DTIC FILE COPY

PROCEEDINGS

28th Annual Conference of the

DTIC ELECTE SEP 19 1990 S B D

MILITARY TESTING ASSOCIATION

AD-A226 551

Coordinated by the

U. S. COAST GUARD ACADEMY

Department of Economics and Management

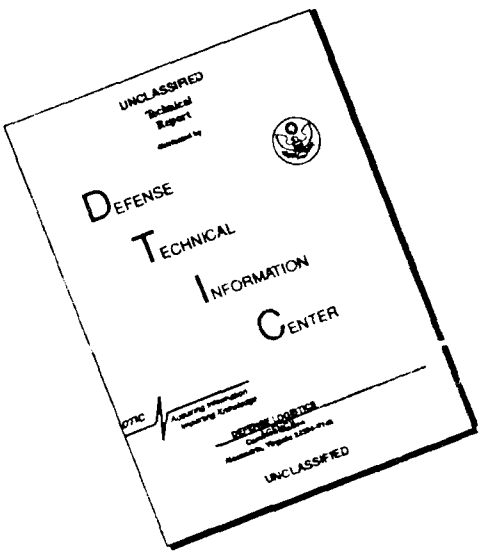
New London, Connecticut

03 - 07 November 1986

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

90 09 19 070

DISCLAIMER NOTICE



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

PROCEEDINGS

28TH ANNUAL CONFERENCE

of the

MILITARY TESTING ASSOCIATION

Coordinated by

the

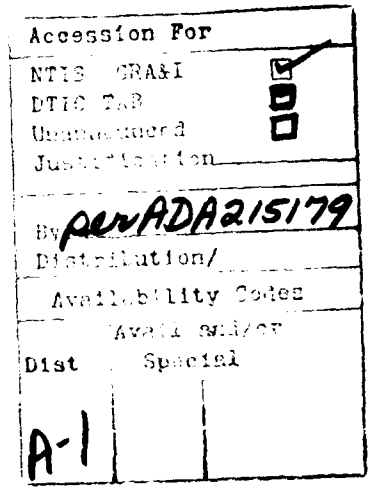
U. S. COAST GUARD ACADEMY

MYSTIC, CONNECTICUT

3-7 NOVEMBER 1986

FORWARD

The papers presented at the Twenty-Eighth Annual Conference of the Military Testing Association came from the business, educational, and military communities, both foreign and domestic. The papers reflect the opinions of their authors only and are not to be construed as the official policy of any institution, government, or branch of the armed services.



CONTENTS

VOLUME 1

SESSION	TITLE	AUTHOR	PAGE
	Foreward		ii
OPENING:	***Call To Order/Welcome***	**POTTER, CDR E.H. III**	xiv
	Keynote:	**CUERONI, RADM R.P.**	xiv
PAPER SESSION:	***Computer Adaptive Testing*** Chairperson: Stevens, G. U. S. Coast Guard Academy		
	An Update on the Computerized Adaptive Screening Test (CAST)	Knapp, D.J. Pliske, R.M.	1
	Personnel Assessment and Technological Advances	Steege, F.W.	7
	Relative Precision of Paper-and-Pencil and Computerized Adaptive Tests	Hetter, R.D. Segall, D.O.	13
	Correlations of Computer-Adaptive Test Time Measures with Thetas and a Criterion Measure	Czarnolewski, M.Y. Martin, C.J.	19
SYMPOSIUM:	***Innovations in Manpower Research Methods: Current Practice and Suggestions*** Chairperson: Wilson, M.J. Westat, Inc.		
	Using Focus Groups in Military Manpower Research	Elig, T.W. Pliske, R.M.	25
	The Use of Factorial Surveys in Military Personnel and Manpower Research	Wilson, M.J.	30
	Applications of Consumer Decision Models to Manpower Research	Romanczuk, A.P. Wilson, M.J.	36
	Using Confirmatory Factor Analysis to Aid in Assessing Task Performance	Harris, J.H. McHenry, J.J. Oppler, S.M.	42
	Estimation of Simultaneous Structural Equation Models: Applications in LISREL	Gilroy, C.L. Horne, D.K.	48
	An Event History Analysis of DEP Contract Losses	Celeste, J.F. Wilson, M.J.	54

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	***Issues in Selection Testing*** Chairperson: Frey, R. Commandant (G-P), U. S. Coast Guard		
	Assessing the Accuracy of the AFOQT Quick Score Procedure	Short, L.O. Wegner, T.G.	60
	The Effects of Reading Burden on SQT Performance	Brittain, C.V. Vaughan, P.R.	66
	Recall Versus Recognition in a Mathematics Test	Elliott, S.J.	72
PAPER SESSION:	***Training Validation*** Chairperson: Lanterman, R. Commandant (G-P), U. S. Coast Guard		
	Enhancing Validity and Reliability in Performance Oriented Training Assessment	Vail, CAPT K.W.	78
	Validation of Training - A Performance Oriented Approach	Donofrio, MAJ R.M. Thomson, CAPT M.W.	84
PAPER SESSION:	***Assessing Values in the U.S. Army*** Chairperson: Wehrenberg, S. Commandant (G-P), U. S. Coast Guard		
	Military Values: Structure and Measurement	Brosvic, G.M. Gilbert, A.C.F. Siebold, G.L. Tremble, T.R.	89
	Moskos' Institutional-Occupational Model: Reliability and Validity of One Operationalization	Brosvic, G.M. Tremble, T.R.	95
	MAR-VALS: A Microcomputer-Based External Evaluation System	Barnett, LT(N) E.G. Kerr, CDR R.H.	101
PAPER SESSION:	***Attitudes Towards Service*** Chairperson: Ferstyl, J. Commandant (G-PTE-4), U. S. Coast Guard		
	How Army Veterans View Their Military Experiences	Kimmel, M.J. Nogami, G.Y.	108
	Measuring Attitudes to Service in the British Army	Morris, V.	113
PAPER SESSION:	***Performance Testing*** Chairperson: Frey, R. Commandant (G-P), U. S. Coast Guard		
	Development and Evaluation of an Interactive Computer-Based Simulated Performance Test	Cantor, J.A. Walker, L.	119

