2500 N. GREENBAY RD
NORTH CHICAGO, IL 60064-3094

*ROBERT J ANGUS
CANADIAN FORCES PERSONNEL APPLIED
RESEARCH UNIT
4900 YONGE STREET, SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

*MR. THOMAS M ANSBRO
CNET N-2XX CNET HQ
NAS
PENSACOLA FL 32508

*DR JANE M ARABIAN
COMMANDER
U S. ARMY RESEARCH INSTITUTE
PERI-RS (ATTN: Dr ARABIAN)
5001 EISENHOWER AVE
ALEXANDRIA, VA 22333-5600

*DR. CAROL V. ASCHERFELD
NAVEDTRAPRODEVCEN (CODE 3143)
SAUFLEY FIELD
PENSACOLA, FL 32509-5000

*LEANNE F ATWATER
NAVY PERSONNEL R&D CENTER
CODE 04
SAN DIEGO, CA 92152-6800

*NANCY K. ATWOOD
ARI, PRESIDIO OF MONTEREY FIELD UNIT
P O BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

"B"

*MS ANNETTE G BAISDEN
NAVAL AEROSPACE MEDICAL RESEARCH INST.
NAMI (CODE 11)
PENSACOLA FL 32505

*DR HERBERT GEORGE BAKER
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*MR THOMAS B BAKER
COMMANDER
U S. ARMY ORDNANCE CENTER & SCHOOL
ATTN ATSL-TD-TDA (MR. BAKER)
ABERDEEN PROVING GROUND, MD 21005-5201

*DR. JAMES H BANKS
U S ARMY RESEARCH INSTITUTE
FIELD UNIT
P O BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*WILLIAM G. BARNES
COMMANDANT
U S ARMY CHEMICAL SCHOOL
ATTN: ATZN-CM (MR. BARNES)
FT MCCLELLAN, AL 36205-5000

*MS. JOANN BARSIS
PERSONNEL PERFORMANCE INC
3304 BUCKEYE LANE
FAIRFAX, VA 22033

*DR WILLIAM M. BART
EDUC PSYCHOLOGY
330 BURTON HALL
UNIVERSITY OF MINNESOTA
178 PILLSBURG DRIVE S.E
MINNEAPOLIS, MN 55455

*HELEN F BELLETTI
1908 CHRIS SCOTT DRIVE
EL PASO, TX 79936

*STEPHEN G. BENIGNI
NODAC
BLDG 150, WASH. NAVY YARD
WASHINGTON, D.C 20374

*CAPT JEANMARIE BESSET
AIR FRANCE DONJ
ADMINISTRATIVE
BP 10201
95703 ROISSY CH DEGAULLE CEDEX
FRANCE

*MAJ CONRAD G. BILLS
INSTRUCTIONAL SYSTEMS DEV DIV
93 BMW/DOS
CASTLE AFB, CA 95342-5000

*WALTER W. BIRDSALL
NAVEDTRAPRODEVCEN (CODE 31)
SAUFLEY FIELD
PENSACOLA FL 32561

*CAPT JACK L. BLACKHURST
AFHRL/MODP
BROOKS AFB, TX 78235-5000

*HARRIET H BLAKELY
US MEPCOM
MILITARY ENTRANCE PROCESSING ST.
(ATLANTA)
80 TENTH ST. N.E.
ATLANTA, GA 30309-3957

*BRUCE BLOXOM
NAVAL POSTGRADUATE SCHOOL
P O BOX 222375
CARMEL, CA 93922

*FREDERICK L. BLUME
CNET
3595 BROOKSHIRE DRIVE
PENSACOLA, FL 32504

*DENNIS M. BYLTH
ARMY PERSONNEL RESEARCH ESTABLISHMENT
C/O RAE FARNBOROUGH
FARNBOROUGH
HAMPSHIRE, UNITED KINGDOM

*ARNOLD C. BOHRER
SELECTIE-EN REKRUTERINGS CENTRUM
SECTIE PSYCHOLOGISCH ONDERZOEK
BRUYNSTRAAT
B-1120 BRUSSEL, BELGIUM

*JOHN BONDARUK
NATIONAL SECURITY AGENCY
FORT GEORGE G. MEADE
MARYLAND, 20755-6000
ATTN: M324

*RICHARD L. BRANCH
HQ US MEPCOM/MEPC-T
2500 GREENBAY RD.
NORTH CHICAGO, IL 60060

*JOHN S. BRAND
CDR USASSC
ATTN: AISG-ES (BRAND)
FORT BENJAMIN HARRISON IN 46216-5370

*HENRY I. BRAUN
ETS (221) - RESEARCH
PRINCETON, NJ 08541

*ROBERT C BREEST
COMMANDER, NAVY RECRUITING AREA FIVE
NTC GREAT LAKES, BLDG 3
GREAT LAKES, IL 60088

*JOEANNE BRIDGE
ALBUQUERQUE MEPS
P.O BOX 103
ALBUQUERQUE, NM 87102-0103

*DR BRENT BRIDGEMAN
EDUCATIONAL TESTING SERVICE
11-R, ROSEDALE ROAD
PRINCETON, NJ 08541

*ARTHUR M. BROWN
US MEPCOM
HQ E SECTOR
FORT MEADE, MD 20755

*DR JOHN R. BRUNI
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*JANICE BUCHHORN
AFHRL/TSOZM
BROOKS AFB TX 78235

*COMMANDING OFFICER
ATTN LINDA BURGUM
VS41, BLDG. 797
NAS, NORTH ISLAND
SAN DIEGO, CA 92135

*LAWRENCE F. BURNS
NAVY RECRUITING COMMAND
4015 WILSON BLVD.
ARLINGTON, VA 22203

*JOHN A. BURT
USCG INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

*LLOYD D. BURTCH
AFHRL/XR
BROOKS AFB TX 78235

*BRIAN J. BUSH
ARI
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*WALTER G. BUTLER
DIRECTOR, US ARMY TRADOC SYSTEMS
ANALYSIS ACTIVITY
ATTN: ATOR-THD
WHITE SANDS MSL RNG, NM 88002-5502

"C"

*WAYNE J CAMARA
HUMAN RESOURCES RESEARCH ORG
1100 SOUTH WASHINGTON ST.
ALEXANDRIA, VA 22314

*EDMUND J CARBERRY
US ARMY ARMOR SCHOOL
7303 RIDAN WAY
LOUISVILLE, KY 40214

*CW03 PETER F. CASE
U.S. COAST GUARD (PMR-5)
2100-2ND STREET S W
ROOM 4406
WASHINGTON, D.C 20593-0001

*JEANNA F. CELESTE
WESTAT, INC.
1650 RESEARCH BLVD.
ROCKVILLE, MD 20854

*DR. FREDERICK R CHANG
NAVPERSRANDCEN
CODE 51
SAN DIEGO, CA 92152-6800

*ROBERT E CHATFIELD
NAVPERSRANDCEN
CODE 623
SAN DIEGO, CA 92152-6800

*JOAN T. CHIPPENDALE
US ARMY RESEARCH INSTITUTE
PERI-IR
BLDG 501
FORT RUCKER, AL 36362

*GAYLE C. CHRISTIANSEN
US COAST GUARD HQ
2100 2ND ST SW
WASHINGTON, DC 20593

*DR. BARBARA R. CHURCH
BLDG. 262 RM. I-2
FEDERAL LAW ENFORCEMENT TRA CTR
GLYNCO, GA 31524

*RAY W. CLARK
NAVAL MILITARY PERSONNEL COMMAND DET
NAVY OCCUPATIONAL DEVELOPMENT AND
ANALYSIS CENTER
WASHINGTON NAVY YARD (ANACOSTIA)
BLDG 150
WASHINGTON, D C 20374-1501