SESSION	TITLE	AUTHOR	PAGE
	Gunnery Indices as Measures of Gunner Proficiency	Witmer, B.G.	125
	Stinger Team Performance in the Realistic Air Defense Engagement System (RADES)	Johnson, D.M. Lockhart, J.M.	131
PAPER SESSION:	***Training Evaluation*** Chairperson: Schoonmaker, LCDR C. U. S. Coast Guard Academy		
	Evaluation of Computer Based Training	Latimer, LT S.	137
	The Course Evaluation System	Ellis, J.A. Knirk, F.G. McDonald, B. Taylor, B.	143
	Validating Vocational Training Materials through the Instructional Quality Inventory (IQI)	Usova, G.M.	150
PAPER SESSION:	***Career Retention*** Chairperson: Wehrenberg, S. Commandant (G-P), U. S. Coast Guard		
	Measuring Factors Related to Reenlistment Decisions	Smith, A.L., Jr.	155
	Primary Motivational Variables in Military Career Decision-Making	McCombs, B.L.	160
	Effects of Varying Item Format on Test Performance	Day, L.E. Kieckhaefer, W.F.	167
	Live-Fire Training	Lister, S.G.	172
PAPER SESSION:	***Training in the Marine Corps*** Chairperson: Bonneau, LT A. U. S. Coast Guard Academy		
	Front-End Analysis for U.S. Marine Corps Training: Individual Training Standards	Davis, D.	178
	MATMEP: Individual Training Standards for Aviation Maintenance Military Occupational Specialties	Rogers, COL L.F., USMCR	182
	The Emergence of Collective Training Standards (CTS) in the Marine Corps	Main, R.	188
	Logistics Support Analysis (LSA) Validation through Training Evaluation	Dilg, M.	194

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	***Leadership*** Chairperson: Blake, LT R. U. S. Coast Guard Academy		
	Effects of Soldier Performance and Characteristics on Relationships with Superiors	Gast, I.F. White, L.A.	200
	Measurement of Leader Effectiveness in a Tactical Environment	Rachford, D.L. Zimmerman, R.A.	206
	Leader Requirements Task Analysis	Hunter, F.T. Steinberg, A.G. van Rijn, P.	212
SYMPOSIUM:	***Factor Analysis of Composite Scores from the Armed Services Vocational Aptitude Battery (ASVAB)*** Chairperson: Kennedy, R.S. Essex Corporation	Bittner, A.C., Jr. Dunlap, W.P. Jones, M.B. Kennedy, R.S.	218
SYMPOSIUM:	***Optimizing a Test Battery by Varying Subtest Times*** Chairperson: Kennedy, R.S. Essex Corporation	Dunlap, W.P. Jones, M.B. Kemery, E.R. Kennedy, R.S.	225
PAPER SESSION:	***Issues in Hands-On Performance Testing*** *Chairperson: Stevens, G.* U. S. Coast Guard Academy		
	Skill Requirement Influences on Measurement Method Intercorrelations	Campbell, C.H. Rumsey, M.G.	231
	Post Differences in Hands-On Task Tests	Hoffman, R.G.	237
PAPER SESSION:	***Automated Training Management*** Chairperson: Kerry, J. Commandant (G-PTE), U. S. Coast Guard		
	The Electronic Clipboard: A Central Requirement for Effective Automation of Training Management in Military Units	Atwood, N.K. Herman, J. Hiller, J.H.	242
	Automation of Army Unit Training	Goehring, D.J.	248
PAPER SESSION:	***CODAP*** Chairperson: Lanterman, R. Commandant (G-P), U. S. Coast Guard		
	Using CODAP Job Analysis for Training and Selection: Retrospective Considerations	Fisher, G.P. Hough, L. Lilienthal, R.	254
	IBM CODAP 370 - Alive and Well in Canada	Owen, D.	260

SESSION	TITLE	AUTHOR	PAGE
	Royal Navy Officers' Selection Scores and Success Beyond Initial Training	Drakeley, R.J.	267
PAPER SESSION:	***Improving Training Performance*** Chairperson: Kerry, J. Commandant (G-PTE), U. S. Coast Guard		
	The Effects of Remedial Training on Classroom Performance at the U.S. Army Ordnance Missile and Munitions Center and School	Illes, J.W.	273
	Learning Styles Inventories - Their Value and Use in the Navy Training Classroom	Dickson, A.M.	277
	Integrating Cognitive Learning Strategies into Training	Usova, G.M.	283
	Some Conditions Affecting Assessment of Job Requirements	Rossmeissl, P.G. Smith, E.P.	289
PAPER SESSION:	***Aptitude Testing*** Chairperson: Jones, K.N. U. S. Coast Guard Institute		
	Computerized Measurement of Vigilance Abilities	Cory, C.H.	295
	Aptitude Selectors for Air Force Officer Non-Aircrew Jobs	Arth, 1LT T.O. Skinner, M.J.	301
PAPER SESSION:	***Evaluating Performance in Training*** Chairperson: Kerry, J. Commandant (G-PTE), U. S. Coast Guard		
	Troubleshooting Proficiency Evaluation Project for the NATO Seasparrow Surface Missile System	Conner, H.B.	307
PAPER SESSION:	***Issues in Job Analysis*** Chairperson: Stevens, G. U. S. Coast Guard Academy		
	Developing Outlines for Specialty Knowledge Tests Using Occupational Survey Data	Longmire, 2LT K.M. Phalen, W.J. Weissmuller, J.J.	313
	Estimates of Task Parameters for Test and Training Development	Ford, P. Hoffman, R.G.	314
PAPER SESSION:	***Personnel Selection II*** Chairperson: Slimak, R. U. S. Coast Guard Academy		
	Royal Air Force Navigator Selection: Resolving an Old Problem	Burke, E.F.	320
	Conditional Logit Analysis for Personnel Selection	McLaughlin, D.H.	325

SESSION	TITLE	AUTHOR	PAGE
PAPER SESSION:	***Training Design*** Chairperson: Belmondo, LT W.J. U. S. Coast Guard Academy		
	A Methodology to Identify the Relative Importance of Training Requirements	Hawrysh, F.J. McMenemy, J.P.	331
	Retention of Common Military Skills	Brown, MAJ G.	337
	Review of Air Force Task Identification Methods and Data Bases	Garcia, S.K.	343
PAPER SESSION:	***Test Validation*** Chairperson: Blake, LT R. U. S. Coast Guard Academy		
	Relationship of SQT Scores to Project A Measure	Arabian, J.M. Mason, J.K.	348
	Test Validity in RAAF Air Traffic Controller Selection	Elliott, S.J.	354
PAPER SESSION:	***Evaluating Training Performance*** Chairperson: Belmondo, LT W.J. U. S. Coast Guard Academy		
	Two for the Price of One: Procedural Tutorial and Testing in Aircrew Training	Randel, J.M.	359
	Development of a New Systemof Measurement	Dillon, R.F. Reznick, R.K.	365
	Effectiveness of the Linking Format in a Technical Training Pamphlet	Bothwell, C. Jones, K.	369
PAPER SESSION:	***Design Issues in Job Analysis*** Chairperson: Slimak, R. U. S. Coast Guard Academy		
	Examination of Alternate Scales for Job Incumbent Occupational Surveys	Goldman, L.A. Worstine, D.A.	375
	Preliminary Holland Code Classification of Navy Entry-Level Occupations	Baker, H.G. Holland, J.L.	381
	The Job Difficulty Indes - Revalidating the Equationfor Supervisory Jobs	Given, SQNLDR K.C.	387
PAPER SESSION:	***Aptitude Testing*** Chairperson: Case, CWO4 P.F. Commandant (G-PMR), U. S. Coast Guard		
	The Development and Validation of the Peters Personnel Test	Peters, G.E.	393
	A Refined Item Digraph Analysis of a Cognitive Ability Test	Bart, W.M. Williams-Morris, R.	396