*ANNE COATES-RADLOFF
HUMRRO
1100 SO. WASHINGTON STREET
ALEXANDRIA, VA 22314

*LCDR PAUL M. CONNOLLY
CNET CODE N-3324
NAS PENSACOLA, FL 32508

*MAJ PAUL COOK
HQ AFMPC/DPMYOT
RANDOLPH AFB, TX 78150-6001

*MERRI-ANN COOPER
ADVANCED RESEARCH RESOURCES
ORGANIZATION
4330 EAST-WEST HIGHWAY
BETHESDA, MD 20814

*DR CHARLES H. CORY
NAVY PERSONNEL R&D CENTER
CODE 621
SAN DIEGO CA 92152-6800

*DR. KENT CRAWFORD
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

*NEAL R CROWLEY
5111 SOUTH 8TH ROAD #101
ARLINGTON, VA 22204

"D"

*MAJ MICKEY R DANSBY
LMDC/AN
BLDG 678
MAXWELL AFB, AL 36112-5712

*DR CHARLES E DAVIS
OFFICE OF NAVAL RESEARCH
800 NORTH QUINCY STREET, CODE 442
ARLINGTON, VA 22217

*GENEVA DAVIS
3617 DALE HOLLOW RD.
ANNISTON, AL 36201

*RENEE DAVIS
6940 FORUM STREET
SAN DIEGO, CA 92111

*LEE ELLYN DAY
REHAB GROUP INC
1360 ROSECRANS, SUITE D
SAN DIEGO, CA 92106

*DR ESTHER F. DIAMOND
EDUCATIONAL & PSYCHOLOGICAL
CONSULTANT
721 BROWN AVENUE
EVANSTON, IL 60202

*CAPT BEN L. DILLA
AFIT/LSB
WRIGHT-PATTERSON AFB, OH 45433

*RONNA F DILLON
DEPT OF GUID. & EDUC PSYCH
SOUTHERN ILLINOIS UNIVERSITY
CARBONDALE, IL 62901

*DR STEVE DOCKSTADER
NAVY PERSONNEL R&D CENTER
CODE 72
SAN DIEGO CA 92152-6800

*DR LINDA M DOHERTY
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO CA 92152-6800

*WILLIAM J DOHERTY
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

*RICHARD D DOORLEY
USA MILPERCEN (DAPC-MSE)
200 STOVALL STREET
ALEXANDRIA, VA 22332

*DR. WALTER F DRISKILL
6422 FALLS CHURCH
SAN ANTONIO, TX 78247

*EUGENE H. DRUCKER
HUMRRO - FT. KNOX OFFICE
PO BOX 293
FORT KNOX, KY 40121

*CAPT R ERIC DUNCAN
AFIT/CIRP OL
UNIVERSITY OF TEXAS AT AUSTIN
1612 PLATEAU RIDGE
CEDAR PARK, TX 78613

"E"

*HEINZ-JURGEN EBENRETT
GERMAN ARMED FORCES OFFICE
(STREITKRAEFTEAMT)
POSTFACH 20 50 03
D-5300 BONN 2
FEDERAL REPUBLIC OF GERMANY

*JOHN M. EDDINS
CERL, UNIV OF ILLINOIS
103 S. MATHEWS ST , RM. 252
URBANA, IL 61801

*WILLIAM A. EDMUNDSON
ARMY EDUCATION CENTER
MACOMB & ARMISTREAD STREETS
FORT BRAGG, NC 28307-5000

*LINDA J EISMAN
302 DRAKE ST.
BENICIA, CA 94510

*ANA G. EKSTROM
ARI FIELD UNIT
P O BOX 223133
CARMEL, CA 93922

*DR. RALPH W. EKWALL
CAS3-USACGSC
ATTN: ATZL-SWB (DR RALPH EKWALL)
FORT LEAVENWORTH, KS 66027

*CAPT HOWARD S ELDREDGE
NAVY PERSONNEL R&D CENTER
CODE 00
SAN DIEGO, CA 92152-6800

*DR JOHN A ELLIS
NAVY PERSONNEL R&D CENTER
CODE 51
SAN DIEGO CA 92152-6800

*TATIANA S. EROHINA
830 C AVE , APT. F
CORONADO, CA 92118

"F"

*BENJAMIN A. FAIRBANK
C/O PERFORMANCE METRICS, INC
5825 CALLAGHAN ROAD, SUITE 225
SAN ANTONIO, TX 78228

*FIORENZO FERRARI
ITALIAN AIR STAFF
STATO MAGGIORE AERONAUTICA
VIE UNIVERSITA ROMA
ITALY

*ROBERT L FEUGE
MANTECH MATHETICS CORP.
P.O. BOX 26655
SAN DIEGO, CA 92126

*BERNARD J. FINE
HEALTH & PERFORMANCE DIVISION
USARIEM
ARMY R & D CENTER
NATICK, MA 01760

*DR GERALD P. FISHER
SENIOR SCIENTIST
HUMRRO
1100 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

*ELI FLYER
HUMRRO
10 SIERRA VISTA DRIVE
MONTEREY, CA 93940

*THOMAS K. FORSYTHE
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

*RONALD F FREEMAN
LOS ANGELES MEPS
5051 RODEO ROAD
LOS ANGELES, CA 90016

*ROBERT L FREY, JR.
HEADQUARTERS, US COAST GUARD
G-P-1/2, ROOM 4200 8
2100 2ND STREET, S.W
WASHINGTON, DC 20593

"G"

*DENNIS GAYNOR
TESTING DIRECTORATE
MEPCOM
1140 INVERRARY LANE
DEERFIELD, IL 60015

*LT.COL FRANK C. GENTNER, USAF
DEOMI/DSE
PATRICK AFB, FL 32925

*CHARLES M GIBBONS
U S ARMY ORDNANCE MISSILE SCHOOL
ATSK-TIL
REDSTONE ARSENAL, AL 35897-6600

*DR. KAROL GIRDLER
U S. ARMY RESERACH INSTITUTE
HQ AMC EUROPE
BOX 228
APO NY 09333

*KENNETH C GIVEN
AIR FORCE HUMAN RESOURCES LAB
AFHRL/MODJ
BROOKS AFB
SAN ANTONIO, TX 78235

*DR. DWIGHT J. GOEHRING
ARMY RESEARCH INSTITUTE
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*ROGER G GOLDBERG
DEPARTMENT OF THE NAVY, DANTES
CODE 10
PENSACOLA, FL 32509

*DR. LAWRENCE A. GOLDMAN
US ARMY SOLDIER SUPPORT CENTER-NCR
ATTN: ATZI-NOS-C (DR. GOLDMAN)
200 STOVALL STREET
ALEXANDRIA, VA 22332-0400

*DR ALEXANDER M. GOTTESMAN
CHIEF, CURRICULUM SUPPORT DIVISION
CODE 221
NAVAL HEALTH SCIENCES EDUCATION
AND TRAINING COMMAND
NAVAL MEDICAL COMMAND,
NAT'L CAPITOL REG
BETHESDA MD 20814-5022

*LINDA J. GRAHAM
HELICOPTER ANTI-SUBMARINE SQUADRON 10
TRAINING (ATTN: EDUCATION SPECIALIST)
BUILDING 626
NORTH ISLAND NAVAL AIR STATION
SAN DIEGO, CA 92103

*SCOTT E. GRAHAM
U.S ARMY RESEARCH INSTITUTE
9109 POMONA PL.
LOUISVILLE, KY 40272

*MR GERRY L GREWELL
MILITARY ENTRANCE PROCESSING STATION
1755 FIFTH AVE.
SAN DIEGO, CA

*LAURA W. GRIEGER
COMMANDANT, USAIS
ATTN: ATSH-I-V-C-O (MS. GRIEGER)
FORT BENNING GA 31905-5593

*CAROL A. GRIFFITHS
TESTING SECTION, ROOM 1708
FEDERAL BUILDING
1000 LIBERTY AVENUE
PITTSBURGH, PA 15222-4101

*WULF GRONWALD
GAF INSTITUTE OF AVIATION MEDICINE
DEPT VI AVIATION PSYCHOLOGY
POSTFACH 1264/FLI
8080 FURSTENFELDBRUCK
FEDERAL REPUBLIC OF GERMANY
kp off

"H"

*LCDR JANE P K. HAMMOND
NAVCMSCOL, CODE 011
DAM NECK
VIRGINIA BEACH, VA 23462

*JAMES P HANLON
230 E. QUEUR STREET
CHAMBERSBURG, PA 17201

*LAWRENCE M. HANSER
U.S ARMY RESEARCH INSTITUTE
ATTN: PERI-RS (DR LAWRENCE M. HANSER)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*FRANK D. HARDING
ADVANCED RESEARCH RESOURCES ORG.
4330 EAST-WEST HIGHWAY
BETHESDA, MD 20814

*2ND LT DAVID I HARDY
USAF OMC/OMYA
RANDOLPH AFB, TX 78150-5000

*GAIL R. HARDY
ROOM 423, SENIOR PSYCHOLOGIST (NAVAL)
ADMIRALTY ARCH IV
SPRING GARDENS, LONDON
SW1A2BL, ENGLAND

*DR JOAN HARMAN
US ARMY RESEARCH INSTITUTE
ATTN: PERI-IC
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*W ANDREWW HARRELL, PH D
DEPARTMENT OF SOCIOLOGY
THE UNIVERSITY OF ALBERTA
TURY 5-21
EDMONTON, ALBERTA T6G 2H4
CANADA

*JANE HATCH
NATIONAL COMPUTER SYSTEMS
1200 NEW HAMPSHIRE AVE. N.W
SUITE 360
WASHINGTON, D C 20036

*FREDERICK J HAWRYSH
NATIONAL DEFENCE HEADQUARTERS
ATTN DMDS 3
101 COL BY DRIVE, OTTAWA
ONTARIO, CANADA K1A 0K2
CANADA

*DR ALLYN HERTZBACH
US ARMY RESEARCH INSTITUTE
ATTN. USARI-PERI-RP
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*REBECCA D HETTER
NAVY PERSONNEL R&D CENTER
CODE 622
SAN DIEGO, CA 92152-6800

*DR. EDWARD N HOBSON
11713 LARIAT LANE
OAKTON, VA 22124

*PAUL W HOLLAND
21-T
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON, NJ 08541

*DR. FRED D HOLT
COMMANDANT
U S. ARMY CHEMICAL SCHOOL
ATTN. ATZN-CM-ES (DR. FRED HOLT)
FT MCCLELLAN, AL 36205-5000

*ROBERT F. HOLZ
US ARMY RESEARCH INSTITUTE
PERI-PO
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*MR CHARLES R HOSHAW
5920 BROOKVIEW DRIVE
ALEXANDRIA, VA 22310

*LEAETTA M HOUGH
PERSONNEL DECISIONS RESEARCH
INSTITUTE
43 MAIN ST SE #405
MINNEAPOLIS, MN 55414

*DR GLORIA B. HOUSTON
NAVAL EDUCATION TRAINING
& SUPPORT CENTER, PACIFIC (N-7)
SAN DIEGO CA 92132

*WILLIAM K HUNTSINGER
NAVY OCCUPATIONAL DEVELOPMENT
AND ANALYSIS CENTER
NAVAL MILITARY PERSONNEL COMMAND DET
BLDG. 150, CODE 30
WASHINGTON, D.C 20374

"I"

*ROBERT A. IRELAND, JR
U S. NAVY RECRUITING DISTRICT
5051 RODEO ROAD
LOS ANGELES, CA 90016-4795

"J"

*IAN L JACKSON
PUBLIC SERVICE COMMISSION
PSC/STAFF DEVELOPMENT BRANCH
L'ESPLANADE LAURIER, WEST TOWER
300 LAURIER AVENUE WEST
OTTAWA, ONTARIO
CANADA K1A 0M7

*DR JOHN W. JONES
ST PAUL INSURANCE CO
RMSD-THE ST. PAUL COMPANY
385 WASHINGTON STREET
ST. PAUL, MN 55102

*KAREN N JONES
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

*KEN JONES
ROYAL NAVY, UNITED KINGDOM
RNSETT
HMS NELSON
PORTSMOUTH, HANTS PO1 3HH
ENGLAND

*GLORIA JONES-JAMES
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

"K"

*PROF MICHAEL KASTNER
UNIVERSITY OF THE ARMED FORCES OF THE FEDERAL REPUBLIC OF GERMANY, FB WOW
WERNER-HEISENBERG-WEG 39
8014 NEUBIBERG, WEST GERMANY

*ROBERT S. KENNEDY
ESSEX CORPORATION
1040 WOODCOCK ROAD
SUITE 227
ORLANDO, FL 32803

*DALE J KERKMAN
TRAINING DEVELOPMENT UNIT
BLDG 90, NTC
ATTN NEW TECHNOLOGY OFFICE
GREAT LAKES, FL 60088-5704

*CDR ROBERT H. KERR
323 COLBY DR.
DARTMOUTH, NOVA SCOTIA
CANADA, B2V2B7

*WILLIAM F. KIECKHAEFER
RGI INCORPORATED
1360 ROSECRANS, SUITE D
SAN DIEGO, CA 92106

*DR. MELVIN J. KIMMEL
U S ARMY RESEARCH INSTITUTE
ATTN PERI-RP
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22310

*DR. GRENVILLE C KING
DIRECTOR
U S ARMY MANAGEMENT ENGINEERING TRAINING ACTIVITY
ATTN AMXOM-MT (DR. KING)
ROCK ISLAND, IL 61299-7040

*DEIRDRE J KNAPP
U S ARMY RESEARCH INSTITUTE
ATTN PERI-RF
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*DR EUGENIA M KOOS
CODE 772, NAS, NORTH ISLAND
COMASWWINGPAC
SAN DIEGO, CA 92135

*BRUCE KRAMER
CTB/MCGRAW-HILL
2500 GARDEN RD
MONTEREY, CA 93940

*DR LEONARD P. KROFKER
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*HANS J KUESSNER
U S ARMY RESEARCH INSTITUTE
TRL
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*JAMES H KVICALA
3748 WINKFIELD PL
COLUMBUS, GA 31909

"L"

*DR. G J. LAABS
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*DR ANITA LANCASTER
1325 CALVIN FOREST DRIVE
VIENNA, VA 22180