SESSION	TITLE	AUTHOR	PAGE
	Cigarette Smoking, Field-Dependence and Contrast Sensitivity	Fine, B.J. Kobrick, J.L.	401
	Personnel Variables and Organization/Mission Performance	Waldkoetter, R.O.	407
PAPER SESSION:	***Organizational Effectiveness*** Chairperson: Warm, T.A. U. S. Coast Guard Institute		
	Building Cohesion the Old Way: From the Ground Up	Holz, R.F.	413
	Influence of Environment, Ability and Temperament on Performance in Army MOS	Borman, W.C. Olson, D.M.	419
	Characteristics of Cognitive Performance in Stressful Environments	Banderet, L.E. Burse, R.L. Crohn, E.A. Cymerman, A. Roberts, D.E. Shukitt, B.L.	425
PAPER SESSION:	***Selection and Prediction for Air Traffic Safety*** Chairperson: Case, CWO4 P.F. Commandant (G-PMR), U. S. Coast Guard		
	Simulation Based Testing: A New Approach in Selecting ATC-Applicants	Haettig, J.H.	431
	The Application of a Human Factors Database in Army Aircraft Accident and Incident Investigation	Feggetter, A.J.W. McIntyre, H.M. Mortenson, L. Pritchard, B.	437
	Job Knowledge Test for Navy and Marine Jet Engine Mechanics	Alba, P.A. Baker, H.G.	443
	Inter-Service Technology Transfer: Performance Testing of Jet Engine Mechanics	Baker, H.G. Blackhurst, MAJ J.L.	448
SYMPOSIUM:	***Job Performance: What Do Soldiers Know, What Can They Do?*** Chairperson: Felker, D.B. American Institutes for Research		
	Patterns of Skill Level One Performance in Representative Army Jobs: Common and Technical Task Comparisons	Campbell, C.H. Campbell, R.C. Doyle, E.L.	454
	Effect of Practice on Soldier Task Performance	Edwards, D.S. Radtke, P.	459
	Effects of Test Programs on Task Proficiency	Ford, P. Hoffman, R.G.	465

SESSION	TITLE	AUTHOR	PAGE
SYMPOSIUM:	***Microcomputer-Based Occupational Information, Counseling, and Assignment for Naval Shipyards*** Chairperson: Mattson, J. Navy Personnel R&D Center		
	Clarifying Expectations through Audio-visual Job Preview	Baker, H.G. Wanous, J P.	470
	Microcomputer-Based Occupational Information and Counseling for Naval Shipyards	Norris, L.	474
	Developing a Microcomputer-based Assignment System for Shipyard Apprentices	Mattson, J.D.	480
PAPER SESSIONS:	***Selection*** Chairperson: Warm, D.A. U. S. Coast Guard Institute		
	A Testing Time for Naval Officers	Jones, A.	486
	Revision of Psychological Service of the Federal Armed Forces	Mademann, P.W. Puzicha, K.J.	492
PAPER SESSION:	***Computerized Testing*** Chairperson: Quedens, LT C. U. S. Coast Guard Academy		
	Does Microcomputer-Based Testing Encourage Truthful Responses?	Carroli, L. Doherty, L. Kantor, J. Rosenfeld, P. Thomas, M.	498
	Using a Shiphandling Simulator to Measure Sail, Boat, and Ship Experience	Evans, R.M.	504
PAPER SESSION:	***Organizational Effectiveness*** Chairperson: Warm, T.A. U. S. Coast Guard Institute		
	Army Civilian Personnel Research Program	van Rijn, P.	510
	Tradeoff Considerations in Providing Immediate Feedback to Organizations	Austin, CAPT J.S. Williams, CAPT M.S. Wood, LTCOL F.R.	516
	Mentoring, Chapter Three: Perceptions of Potential Proteges in the USAF	Dilla, CAPT B.L. Gouge, CAPT J.A.	522
PAPER SESSION:	***Validating Selection Criteria*** Chairperson: Quedens, LT C. U. S. Coast Guard Academy		
	Predicting Academic Success of Officer Candidates	Melter, A.H.	528

SESSION	TITLE	AUTHOR	PAGE
	Short- Versus Long-Term Tenure as a Criterion for Validating Biodata	Smith, E.P. Walker, C.B.	534
PAPER SESSION:	***Issues for Personnel Managers*** Chairperson: Case, CWO4 P.F. Commandant (G-PMR), U. S. Coast Guard		
	The Impact of Increased Training Time on National Guard Retention	Grissmer, D.W. Nogami, G.Y.	540
SYMPOSIUM:	***Project A Concurrent Validation: Preliminary Results*** Chairperson: Rossmeissl, P.G. Hay Systems, Inc.		
	The Project A Concurrent Validation Data Collection	Campbell, C.H. Campbell, J.P. Harris, J.H.	544
	The Development of a Model of the Project A Criterion Space	Campbell, J.P. Hanser, L.M. Wise, L.	550
	New Predictors of Soldier Performance	Ashworth, S. Hough, L. Peterson, N. Toquam, J.	556
	ASVAB Validities Usintg Improved Job Performance Measures	McHenry, J.J. Opper, S.H. Rossmeissl, P.G. Wise, L.L.	562
SYMPOSIUM:	***Assessing the Effects of Environmental Stressors and Treatment Strategies*** Chairperson: Banderet, L.E. U. S. Army Research Institute of Environmental Medicine		
	Development of Behavioral Assessment Protocols for Varied Repeated-Measures Testing Paradigms	Banderet, L.E. Kennedy, R.S. Lane, N.E. Wilkes, R.L.	568
	The Underestimation of Treatment Effects: Possible Clauses	Lieberman, H.R.	574
	The Use of Subjective Measures for Basic Problem-Definition	Munro, I. Rauch, MAJ T.M.	580
	Mood States at 1600 and 4300 Meters High Terrestrial Altitude	Banderet, L.E. Shukitt, B.L.	586

SESSION	TITLE	AUTHOR	PAGE
	Effects of Various Environmental Stressors on Cognitive Performance	Banderet, L.E. Burse, R.L. Crohn, E.A. Cymerman, A. Roberts, D.E. Shukitt, B.L.	592
SYMPOSIUM:	***Compterized Adaptive Testing Hardware/Software Development for the U.S. Military*** Chairperson: Sands, W.A. Navy Personnel R&D Center		
	Computerized Adaptive Testing Hardware/ Software Development for the U.S. Military: Introduction and Overview	Sands, W.A.	598
	Design and Development of the ACAP Test Item Data Base	Wilbur, E.R.	601
	Development of the Test Administrator's Station in Support of ACAP	Rafacz, B.A.	606
	Design and Development of the ACAP Test Administration Software	Jones-James, G.	612
	Communication of Computerized Adaptive Testing Results in Support of ACAP	Folchi, J.S.	618
SYMPOSIUM:	***Precommissioning Training in Army ROTC: Research and Policy Issues*** Chairperson: Hunter, F.T. U.S. Army Research Institute for the Behavioral Sciences		
	ROTC Cadet Subpopulations: Scholarship or Not	Elig, T.W. Hertzbach, A.	624
	Relationships Among Precommissioning Indicators of Army Officer Performance	Hunter F.T.	630
	Issues Concerning ROTC Intervention Programs	Twohig, P.	636
	Issues Involved in Establishing Basic Skills Standards	Oliver, L.W.	639
	Bylaws of the Military Testing Association		645
	Minutes of the Steering Committee		650
	MTA 28th Annual Conference Staff		654