*RICHARD S LANTERMAN
COMMANDANT, U S COAST GUARD, G-P-1/2
2100 SECOND STREET SW
WASHINGTON, DC 20593

*MS JANICE LAURENCE
HUMAN RESOURCES RESEARCH ORGANIZATION
1100 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

*MS LINDA C LAWING
CHARLOTTE MEPS
P O. BOX 34129
CHARLOTTE, NC 28234

*DR RICHARD A LILIENTHAL
U.S ARMY CIVILIAN PERSONNEL CENTER
PECC-CMP
200 STOVALL STREET
ALEXANDRIA, VA 22332-0300

*LTC LARRY R LIPPINCOTT
CHIEF, OTEA FIELD OFFICE
ATTN: CSTE-FOB, BLDG T-48
FORT BLISS, TX 79916-7619

*GENERAL HUGH LONG
EDUCATIONAL TESTING SERVICE
1825 I. STREET, N W
WASHINGTON, D.C 20006

*MICHAEL W LORENZO
3224 YORKTOWN DRIVE
TALLAHASSEE, FL 32312

*ROBERT M. LUND
HQ, USMEPCOM
2500 GREEN BAY RD
NORTH CHICAGO, IL 60064

"M"

*DR. MILTON MAIER
CENTER FOR NAVAL ANALYSES
4401 FORD AVENUE
P O. BOX 16268
ALEXANDRIA, VA 22302-0268

*DR. WILLIAM L MALOY
CHIEF OF NAVAL EDUC AND TRAINING
CODE 00A
NAVAL AIR STATION
PENSACOLA, FL 32508

*CLESSEN J MARTIN
U S ARMY RESEARCH INSTITUTE
PERI-RS
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*JOHN J. MATHEWS
NTEC, CODE 10
NAVAL TRAINING CENTER
ORLANDO, FL 32813-7100

*JOYCE D MATTSON
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

*PAUL W MAYBERRY
CENTER FOR NAVAL ANALYSES
4401 FORD AVENUE
P O BOX 16268
ALEXANDRIA, VA 22302-0268

*MR. JERRY MCCABE
TEACHING & LEARNING TECHNOLOGIES
3110 WOODCREEK DRIVE
DOWNERS GROVE, IL 60515

*BARBARA L. MCCOMBS
DENVER RESEARCH INSTITUTE
SSRE DIVISION
UNIVERSITY OF DENVER
DENVER CO 80208

*DR CLARENCE C MCCORMICK
HQ USMEPCOM/MEPCT-P
2500 GREEN BAY ROAD
NORTH CHICAGO, IL 60064-3094

*PAMELA B MCCRARY
FEDERAL LAW ENFORCEMENT TRA CENTER
BLDG. 262
GLYNCO, GA 31524

*BARBARA A. MCDONALD
NAVY PERSONNEL R&D CENTER
CODE 51
SAN DIEGO, CA 92152-6800

*JEFFREY J MCHENRY
PERSONNEL DECISIONS RESEARCH
INSTITUTE
43 MAIN STREET SOUTHEAST
SUITE 405
MINNEAPOLIS, MN 55414

*KENNETH W. MEACHAM
COMMANDANT
USAOMMCS (ATTN ATSK-ES) DOES
STANDARDIZATION & ANALYSIS DIVISION
REDSTONE ARSENAL, AL 35897-6700

*DR. BARBARA M MEANS
HUMAN RESOURCE RESEARCH ORGANIZATION
(HUMRRO)
1100 S. WASHINGTON STREET
ALEXANDRIA, VA 22314

*DOUGLAS H MEBANE
COMMANDANT
DEFENSE LANGUAGE INSTITUTE
DLIELC/LEACT (MR. MEBANE)
LACKLAND AFB, TX 78236-5000

*LARRY L. MELIZA
ARI FIELD UNIT
PO BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*DR. ALBERT MELTER
PERSONALSTAMMAMT DER BUNDESLSEHR
ABT. II/3(OP2), MUDRA-KASERNE
KOLNER STR 262
D-5000 KOLN 90
FEDERAL REPUBLIC OF GERMANY

*CHARLES MICHALK
CNATRA CODE 3122
NAVAL AIR STATION
CORPUS CHRISTI, TX 78419

*DR ANGELO MIRABELLA
ARMY RESEARCH INSTITUTE
12508 KUHL ROAD
SILVER SPRING, MD 20902

*DR JAMES L. MITCHELL
MCDONNELL DOUGLAS ASTRONAUTICS CO
DEPT E422
926 TOEPPERWEIN ROAD
CONVERSE, TX 78109

*ANTHONY E MIZEN
CINCNAVHOME
HM NAVAL BASE
PORTSMOUTH, HANTS, PO1 3LR
ENGLAND

*DEBORAH MOHR
NAVY PERSONNEL R&D CENTER
CODE 72
SAN DIEGO, CA 92152-6800

*WILLIAM E. MONTAGUE
NAVY PERSONNEL R&D CENTER
CODE 05
SAN DIEGO, CA 92152-6800

*JUDY M MORACCO
3510 CARLOTTA STREET
PENSACOLA, FLORIDA 32503

*KATHLEEN MORENO
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

*THOMAS MUIR
CHIEF OF NAVAL OPERATIONS
(O. 14)
NAVY DEPT
WASHINGTON, DC 20350-2000

*DIXIE L. MURDOCK
306 AUSTIN LOOP
FT BENNING, GA 31905

*KEVIN R MURPHY
DEPT OF PSYCHOLOGY
COLORADO STATE UNIVERSITY
FORT COLLINS, CO 80523

"N"

*DELBERT M NEBEKER
NAVY PERSONNEL R&D CENTER
CODE P72
SAN DIEGO, CA 92152-6800

*WOLFGANG NEEF
UNIVERSITY OF THE ARMED FORCES OF THE
FEDERAL REPUBLIC OF GERMANY, FB WOW
WERNER-HEISENBERG-WEG 39
D-8014 NEUBIBERG, WEST GERMANY

*DR PETER F. NEWTON
NATIONAL SECURITY AGENCY
BOX 1, FANX III
FORT GEORGE G MEADE, MD 20755-6000

*JUDITH NICHOLS
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

"O"

*DARLENE M OLSON
19654 FRAMINGHAM DRIVE
GAITHERSBURG, MD 20879
ALEXANDRIA, VA 22333

*CDR. J E OLSON
NETPDC, CODE 03
PENSACOLA, FL 32509

*DR. ROLF OTTE
PERSONALSTAMMAMT DER BUNDESWEHR
MUDRAKASERNE
KOLNER STR. 262
5 KOLN 90
GERMANY

*DAVID J OWEN
DIRECTORATE OF MILITARY
OCCUPATIONAL STRUCTURES
NATIONAL DEFENCE HEADQUARTERS
101 COLONEL BY DRIVE
OTTAWA, ONTARIO
CANADA K1A0K2

"P"

*CAPT MARK L PARIS, PH.D
CHIEF, PSYCHOLOGY DIVISION
DIRECTORATE OF MENTAL HEALTH
U S DISCIPLINARY BARRACKS
FT. LEAVENWORTH, KS 66027-7100

*DR JOHN J. PASS
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO CA 92152-6800

*DR DAVID L. PAYNE
TDAC
3280 PROGRESS DR.
ORLANDO, FL 32826

*DR EARL PENCE
U S ARMY RESEARCH INSTITUTE
ATTN PERI-RL (DR EARL PENCE)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*DR NANCY N. PERRY
CNET CODE 00A2
NAS PENSACOLA
PENSACOLA FL 32514