SESSION	TITLE	AUTHOR	PAGE
	Military Testing Association Conference Registrants		655
	Author Index		665

OPEN SESSION OF THE 28TH ANNUAL MILITARY TESTING ASSOCIATION CONFERENCE 3 NOVEMBER 1986

The 28th Annual Meeting of the Military Testing Association was convened at 1300 on 3 November 1986 in the Hilton Inn in Mystic, Connecticut. CDR Earl H. Potter III, representing the host institution, the U. S. Coast Guard Academy in New London, Connecticut, welcomed the attendees. A slide introduction to the Coast Guard Academy and the surrounding area was presented by LT Paul Preusse.

KEYNOTE ADDRESS

The keynote address was given by Rear Admiral R. P. Cueroni, Superintendent, U. S. Coast Guard Academy and 1986 President of the Association. Rear Admiral Cueroni addressed his remarks to the relationship between the professional community represented by the MTA and operational commanders. The highlights of his remarks were:

- Welcome to Mystic and the U. S. Coast Guard Academy.

- Your presence here is evidence that commanding officers are concerned about people. They, as do I, surely want to select the best people and equip them with the skills needed to do the job. We know the role you play in meeting that objective.

- I know that sometimes it seems that your work is taken for granted. But a moment's reflection shows us that no one enters the service without being screened for fitness and ability through elaborate procedures which you and your predecessors have developed. No one serves without being trained in programs which you develop. I know your work touches the lives of our people and plays an important role in fulfilling our mission.

- Still, today we seem to have fewer and fewer dollars to support your work. In the Coast Guard we are used to doing more with less, so I feel I may have the license to suggest some priorities as we face an increasingly complex world, a declining personnel pool, and shrinking financial resources.

- Communication among those in your field is essential. We can't afford to duplicate research or planning efforts.

- We must consider the practical benefits of every study, every research project. Too much good work remains on the shelf because it wasn't connected to the goals of those in command.

- You have to get out in front of change and help your commanding officers to meet challenges that will develop. Too many times we get stuck making refinements on solutions to which we have a personal commitment.

- Accept communication with the line commanders in your service as a challenge. When specialists and decision makers work well together the combination is tough to beat.

- As Superintendent of the U. S. Coast Guard Academy you must know that I am committed to education and training. As a commanding officer, I am passionately concerned about practical results.

- Work at your meeting--listen to each other, challenge each other's ideas, teach each other. If you each take home one idea that will "make a difference", your boss's money will have been well spent.

AN UPDATE ON THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)[1]

Deirdre J. Knapp and Rebecca M. Pliske
U.S. Army Research Institute for the
Behavioral and Social Sciences

The U.S. Army implemented the Computerized Adaptive Screening Test (CAST) in 1984. CAST was designed to predict performance on the Armed Forces Qualification Test (AFQT) composite of the Armed Services Vocational Aptitude Battery (ASVAB). It is used by recruiters to prequalify prospective Army recruits. The purpose of this paper is to discuss some of the operational concerns involved in a large scale application of computerized adaptive testing. After presenting background information on the development and validation of CAST, we discuss operational concerns regarding test interpretation, testing environment, and computer hardware.

Background

The Enlistment Screening Test (EST) is a traditional paper-and-pencil test that requires a maximum administration time of 45 minutes, as well as hand-scoring and hand-conversion to the AFQT metric (Mathews & Ree, 1982). Given the time constraints imposed upon Army recruiters, there was a need to develop a screening test that would be quicker and easier to administer. Accordingly, the Army funded the Navy Personnel Research and Development Center to construct CAST using test items that had been calibrated in work related to the development a computerized adaptive version of ASVAB.

CAST currently consists of 78 word knowledge (WK) and 225 arithmetic reasoning (AR) multiple-choice test items. The

---

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.

items were calibrated using a 3-paramenter logistic ogive item response model. CAST uses an Owens-Bayesian theta estimation procedure and its stopping rule is 10 WK and 5 AR items. The CAST score is a weighted combination of the final WK and AR theta estimates that results in an estimate of the examinee's AFQT percentile score.

The vehicle used to administer CAST is known as the Joint Optical Information Network (JOIN). Each JOIN system consists of a Z-80 microprocessor, 2 logically distinct keyboards, a modem, a video disk player, a color monitor, and a dot matrix printer. Each Army recruiting station (of which there are over 2,000) has at least one JOIN system on site.

Relationship to AFQT Scores

There are three validation efforts associated with CAST. The initial validation study was conducted at the Los Angeles Military Entrance Processing Station (MEPS) with a sample of 312 Army applicants (Sands & Gade, 1983). The correlation between optimally weighted CAST subtest scores and AFQT scores was .85. In the second data collection effort, Army recruiters in the Midwest recorded CAST scores and forwarded these data to the Army Research Institute (ARI) for analysis (Pliske, Gade, & Johnson, 1984). ARI researchers obtained the AFQT scores of applicants for whom they had CAST scores by matching social security numbers with computerized records maintained by MEPS. The resulting cross-validation estimate was .80 (n=1,962).

The most recent evidence of CAST's validity was based on data collected from a national sample of 60 Army recruiting stations during January through December 1985 (Knapp & Pliske, 1986). CAST data were collected via experimental data collection software that recorded item level information (e.g., item identification numbers and successive theta estimates) onto diskettes that were forwarded to ARI for analysis. The simple bivariate correlation between CAST scores and AFQT scores was comparable to those obtained in the earlier studies (r=.79; n=5,929). The correlation corrected for restriction in range was .83.

Operational Concerns

Test Interpretation

There are at least three factors that bear on the issue of inferences based on CAST performance. The first factor is related to the lack of an Army-wide policy regarding action to be taken given a prospect's performance on CAST. In some recruiting battalions specific guidelines are given to recruiters, in other battalions prescreening is under the individual recruiter's control. In addition to CAST performance, considerations such as distance to the nearest ASVAB testing location and the current status of the recruiter's quota of accessions will be used to determine the action to be taken with a prospect. The second factor is related to the psychometric naivete of the test interpreters (i.e., the recruiters). Despite efforts to train them to the contrary, recruiters tend to interpret CAST performance at face value. For example, they are likely to predict that an examinee receiving a predicted AFQT percentile score of 51 will subsequently qualify for options only available to those who score at least 50 on AFQT, whereas they will predict that an examinee receiving a predicted percentile score of 49 will not. Third, Army recruiters use CAST scores to predict AFQT categories rather than AFQT scores. At the present time, they are primarily interested in the cutpoint between AFQT categories 3B and 4A (31st percentile) and the cutpoint between AFQT categories 3A and 3B (50th percentile).