*DR NORMAN G PETERSON
PERSONNEL DECISIONS RESEARCH INSTITUTE
43 MAIN ST. S E #405
MINNEAPOLIS, MN 55414

*DR ALLAN L. PETTIE
USATSCFE
ATIC-ITP
ATTN: DR ALLAN PETTIE
FORT EUSTIS, VA 23604-5206

*WILLIAM J PHALEN
AFHRL/MOMM
BROOKS AFB, TX 78235-5601

*EARL R PHILLIPS
M C AIR STATION CHERRY POINT
EDUCATION OFFICER (CODE TE)
MCAS CHERRY POINT, NC 28533

*VIRGINIA M PHILLIPS
U S. ARMY RECRUITING COMMAND
FEDERAL OFFICE BLDG 4-1-150
24000 AVILA ROAD
LAGUNA NIGUEL, CA 92677-1001

*JOHN D PICKETT, JR
AFROTC/XPX
MAXWELL AFB, AL 36112-6663

*DON PITTMAN
DOD TEST SPECIALIST
SPOKANE MEPS
U S COURT HOUSE
SPOKANE, WA 99201

*COR EARL H POTTER
DEPT OF ECONOMICS AND MANAGEMENT
US COAST GUARD ACADEMY
NEW LONDON, CT 06320

*STEPHEN PRESTWOOD
ASSESSMENT SYSTEMS CORPORATION
2233 UNIVERSITY AVENUE, SUITE 310
ST PAUL, MN 55114

*TERRY J PROCIUK
CANADIAN FORCES PERSONNEL
APPLIED RESEARCH UNIT
SUITE 600, 4900 YONGE STREET
WILLOWDALE, ONTARIO
CANADA M2N 6B7

*TONY P PRUCNAL
ALTERNATE TEST CONTROL OFFICER
ARMY EDUCATION CENTER
BUILDING 219
FT MYER, VA 22211-5050

*DR ALBERT P. PRUETT
NAMTRAGRU NAS MEMPHIS
P.O BOX 28731
MEMPHIS, TN 38128

"Q"

*RONALD J. QUAYLE
NATIONAL COMPUTER SYSTEMS
1101 30TH ST N W , SUITE 500
WASHINGTON, DC 20007

*MAJ JOHN C QUEBE
9851 FORTUNE RIDGE
CONVERSE, TX 78109

"R"

*EUGENE M RAMRAS
NPRDC, CODE 01A
SAN DIEGO, 92152-6800

*MARTIN L. RAUCH
MINISTRY OF DEFENSE - P II 4
POSTFACH 1328
53 BONN 1
FEDERAL REPUBLIC OF GERMANY

*LTCOL BRUCE J REED
COMMANDANT OF THE MARINE CORPS
(CODE TAP-30)
HQ, U.S. MARINE CORPS
WASHINGTON, D C 20380-0001

*DR JAMES A RIEDEL
NAVY PERSONNEL R&D CENTER
CODE 72
SAN DIEGO, CA 92152-6800

*BARRY J. RIEGELHAUPT
HUMRRO
1100 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

*COL PAUL T. RINGENBACH
COMMANDER
USAF OCCUPATIONAL MEASUREMENT CENTER/CC
RANDOLPH AFB, TX 78150

*MR FRANK L RIPKIN
NAVAL CIVILIAN PERSONNEL COMMAND
NCPC-132 (OP-140F2)
ARLINGTON ANNEX, ROOM 6824
WASHINGTON, DC 20350-2000

*DR GERD W RODEI
FREIW.-ANNAHMEZENTRALE
DER MARINE
EBKERTEGE-KASERNE
D - 2940 WILHELMSHAVEN
FEDERAL REPUBLIC OF GERMANY

*BOB B. RODGERS
8018 SUNSET PATH COURT
SPRINGFIELD, VA 22153

*KENDALL L. ROOSE
TRAINING AIRWING FIVE
ACADEMIC TRAINING DEPARTMENT
MILTON, FL 32570

*GERARDUS J ROOZENDAAL
ROYAL NETHERLANDS EMBASSY
OFFICE OF THE ARMY ATTACHE
4200 LINNEAN AVENUE N.W.
WASHINGTON, DC 20008

*HARVEY ROSENBAUM
AMERICAN INSTITUTE FOR RESEARCH
AIR
1055 THOMAS JEFFERSON ST, NW
WASHINGTON, DC 20007

*RODNEY L. ROSSE
PERSONNEL DECISIONS RESEARCH
INSTITUTE
43 MAIN STREET SE #405
MINNEAPOLIS, MN 55414

*PAUL G ROSSMEISSL
US ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333-5600

*GREGORY B RUNYAN
501 TAMPICO BLVD
PENSACOLA, FL 32506

"S"

*DENNIS P SACCUZZO
13253 OAX COURT
SAN DIEGO, CA 92129

*WILLIAM A SANDS
NAVY PERSONNEL RESEARCH AND
DEVELOPMENT CENTER (CODE 63)
SAN DIEGO, CA 92152-6800

*GARY G SARLI
US ARMY RESEARCH INSTITUTE
PO BOX 6057
FORT BLISS, TX 79906-0057

*CLARE N SCHOFER
14 TRAINING GROUP HQ
CANADIAN FORCES BASE WINNIPEG
WESTWIN, MAN
R2R OTO CANADA

*DR MARY A SCHRATZ
NAVY PERSONNEL R&D CENTER
CODE 632
SAN DIEGO, CA 92152-6800

*MARK SCHWARTZ
COMMANDER
U S. ARMY INTELLIGENCE SCHOOL
ATTN ATSI-EFS (DR SCHWARTZ)
FT. DEVENS, MA 01433-6301

*DANIEL O. SEGALL
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

*WILLIAM L SHACKELFORD
THE BDM CORPORATION
NORTH BUILDING
2600 GARDEN ROAD
MONTEREY, CA 93940

*NORMAN P SHERWOOD
ARMY RECRUITING COMMAND
HQ USAREC
ATTN. RO-RS
FT SHERIDAN, IL 60037

*DR JOYCE SHIELDS
THE HAY ASSOCIATES
1110 VERMONT AVE , SUITE 710
WASHINGTON, D C 20005

*THEODORE M. SHLECHTER
U S ARMY RESEARCH INSTITUTE
FT. KNOX FIELD UNIT
STEELE HALL
FT. KNOX, KY 40121-5620

*WAYNE SHORE
PERFORMANCE METRICS, INC.
5825 CALLAGHAN, SUITE 225
SAN ANTONIO, TX 78228

*LT COL LAWRENCE Q SHORT
AFHRL/MOAO
BROOKS AFB, TX 78235-5601

*DR ALLEN N SHUB
SHUB & ASSOCIATES
P O. BOX 572
BUFFALO GROVE, IL 60090-0572

*DR NORMAN M. SHUMATE
DIRECTORATE OF TRAINING AND DOCTRINE
US ARMY ARMOR SCHOOL
ATTN. ATSB-DOTD
FORT KNOX, KY 40121-5200

*CLARENCE V. SIEGNER
FLEASWTRACENPAC
HARBOR DRIVE & NIMITZ BLVD
ATTN. DR. C. V SIEGNER N8B
SAN DIEGO, CA 92147

*LOYD D SINGLETARY
CHIEF OF NAVAL EDUCATION AND TRAINING
CODE N-5321 NAS
PENSACOLA, FL 32508

*ELIZABETH P. SMITH
COMMANDER
U.S ARMY RESEARCH INSTITUTE
ATTN: PERI-RS (DR ELIZABETH SMITH)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*DR. MARGARET J SMITH
CODE 301, BLDG 2435, RM 137
NETPDC
PENSACOLA, FL 32509

*DUANE L. SOWERS
MEPS, PHOENIX
ATTN DUANE SOWERS
215 N. 7TH STREET
PHOENIX, AZ 85034

*GARNETT L. SPEARMAN
2328 TRAVIS PINES ROAD
AUGUSTA, GA 30906

*DR GLEN L SPIVEY
ACSC/CAE
BLDG 1402
MAXWELL AFB, AL 36112-5542

*DR. WILLIAM D SPRENGER
HQS, TRADOC
ATRO-TA (SPRENGER)
FT MONROE, VA 23651

*MICHAEL R. STALEY
TEXAS MAXIMA CORP.
8301 BROADWAY, SUITE 212
SAN ANTONIO, TX 78209

*PAUL P STANLEY II
USAFOMC/OMD (PAUL STANLEY)
RANDOLPH AFB, TX 78150

*BRIAN M STECHER
(EDUCATIONAL TESTING SERVICE)
815 COLORADO BLVD
SUITE #606
LOS ANGELES, CA 90041

*ROBERT P. STEEL
AFIT/LSB
WRIGHT-PATTERSON AFB, OH 45433

*STAN STEPHENSON
SOUTHWEST TEXAS STATE UNIVERSITY
SAN MARCOS, TX 78666

*CHARLES R STEWART III
COMMANDER
US ARMY INTELLIGENCE CTR & SCHL
ATTN AISI-ES
FORT HUACHUCA, AZ 85613-7000

*MAJ WILLIAM J STRICKLAND
15880 CLIFFBROOK CT.
DUMFRIES, VA 22026

*LTCOL WILLIAM J SUBLETTE
COMMANDANT OF THE MARINE CORPS
(CODE TDA-20)
HEADQUARTERS USMC
WASHINGTON, DC 20380-0001

*JAMES B. SYMPSON
NAVY PERSONNEL R&D CENTER
CODE 63
SAN DIEGO, CA 92152-6800

"T"

*JOESPH S TARTELL
USAFOMC/OMY
RANDOLPH AFB, TX 78150-5000

*MAURICE M TATSUOKA
CERL, UNIV OF ILLINOIS
AT URBANA-CHAMPAIGN
220 EDUCATION BUILDING
1310 SOUTH SIXTH STREET
CHAMPAIGN, IL 61820

*CARL J TAYLOR
AFHRL/MODP
BROOKS AFB, TX 78235-5000

*DAVID M THISSEN
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF KANSAS
LAWRENCE, KA 66045

*DR. JOHN P. THOMAS
RD#2, BOX 319
DELANSON, NY 12053

*JERRY L. THOMPSON
EDUCATIONAL TESTING SERVICE
PRINCETON, NJ 08541

*DR. RONALD B. TIGGLE
NAVY PERSONNEL R&D CENTER
CODE 631
SAN DIEGO, CA 92152-6800

*SHAPON TKACZ
ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE.
ALEXANDRIA, VA 22333

*JODY L. TOQUAM
PERSONNEL DECISIONS
RESEARCH INSTITUTE
43 MAIN STREET S E , SUITE #405
MINNEAPOLIS, MN 55414

*DR. MARVIN H. TRATTNER
US OFFICE OF PERSONNEL MANAGEMENT
1900 E STREET NW
WASHINGTON, DC 20413

*JAMES M. TRIPP
CDR, USA TRAINING SUPPORT CENTER
ATTN: ATIC-ITT (MR TRIPP)
FORT EUSTIS, VA 23604

*DR J. TWEEDDALE
TECHNICAL DIRECTOR
NAVY PERSONNEL R&D CENTER
CODE 01
SAN DIEGO, CA 92152-6800
.kp off

"U"

*DR GEORGE M. USOVA
110 NELSON DRIVE
NEWPORT NEWS, VA 23601

"V"

*DAVID VALE
ASSESSMENT SYSTEMS CORPORATION
2233 UNIVERSITY AVE., SUITE 310
ST PAUL MN 55114

*NICHOLAS H. VANMATRE
NAVY PERSONNEL R&D CENTER
CODE 52
SAN DIEGO CA 92152-6800

*PAUL P VAN RIJN
ARI (PERI-RL)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*DR. DAVID S VAUGHAN
DEPT E422, BLDG 277C/1/F7
MCDONNELL DOUGLAS ASTRONAUTICS CO.
951 HORNET DRIVE
HAZELWOOD, MO 63042

*SALLY I. VICKERS
HQ U.S. WESTERN MEPCOM
BLDG 1808
PRESIDIO OF S.F., CA 94129

*DR ROBERT VINEBERG
HUMRRO
27857 BERWICK DRIVE
CARMEL, CA 93923

*SUSAN J VOIDAHL
MILWAUKEE MEPS
1711 BIRCH RD
KENOSHA, WI 53140

"W"

*DR LLOYD W. WADE
MARINE CORPS INSTITUTE
SPECIAL PROGRAMS DEPARTMENT
PO BOX 1775
ARLINGTON, VA 22222-0001

*KATHLEEN E. WAGNER
ARMY CONTINUING EDUCATION CENTER
BLDG 2248
AFZG-PTS-TRE
FT. SAM HOUSTON, TX 78234-5000

*DR. MICHAEL P WAGNER
DYNAMICS RESEARCH CORPORATION
60 CONCORD STREET
WILMINGTON, MA 01887

*DR. HOWARD WAINER
RESEARCH AND STATISTICS GROUP
EDUCATIONAL TESTING SERVICE (21T)
PRINCETON, NJ 08541

*DR. RAYMOND O. WALDKOETTER
US ARMY SOLDIER SUPPORT CENTER
ATTN. ATSG-DSS (DP. R.O. WALDKOETTER)
BUILDING 1
FORT BENJAMIN HARRISON, IN 46216-5060

*CLINTON B WALKER
ARMY RESEARCH INSTITUTE (PERI-RS)
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

*GARY H. WANAMAKER
ALLEN CORPORATION
4053 8th Ave #27
SAN DIEGO, CA 92103

*DR. THOMAS A WARM
US COAST GUARD INSTITUTE
PO SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

*JONATHAN WARREN
2360 EUNICE ST
BERKLEY, CA 94708

*DR BRIAN K. WATERS
MANPOWER ANALYSIS PROGRAM
HUMRRO
1100 S. WASHINGTON ST.
ALEXANDRIA, VA 22314

*HARVEY WEBB
HQ U S. MEPCOM
2500 GREEN BAY ROAD
NORTH CHICAGO, IL 60064-3094

*JOSEPH L WEEKS
AFHRL/MODJ
BROOKS AFB
SAN ANTONIO, TX 78235

*CAPT TONI G WEGNER
AFHRL/MOAE
BROOKS AFB, TX 78235-5601

*JOHNNY J WEISSMULLER
TEXAS MAXIMA CORP
8301 BROADWAY STE 212
SAN ANTONIO, TX 78209

*ROBERT WELLS
NAVY PERSONNEL R&D CENTER
CODE 06
SAN DIEGO, CA 92152-6800

*MAJ JOHN R WELSH
AFHRL/MOAE
BROOKS AFB, TX 78235-5601

*MAJ KAROL W J WENEK
CANADIAN FORCES PERSONNEL APPLIED
RESEARCH UNIT
4900 YONGE STREET, SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