Since CAST was implemented in 1984, the test results have been displayed in the form of a bar graph representing the examinee's predicted AFQT percentile score. Obviously, such a display does not address the considerations discussed above. Accordingly, we have proposed two alternative approaches to the display of CAST performance information. The first alternative still provides the recruiter with a predicted AFQT score, but it also provides information about CAST's prediction error. Specifically, the point prediction is depicted on a line that represents the AFQT percentile score continuum. A shaded area that encompasses 21 points on either side of the point estimate shows where 90% of the examinees receiving that score are likely to fall. The area in which 68% of those examinees will likely fall is also shown. The accompanying text explains that the subsequent

performance on AFQT of most examinees performing at this level on CAST will be clustered about the point estimate and that fewer cases will be found as one gets further from that estimate. The second alternative to the display of CAST results is the presentation of the predicted probability of performing within certain ranges of AFQT percentile scores (i.e., below 21, 22-30, 31-49, and 50 or above). These probabilities would be based on validation data collected by Knapp and Pliske (1986).

The Army Recruiting Command is currently considering the two suggested alternative display screens. They are canvassing field recruiters to determine the alternative that seems to be the most useful. More than likely, recruiters will find both alternatives less appealing than a simple point prediction because their decisions will be less clear-cut. These decisions will, however, be better informed.

Testing Environment

The CAST administration environment is not ideal. Although recruiters are told to provide examinees with a quiet setting in which to take the test, there is no check to see that this is done in all 2,000 recruiting stations. The recruiters are also told that examinees are to receive all the time they need to complete the test, but again, there is no way to verify that this is always done. Thus there is a lack of standardization to CAST testing procedures.

Test security also becomes a problem when testing occurs on an indefinite basis at so many locations. Fortunately, CAST's current status as an informal screening device means that it is rarely worth one's while to cheat on the test. For example, an examinee who scores very low on CAST will most likely be allowed to take the ASVAB if he insists. Further, recruiters are generally evaluated on the number of people whom they access rather than on the number of people whom they send for ASVAB testing. Therefore, the problem of test security has been one which we have not fully addressed.

Given the lack of standardization of test administration procedures and minor concerns about test security, it is very important that the validity of the test be estimated under operational conditions. Accordingly, the two CAST cross-validation efforts have been conducted using data

collected from recruiting stations. Thus any decrease in CAST's validity due to problems associated with suboptimal testing environments is presumably reflected in these cross-validation estimates.

Computer Hardware

One of the problems associated with large-scale computerized testing is related to hardware maintenance. The frequency with which recruiters use their JOIN computer systems and their general unfamiliarity with computer hardware necessitates that the computers be reliable. To the extent that the equipment does require repair and regular maintenance, this must be accomplished with minimal delay. The Army has addressed this problem by awarding a JOIN repair/maintenance contract that requires the contractor to respond within 24 hours of the request for service.

Another problem associated with the use of computerized testing arises when examinees are unfamiliar with, and possibly intimidated by, the computer equipment. Although this situation is less likely to occur as computers become more commonplace, at present it cannot be ignored. As a result of this concern, the JOIN system was designed to incorporate two keyboards. The main keyboard closely resembles conventional typewriter keys. The recruiter uses this keyboard to control the software. After the recruiter boots the CAST software, however, control is switched to a much smaller keypad. This detachable keypad has 5 blue keys that are labeled A-F; a green key labeled "GO," a red key labeled "ERASE, and black keys numbered 1-9. The examinee is given this simplified keypad and proceeds to take the test. Once the test is completed, the recruiter enters a code that switches control back to the main keyboard.

Allowing examinees to respond to test questions using a nonthreatening and easily understood keypad alleviates the apprehension that a novice computer user may bring to the testing situation. With respect to examinees who may be highly computer literate, this procedure insures that they do not gain access to the computer software.

Conclusion

It should be clear that many of the problems associated with the implementation of CAST are also applicable to EST. Specifically, those concerns that are not related to the computerized nature of the test are shared by EST. With respect to the steps that can be taken to improve the interpretation of test performance, the computer provides a much better vehicle for improvements than can be obtained using traditional means (e.g., hand conversion charts). A contract effort is underway to incorporate major improvements into CAST. CAST II will take advantage of improved item pools and modifications that are based on research that was not available when the original CAST was developed.

References

Knapp, D.J., & Pliske, R.M. (1986). Preliminary Report on a National Cross-validation of the Computerized Adaptive Screening Test (CAST) (ARI Research Report No. 1430). Alexandria, VA.: U.S. Army Research Institute for the Behavioral and Social Sciences.

Mathews, J.J. & Ree, M.J. (1982). Enlistment Screening Test Forms 81a and 81b: Development and calibration (ARHRL Report No. 81-54). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

Pliske, R.M., Gade, P.A., & Johnson, R.M. (1984). Cross-validation of the Computerized Adaptive Screening Test (CAST) (ARI Research Report No. 1372). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

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

PERSONNEL ASSESSMENT AND TECHNOLOGICAL ADVANCES

Friedrich W. Steege
Federal Ministry of Defense, Bonn
Federal Republic of Germany

INTRODUCTION

In this presentation I will provide recent findings on how computer assisted and adaptive testing (CAT) performed by the Psychological Service of the German Federal Armed Forces can support personnel assessment and counseling in a meaningful way. CAT is regarded as one element in a system of measures of personnel psychology, by which I mean all applications of psychological knowledge and expertise to the manpower and personnel selection, classification and training arena. It will primarily be applied in assessing individual ability and achievement, and in special cases, personality characteristics. It is the responsibility of the psychologist to interprete the data gathered with help of the computer, and to decide on measures or treatments to be taken.

CAT is applied to the classical fields of personnel assessment in the GFAF, i.e. the selection of volunteers including officer candidates, the placement of draftees, and the selection of specialists.

The crucial point in the future development will be to more explicitly orient these psychological measures towards the career of the soldiers, that is, using these measures as an integral part of a dynamic personnel system which will include counseling interactions. Additionally, technological advances (e.g. the laser-disc) will without doubt open wider opportunities in this field.

EXPECTATIONS FROM CAT BY THE PERSONNEL MANAGEMENT OF THE GFAF

The development of computer assisted test applications coincides with a phase in which the Armed Forces of the Federal Republic, like those of many other countries, face major difficulties with respect to the recruitment of personnel in the 90s. The decreasing personnel resources in the years to come require a more differentiated use of the manpower pool for military personnel management. Primarily, it will be necessary to identify what performance level will minimally suffice to recommend the prospective soldiers for a military training in a special occupational area. Accordingly, this situation has produced considerable pressure and great expectations from new CAT developments which are anticipated to deliver better data than currently achieved with paper and pencil tests.

Compared with the cost incurred by the development and subsequent implementation of CAT installations, benefits are expected in the following areas:

- Improved psychometric properties, i.e. more differentiated data and inferences;
- Organizational facilitations, i.e. more flexibility to execute personnel-oriented psychological measures;

- Improvements regarding attractiveness and acceptance of of the GFAF through a positive initial experience.

Psychometric Improvements

Primarily, the following improvements are expected using CAT:
- A greater multitude and differentiation of test items
- Test items with a wider range of difficulty, particularly with regard to testees showing exceptionally good or poor test results;
- A test and test-item sequence that is tailored to the individual testee (sequential and/or adaptive testing); and
- A faster and more differentiated scoring of test results and placement of soldiers of all categories.

Organizational Facilitation

It is also expected that by applying CAT networks it will be possible to design the induction process in a way that every draftee or volunteer is administered the test at the point in time when he is available in the course of the selection or muster examination. Group sessions will no longer be necessary. This is expected to provide the selection and placement centers with more flexibility. It is also expected that the time needed for the procedure may decrease for certain groups of applicants.

Acceptance of CAT

Finally, it is expected that future personnel-oriented psychological measures will be designed in a way that they better the initial experience of the citizen soldier. This would be a contribution to the political aim to make the future soldier adequately aware of the meaningfulness of his military service. Computer-aided test applications are regarded as one means to fulfill that goal.

The total thrust of these expectations has directed us toward the objective of applying CAT routinely to the selection of all volunteers and draftees by 1988.

STATUS OF CAT DEVELOPMENT IN NOVEMBER 1986

Here, I will briefly summarize the current status of the CAT development in the GFAF. Again, the aspects of psychometrics, organization, attractiveness and acceptance will be addressed.

Psychometric Developments

First, let us look at some psychometric characteristics of computer assisted and adaptive testing. Presently, we are applying on CAT terminals standard batteries including apparatus type tests. Adaptive applications are also being tried.

In this connection the following problem areas or development lines are of particular importance in my opinion:

- Optimism with respect to practical applications of adaptive testing is nourished by experiences reported by international experts. One of the first test batteries on adaptive basis in operational use is the Computerized Adaptive Screening Test (CAST) of the U.S. Forces (Knapp et al. 1985; Hakel 1986). In this context positive results regarding the reliable parallel measurement of tests based on classical methods vs. those based on parameter-free construction rationals have to be mentioned (See among others Martin, McBride & Weiss 1982; Sympson 1984). Further, reports on the validity of computerized adaptive test applications have to be pointed to.

- The two last points both lead to the general finding that CAT helps reduce measurement errors, and that the quality of the measurement of abilities and achievements will be increased. Results of this kind are among others reported by Weiss (1985). See also Wottawa (1986).

- Of special importance for the CAT implementation are meaningful criteria for choosing the type of item-response-theory (IRT) or parameter estimation model (among others Hambleton, Ed., 1983; Reckase 1985). There is a wide-spread criticism of the Rasch model, especially in the U.S. (see Wood 1984; Sympson 1984). On the other hand, the practicality of the Rasch model or the theoretical advantages of it as well as of its extensions are emphasized (Hornke & Habon 1985).

- It has been emphasized as a special advantage of an extension of the Rasch model, the LLM (linear logistic model; see Bejar 1983), that it alleviates the rule oriented construction of test items, thus assuring that these items conform with personality theory. This advantage has been pragmatically demonstrated by the research work funded by the Federal Ministry of Defense (Hornke & Habon 1984; 1985). A first crossvalidation of parameter estimations does support this line of research.

Technical Equipment and Organizational Trials

I will now briefly describe areas in which personnel assessment is presently being supported by computers.

The following practical applications already exist today:
- Optical marker reading of answer sheets
- Scoring of individual test data
- Databanks for conscripts, volunteers, and officer candidates
- Scientific subroutine packages on mainframe computers and on personal computers (IBM AT or compatibles).

The following experimental or pilot applications are designed for use in personnel assessment in the GFAF:
- Computerized Testing Stations/Terminals, Versions Ia/Ib (outdated)
- Computerized Testing Stations/Terminals, Version II
  * Host: IBM AT 02 with 20 MB harddisk
  * Back-up and administrative PC: IBM AT 02 with 20 MB harddisk
  * Local area network (LAN) for up to 15 terminals
  * Each testing station (terminal): 20 MB harddisk, white/black screen of 768 x 1024 pixels, image frequency 70 cycles, headphones for voice output, special keyboard

- CAT Center Federal Armed Forces Office, Bonn: 1 central unit, 1 terminal, 1 tape drive, special software for item generation
- CAT experimental station, Volunteer Selection Center, Munich
- Linking of testing unit to databanks or other information systems via intercenter connections.

We started our field trials with our first CAT versions in 1982 (see Wildgrube 1986). Using the advanced CAT II system, trials are presently performed in two recruitment centers (in Munich and in Hildesheim, near Hannover) to determine whether CAT procedures will facilitate conduct of both medical examination and psychological testing of conscripts at a single location on a single day.

Acceptance of CAT

To be user friendly is a key aspect and a high priority of our CAT development. In this regard, we are paying particular attention to:
- Design of the testee keyboard,
- Quality of the graphic representation,
- Extensive use of additional digitalized speach partitions, offered parallel to the written instructions on the terminal screen, and
- Menue design of the whole testing system, written in "C".

These design features enhance the psychometric quality as well as they contribute to the wider acceptance of CAT as part of the personnel-oriented measures of the GFAF.

In addition to the budgetary advantage of a single visit to the recruitment centers by conscripts, the trial "Medical Examination and Psychological Assessment at the same day" is expected to yield an even greater benefit. Because of this streamlined procedure, each conscript will be provided information at the end of that day which heretofore had not been available as rapidly. He will be told to which military unit he will be assigned, as well as its location. Perhaps more importantly, he will have the opportunity to get - at least a brief - counseling by a psychologist.

EVALUATION OF THE IMPLEMENTATION AND THE FUTURE DEVELOPMENT OF CAT IN THE GFAF

In summing one can say: Our experimental installation is yielding first results now. Our experiences with respect to supervising the development have tought us to organize the management for the design of the operational versions (CAT III) differently in future. We are considering contracting a division of knowledge-based systems from a private firm and a larger producer of hardware. We are also considering involvement of a scientific group for the software possibly to include the Armed Forces and civilian universities.

Criteria for evaluation of future applications are again technical (i.e. primarily psychometric) quality, psychological acceptability, and cost-effectiveness considerations, such as requirements of the employing organization.

Evaluation of the technical quality

Criteria of the technical quality are (see also Green et al. 1982):
- Ability to construct test items
- Development of adequate adaptive algorithms
- Adequate analysis of decision procedures.

Trials already performed confirm in general that CAT will render more differentiated data about the testee in shorter time.

It has to be reemphasized here that CAT is part of a system. CAT is in technical perspective not more (but not less) than a means or a tool for diagnostic classification purposes. Basic research with respect to psychometric quality has to be supplemented urgently. This extends beyond our organization and country.