*CHARLES M WEST
CNET, CODE N-1A
NAS
PENSACOLA, FL 32508

*DOUGLAS C. WETZEL
NAVY PERSONNEL R&D CENTER
CODE 51
SAN DIEGO CA 92152-6800

*MAJ COLIN P WHEELER
ARMY SCHOOL OF TRAINING SUPPORT
RAEC CENTRE
BEACONSFIELD, BUCKS
ENGLAND

*LEONARD A WHITE
U S ARMY RESEARCH INSTITUTE
MPRL-PERI-RS
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

*MR PATRICK J WHITMARSH
U.S. ARMY RESEARCH INSTITUTE FIELD UNIT
P O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

*DR GERRY L WILCOVE
NAVY PERSONNEL R&D CENTER
CODE 62
SAN DIEGO, CA 92152-6800

*WOLFGANG WILDGRUBE
STREITKRAEFTEAMT
POSTFACH 20 50 03
D-5300 BONN 2
WEST GERMANY

*RICHARD C WILLING
US COAST GUARD INSTITUTE
PO SUBSTATION 18 (MVP)
OKLAHOMA CITY, OK 73169-6999

*MAJ F. P WILSON
CANADIAN FORCES.. DPSRSC
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, CANADA K1A 0K2

*ROBERT J WILSON
NAVAL MILITARY PERSONNEL COMMAND DET
NAVY OCCUPATIONAL DEVELOPMENT AND
ANALYSIS CENTER
BLDG 150, WASHINGTON NAVY YARD
ANACOSTIA
WASHINGTON, D C. 20374-1501

*SUSAN T WILSON
EDUCATIONAL TESTING SERVICE 12-R
PRINCETON, NJ 08541

*DR WINFORD D. WIMMER
636 BRENTWOOD DRIVE
ANNISTON, AL 36206

*S. WINDLE
ALLEN CORPORATION
10080 CARROLL CANYON RD
SAN DIEGO, CA 92131

*DR. HILDA WING
809 THAYER PLACE
SILVER SPRING, MD 20910

*DR. MARTIN F. WISKOFF
NAVY PERSONNEL RESEARCH
& DEVELOPMENT CENTER
CODE 06
SAN DIEGO CA 92152-6800

*DR. JOHN WOLFE
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

*MR. ROBERT J. WOLFINGTON
COMMANDER
U.S. ARMY AVIATION CENTER
ATTN ATZQ-TD-IV-E (MR. WOLFINGTON)
FORT RUCKER, AL 36360-5000

*JO ANNA WOOD
DEPARTMENT OF PSYCHOLOGY
SOUTHERN ILLINOIS UNIVERSITY
CARBONDALE, IL 62901

*DARRELL A WORSTINE
SOLDIER SUPPORT CENTER-NCR
200 STOVALL STREET
ALEXANDRIA, VA 22332-0400

*GREGG J WRIGHT
BOOZ, ALLEN & HAMILTON
7315 WISCONSIN AVE - 1100W
BETHESDA, MD 20814

"Z"

*TIMOTHY C ZELLO
COMMANDER
U.S. ARMY ORDNANCE CENTER & SCHOOL
ATTN: ATSL-DES-E (TIMOTHY ZELLO)
ABERDEEN PROVING GROUND, MD 21005-5201

*COLONEL GERRY A ZYPCHEN
NATIONAL DEFENCE HEADQUARTERS
ATTN. DMOS
2781 SPRINGLAND DRIVE
OTTAWA, ONTARIO
CANADA, K1V 9X2

AUTHOR INDEX


| AUTHOR | PAGE |
|---|---|
| ABRAHAMS, N.M. | 503 |
| ADAMS, J. | 210 |
| AMYOT, CAPT K.A. | 462 |
| ANDERSON, J.W. | 153 |
| ANGUS, CAPT R.J. | 547 |
| ANSBRO, T.M. | 701 |
| ARABIAN, J.M. | 300,603 |
| ARCHER, BG C.J. | xxvi |
| ARMIJO, L. | 559 |
| ATWATER, L. | 113 |
| ATWOOD, N.A. | 798 |
| BAKER, H.G. | 327,567,780 |
| BALLENTINE, R.D. | 306 |
| BANKS, J.H. | 916 |
| BARGE, B.N. | 582,891 |
| BART, W.M. | 707 |
| BEARDEN, R.M. | 317 |
| BELL, 1LT J.M. | 228 |
| BILLS, MAJ C.G. | 666 |
| BLACK, B.A. | 786 |
| BLACKHURST, CPT J.L. | 306,327 |
| BLYTH, D.M. | 741 |
| BOLDOVICI, J.A. | 444,826 |
| BONCZAR, T.P. | 491 |
| BONDARUK, J. | 543 |
| BORMAN, W.C. | 367 |
| BRANDT, D.A. | 730 |
| BRIDGEMAN, B. | 337 |
| BROWN, G.E. | 650 |
| BURT, J.A. | 352 |
| BUSH, B.J. | 838 |
| BUTLER, W.G. | 204 |
| CANTOR, K.A. | 832 |
| CECIL, S. | 486 |
| CELESTE, J.F. | 526 |
| CHANG, F.R. | 553 |
| CHATFIELD, R.E. | 222 |
| COOPER, M. | 621 |
| CORPE, V.A. | 885 |
| CORY, C.H. | 775 |
| COSTELLIC, MAJ M. | 194 |
| CRAWFORD, A. | 113 |
| CRAWFORD, K.S. | 73 |
| CROWLEY, N.R. | 159 |
| DANSBY, MAJ M.R. | 531 |
| DIAMOND, E.E. | 570 |
| DILLA, CPT B.L. | 118,216 |
| DOCKSTADER, S.L. | 61 |
| DOHERTY, L.M. | 480 |
| DOHERTY, W.J. | 928,931 |
| DRISKILL, W.E. | 198 |
| DUNCAN, CPT R.E. | 713 |
| DUNLAP, W.P. | 107 |
| DUNNETTE, M.D. | 885 |
| EATON, N.K. | 769 |
| EBENRETT, H. | 251 |
| EDDINS, J.M. | 13 |
| EKSTROM, A.G. | 792 |
| EKWALL, R.W. | 943 |
| ELLINGSWORTH, M.E. | 420 |
| ELLIS, MAJ R.T. | 547 |
| FAIRBANK, B.A. | 136 |
| FINE, B.J. | 843 |
| FISHER, G.P. | 897,900 |
| FORSYTHE, T.K. | 928 |
| FRENCH, C.M. | 809 |
| FREY, R.L. | 736 |
| GARCIA, S.K. | 262 |
| GAST, I.F. | 147 |
| GIRDLER, K. | 130 |
| GOEHRING, D.J. | 239,655 |
| GOLDMAN, L.A. | 358,537 |
| GORMAN, LTC C.D. | 198 |
| GOTTESMAN, A.M. | 165 |
| GRAHAM, S.E. | 826 |
| GRIEGER, L.W. | 644 |
| HANLON, J.P. | 759 |
| HANSER, L.M. | 300 |
| HARDING, F.D. | 90,467 |
| HARDY, G.R. | 747 |
| HARDY, LT D.L. | 198 |
| HARMAN, J. | 404 |
| HART, R.J. | 655 |
| HEISEY, J.G. | 855 |
| HERTZBACH, A. | 587 |
| HETTER, R.D. | 503 |
| HIGHTOWER, CPT J.M. | 375 |
| HOLLAND, P.W. | 282 |
| HOUGH, L.M. | 582,891,900 |
| HOUSTON, J.S. | 885,891 |
| IBSEN, CPT K.A. | 531 |
| JOHNSON, D.M. | 615 |
| JOHNSON, J.H. | 398 |
| JOHNSON, R.M. | 102,587 |
| JONES, J.W. | 637 |
| JONES, K.W. | 352 |
| JONES, LCDR K. | 341 |
| JOYNER, J.N. | 323 |
| KAMP, J.D. | 891 |
| KANTOR, J.E. | 680 |

# AUTHOR INDEX


| AUTHOR | PAGE |
|---|---|
| KASTNER, N. | 438 |
| KENNEDY, R.S. | 107,398 |
| KERKMAN, D. | 393 |
| KERR, CDR R.H. | 88 |
| KERSHAW, S.W. | 288 |
| KIECKHAEFER, W.F. | 29,38,55 |
| KIMMEL, M.J. | 188 |
| KING, G.C. | 906 |
| KNAPP, D.J. | 102,587 |
| KOBRICK, J.L. | 843 |
| KOFFMAN, N. | 559 |
| KOROTKIN, A.L. | 621 |
| KROEKER, L.PL | 317 |
| LAABS, G.J. | 317,780 |
| LANCASTER, A.R. | 849 |
| LANE, N.E. | 107 |
| LARSON, G.E. | 233 |
| LAU, A. | 769 |
| LAURENCE, J.H. | 861 |
| LEWANDOWSKI, CPT F. | 216 |
| LILIENTHAL, R. | 897,900 |
| LOCKHART, J.M. | 615 |
| LOWE, CPT J.K. | 375 |
| MACPHERSON, D. | 515 |
| MAIER, M.H. | 311 |
| MALOY, W.L. | 124 |
| MATHEWS, J.J. | 809 |
| MATTSON, J.D. | 503 |
| MEANS, B. | 855 |
| MELIZA, L.L. | 661 |
| MERKLE, P.J. | 398 |
| MILLER, C.R. | 599 |
| MIRABELLA, A. | 520 |
| MITCHELL, J.L. | 265,271,276 |
| MIZEN, LCDR A.E. | 341 |
| MOHR, D.A. | 67 |
| MORABITO, CPT M.A. | 118 |
| MORENO, K.F. | 29,38,55 |
| MORRISON, R.F. | 222 |
| MUELLER, H.A. | 686 |
| MUMFORD, M.D. | 90,467 |
| McBRIDE, J.R. | 43 |
| McCOMBS, B.L. | 331 |
| McCORMICK, C. | 627 |
| McDONALD, B. | 450 |
| McGUE, M.K. | 891 |
| McHENRY, J.J. | 879 |
| McLAIN, LTC F. | 130 |
| McLAUGHLIN, D.H. | 730 |
| NEBEKER, D.M. | 78,82 |
| NEEF, W. | 438 |
| NEWTON, P. | 543 |
| NICHOLS, J.J. | 931 |
| NIGAM, A. | 855 |
| NULLMEYER, R.T. | 666 |
| OLIVIER, L.F. | 420 |
| OLSON, D.M. | 367 |
| PARIS, M.L. | 650 |
| PASS, J.J. | 565 |
| PENCE, E.C. | 922 |
| PERRIN, B.M. | 265,271 |
| PERRY, N.N. | 124 |
| PETERSON, N.G. | 867,873 |
| PETTIE, A.L. | 821 |
| PFEIFFER, G.J. | 420 |
| PHALEN, W.J. | 276,414 |
| PIERCE, J.E. | 537 |
| PITZ, G.F. | 171 |
| PLISKE, R.M. | 102 |
| POST, T. | 337 |
| POTTER, E.H. | 473 |
| QUEBE, J.C. | 695 |
| RAFACZ, B.A. | 23,567 |
| REE, M.J. | 672 |
| RICHARDS, J.D. | 210 |
| RIEDEL, J.A. | 78 |
| RILGELHAUPT, B.J. | 491 |
| RIMLAND, B. | 233 |
| RIPKIN, F.L. | 486 |
| ROBERSON, K. | 559 |
| ROSENBAUM, H. | 815 |
| ROSSE, R.L. | 873 |
| ROSSMEISSL, P.G. | 609,730 |
| RUMSEY, M.G. | 147 |
| SACCUZZO, D.P. | 233 |
| SACHS, S.A. | 632 |
| SAKO, S. | 763 |
| SANDS, W.A. | 19,567 |
| SARLI, G.G. | 615 |
| SCHNEIDER, E.F. | 861 |
| SCHRATZ, M.K. | 34 |
| SCHWARTZ, M.M. | 725 |
| SEGALL, D.O. | 29,38,55 |
| SELLMAN, W.S. | 852 |
| SHACKELFORD, W.L. | 937 |
| SHANE, G.S. | 426 |
| SHIFLETT, S. | 621 |
| SHLECHTER, T.M. | 444 |
| SHORT, LTC L.O. | 375 |
| SHUB, A.N. | 432 |

AUTHOR INDEX


| AUTHOR | PAGE |
|---|---|
| SLAUGHTER, LTC W.J. | 763 |
| SMITH, E.P. | 593 |
| SMITH, M.G. | 398 |
| SPRENGER, W.D. | 753 |
| STALEY, M.R. | 414 |
| STEEL, R.P. | 426 |
| STEPHENSON, S.D. | 346,456 |
| STERN, B.M. | 632 |
| STICHT, T. | 559 |
| TARTELL, J.S. | 228 |
| TATSUOKA, M. | 1 |
| TAYLOR, C.J. | 306 |
| THISSEN, D. | 294 |
| TIGGLE, R.B. | 23 |
| TKACZ, S. | 177 |
| TOQUAM, J.L. | 879,885 |
| TRATTNER, M.H. | 509 |
| TYERMAN, D. | 88 |
| USOVA, G.M. | 804 |
| VALE, C.D. | 7 |
| VAN RIJN, P. | 381 |
| VAUGHAN, D.S. | 265,271 |
| VINEBERG, R. | 323 |
| WAGNER, M. | 194 |
| WAINER, H. | 288 |
| WALDKOETTER, R.O. | 387 |
| WALKER, C.B. | 497 |
| WALKER, L. | 832 |
| WARM, T.A. | 142 |
| WARREN, J. | 543 |
| WATERS, B.K. | 719 |
| WEEKS, J.L. | 90,467 |
| WEGNER T.G. | 672,676 |
| WEISSMULLER, J.J. | 414 |
| WELSH, J.R. | 676 |
| WENEK, MAJ K.W.J. | 245 |
| WETZEL, C.D. | 43 |
| WHEELER, MAJ C.P. | 408 |
| WHITE, L.A. | 147,632 |
| WHITEHILL, B. | 450 |
| WHITMARSH, P.J. | 912 |
| WILCOVE, G.L. | 183 |
| WILDGRUBE, W. | 96 |
| WILKES, R.L. | 107 |
| WILSON, MAJ F.P. | 256 |
| WING, H. | 582,632,769 |
| WISE, L.L. | 300,730 |
| WOLFE, J.H. | 49 |
| WOOD, J.A. | 171 |
| WUEBKER, L.J. | 637 |
| YADRICK, R.M. | 265,271 |
| YATES, L.G. | 515 |