Evaluation of psychological acceptability

Using the term "psychological acceptability", we are referring to those ethical standards of psychology that especially regard the rights of the test takers. An essential element of psychological acceptability in this understanding is that the assessment of any individual should include counseling. Testing in itself would be too narrow.

Wottawa (1986) emphasizes that the application of electronic data processing technology in psychological assessment be acceptable in principle, provided that an increase in reliability and efficiency without loss of validity would be guaranteed, and it would be unlikely that additional mistakes occur. For the preparation of test results (e.g. for interpretation), data processing is undoubtedly very useful. However, it should not lead to an impairment of the special relationship of psychologist and testee. For the diagnosis data processing should remain a tool, it never must become an end in itself:

> "The idea of an inhuman, fully automated "testing line" leaving the testee no chance to get an individual treatment and a personal interview with experienced examinators or psychologists might become reality from a technological perspective by use of highly sophisticated computer hard- and software. It would, however, never meet technical requirements and ethical standards of Psychologists. It has to be stated, however, that the present application of computer aids in the GFAF by no means indicate a negative development in that direction" (Wottawa 1986,p. 49).

In any case, developers of specific diagnostic expert systems have to consider these warnings. We intended to make clear that computer aided diagnostic will ever remain a tool.

Evaluation of employer concerns

Economic considerations and concerns of the employing organization are an essential consideration if maximum utilization is to occur. It is critical to search for new organizational models that are cost effective but nevertheless sufficient in view of the criteria stated earlier. This is the predominant aim of our psychological policy in connection with the development and implementation of CAT-installations. It, therefore, requires con-

stant coordination and "reality testing" with the organizations to be served.

CONCLUSION

The need for development of computerized adaptive testing for selection, classification and placement of incoming service members is obvious. This is true both from a psychologists effort to develop a better process and from the management or organizational perspective to work under resource constraints. The impact of a full blown, reliable CAT program will be felt throughout any organization in which it is instituted. Morale, motivation, and training can all benefit and in the long run, the contribution to efficiency will surely be worth the cost in money and perspiration.

REFERENCES

Bejar, I.I. (1983): Introduction to item response models and their assumptions. In: HAMBLETON, R.K., Ed., 1 - 23

Embretson, S., Ed.(1984): Test design: Contributions from psychology, education, and psychometrics. New York, Academic Press

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., Reckase, M.D. (1982): Evaluation plan for the computerized adaptive Vocational Aptitude Battery. Baltimore: Johns Hopkins University (Research Report 82-1)

Hakel, M.D. (1986): Personnel selection and placement. In: Annual Review of Psychology 37, 351-380

Hambleton, R.K., Ed. (1983): Applications of item response theory. Vancouver, Educational Research Institute of British Columbia

Hornke, L.F. & Habon, M.W. (1984) : Regelgeleitete Konstruktion und Evaluation von nicht-verbalen Denkaufgaben. Bonn: Bundesministerium der Verteidigung (Wehrpsychologische Untersuchungen Nr. 4/84)

Hornke, L.F. & Habon, M.W. (1985): Kognitive Lösungsoperationen bei regelgeleitet konstruierten figuralen Denkaufgaben. Bonn: Bundesministerium der Verteidigung (Arbeitsbericht des PsychDstBw Nr. AL-4-85)

Knapp, D.J., Pliske, R.M. & Johnson, R.M. (1985): An introduction to the computerized adaptive screening test (CAST). Paper presented at the APA Convention. Los Angeles

Martin, J.T., McBride, J.R. & Weiss, D.J. (1983): Reliability and validity of adaptive and conventional tests in a military recruit population. Minneapolis: University of Minnesota (Research Report 83-1)

Reckase, M.D. (1985): Trait estimates from multidimensional items. Paper presented at the APA Convention. Los Angeles

Sympson, J.B. (1984): Review of Hambleton (Ed.). In: Applied Psychological Measurement, 8, 467-470

Weiss, D.J. (1985): Final report. Computerized adaptive measurement of achievement and ability. Minneapolis: University of Minnesota (ONR-Report)

Wildgrube, W. (1986): News about CAT in the German Federal Armed Forces (GFAF). In: NPRDC (Ed.), Proceedings of the 27th Annual Conference of the Military Testing Association. San Diego

Wood, R. (1984): Review of Weiss, D.J. (Ed.). In: Applied Psychological Measurement, 8, 463-465

Wottawa, H. (1986): Fachliche Mindestvoraussetzungen für den Einsatz computerstützter Testanlagen im Rahmen wehrpsychologischer Diagnostik. Bonn: BMVg/FMOD (Arbeitsbericht des PsychDstBw Nr. PE-1-86)

RELATIVE PRECISION OF PAPER-AND-PENCIL AND COMPUTERIZED ADAPTIVE TESTS

Rebecca D. Hetter and Daniel O. Segall†

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

The Navy Personnel Research and Development Center is conducting research to design and evaluate a computerized adaptive test (CAT) as a potential replacement for the paper-and-pencil Armed Services Vocational Aptitude Battery (P&P-ASVAB). In support of this effort, the Accelerated CAT-ASVAB Program (ACAP) is evaluating item pools specifically developed for computerized adaptive testing. The objective of this research is to compare the precision of CAT-ASVAB ability estimates with their P&P-ASVAB counterparts.

METHOD

The work described here compared simulated administrations of a 10-item adaptive test and a 15-item conventional test. The comparison was based on score information functions and test reliabilities.

Item Pool

The item pool consisted of paragraph comprehension (PC) items. The conventional test consisted of the 15 items that make up the PC subtest of P&P-ASVAB, Form 9A. For the adaptive test, the items were selected from a PC pool specifically developed for research in support of CAT-ASVAB by Prestwood, Vale, Massey & Welsh (1985). Using a joint calibration approach, Prestwood *et al.* obtained item response theory parameter estimates -- based on the three-parameter logistic model (3PL) (Birnbaum, 1968) -- for all items. These estimates were used in the present simulations.

Exposure Control

The CAT-ASVAB uses item information as a basis for selecting items during the adaptive test. To avoid overexposing certain highly informative items, the system incorporates an *exposure control* algorithm that operates in conjunction with maximum information item selection (Sympson & Hetter, 1985). This algorithm reduces the exposure rate of certain highly informative items, while increasing the exposure rate for other items. The result is an upper ceiling on item exposure.

For this study, the CAT items were divided into two alternate forms (Moreno, 1986). Form 1 consisted of 85 items and Form 2 of 83 items. Exposure control parameters were computed for items within each form. The target exposure rate was set to 1/3. This results in an exposure rate of 1/6 across the two CAT-ASVAB forms. An exposure rate of 1/6 is

† The opinions expressed here are those of the authors and do not necessarily represent those of the Department of the Navy.

comparable to the six forms used in P&P-ASVAB.

Score Information Functions

Score information functions provide one criterion for comparing the precision of the CAT-ASVAB with the P&P-ASVAB. Birnbaum (1968, Section 17.7) defines the information function I for any score y to be

$$I\{\theta, y\} \equiv \frac{\left[\frac{d}{d\theta}\mu_{y|\theta}\right]^2}{V(y|\theta)}. \quad (1)$$

where θ represents ability, μ is the conditional mean of y given θ and V is the conditional variance. This function is, by definition, inversely proportional to the square of the length of the asymptotic confidence interval for estimating ability θ from score y. Three information functions were computed and compared for this study: one from the P&P-ASVAB (Form 9A), and one from each of the two CAT-ASVAB forms. The test with greater information at a given ability level will have a smaller confidence interval for estimating θ.

CAT-ASVAB Score Information Functions

The score information function for the CAT was approximated from simulated adaptive test sessions.

Adaptive Tests. The simulated tests were 10 items in length. The sessions were repeated independently for 500 examinees at each of 31 different θ levels (equally spaced along the [−3,+3] interval). These θ levels are assumed to be true abilities for CAT simulations.

Owens' Bayesian scoring (Owen, 1975) was used throughout the test to update the ability estimate. Items were selected from information tables on the basis of maximum information, in conjunction with the exposure control algorithm. (An information table consists of lists of items by ability level. Within each list, all the items in the pool are arranged in descending order of the values of their information functions computed at that ability level. This study used 37 ability levels equally spaced along the [-2.25 to +2.25] interval).

To simulate examinee responses, a pseudo-random number was drawn from a uniform distribution in the interval (0,1). If the random number was less than the 3PL probability of a correct response, the item was scored correct; otherwise it was scored incorrect. Prestwood *et al.* parameter estimates and true ability were used to generate and/or score responses.

Score Information. At each true θ level, the mean m and variance s^2 of the 500 final scores ($\hat{\theta}$) were computed. The information function I at each selected level of θ can be approximated from these results, using the formula (Lord, 1980, eq. 10-7):

$$I\{\theta, \hat{\theta}\} \approx \frac{[m(\hat{\theta}|\theta_{+1}) - m(\hat{\theta}|\theta_{-1})]^2}{(\theta_{+1} - \theta_{-1})^2 s^2(\hat{\theta}|\theta_o)}, \quad (2)$$

where θ_{-1}, θ_o, θ_{+1} represent the successive levels of θ, and $\hat{\theta}$ represents the Owen's Bayesian estimate. However, the curve produced by this approximation often appears jagged, with many local variations. To reduce this problem, information was approximated by equation (3):

$$I\{\theta, \hat{\theta}\} \approx \frac{\left[\frac{m(\hat{\theta}|\theta_{+1}) + m(\hat{\theta}|\theta_{+2})}{2} - \frac{m(\hat{\theta}|\theta_{-1}) + m(\hat{\theta}|\theta_{-2})}{2}\right]^2}{\left[\frac{\theta_{+1} + \theta_{+2}}{2} - \frac{\theta_{-1} + \theta_{-2}}{2}\right]^2 \left[\frac{1}{5}\sum_{k=-2}^{+2} s(\hat{\theta}|\theta_k)\right]^2}$$

$$= \frac{25\,[m(\hat{\theta}|\theta_{+2}) + m(\hat{\theta}|\theta_{+1}) - m(\hat{\theta}|\theta_{-1}) - m(\hat{\theta}|\theta_{-2})]^2}{(\theta_{+2} + \theta_{+1} - \theta_{-1} - \theta_{-2})^2\,[\sum_{k=-2}^{+2} s(\hat{\theta}|\theta_k)]^2}\,, \qquad (3)$$

where θ_{-2}, θ_{-1}, θ_o, θ_{+1}, θ_{+2} represent successive levels of θ. This approximation results in a moderately smoothed curve with small local differences.

P&P-ASVAB Score Information Functions

The P&P-score information function for a number-right score x was computed by (Lord, 1980, eq. 5-13):

$$I\{\theta, x\} = \frac{\left[\sum_{i=1}^{n} P'_i(\theta)\right]^2}{\sum_{i=1}^{n} P_i(\theta)Q_i(\theta)}\,. \qquad (4)$$

where P_i is the 3PL probability of a correct response, $Q_i = 1-P_i$, and P'_i is the first derivative of P_i. This function was computed by substituting Prestwood *et al.* estimated P&P-ASVAB (Form 9A) parameters for those assumed to be known in (4).

Test Reliabilities

Test reliabilities provided the second criterion for comparing the CAT-ASVAB with the P&P-ASVAB. Lord & Novick (1968) define the *reliability of a test* as the squared correlation between observed score and true score. Since in simulation work the true score (ability) is assumed to be known, this correlation can be computed directly. In real life testing, true scores are unknown and reliability must be estimated. Among the various methods for estimating reliability is the *test-retest method*, where the same test is administered to each person twice. The correlation between the two scores is then taken as an approximation of the reliability.

Adaptive Tests. The tests were 10 items in length. They were administered independently at 1900 true ability levels randomly generated from the (0,1) normal distribution. Scoring, item selection, and response simulation were performed as in the score information analyses. Test reliability was computed as the squared correlation between estimated (observed) ability and true ability.

Paper-and-Pencil Tests. The 15 PC items in P&P-ASVAB (Form 9A) were administered twice at the same 1900 simulated true abilities used in the adaptive tests. Responses were simulated as before, and two number-right scores (x_1, x_2) were computed for each simulee. Test reliability was estimated as the correlation between the two number-right scores.

Figure 1: CAT-Form 1 and P&P-Form 9A

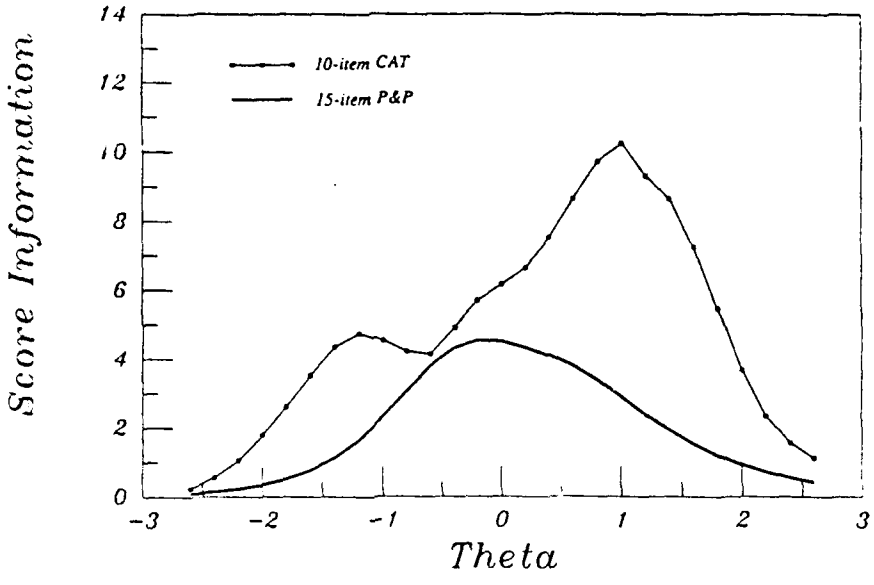
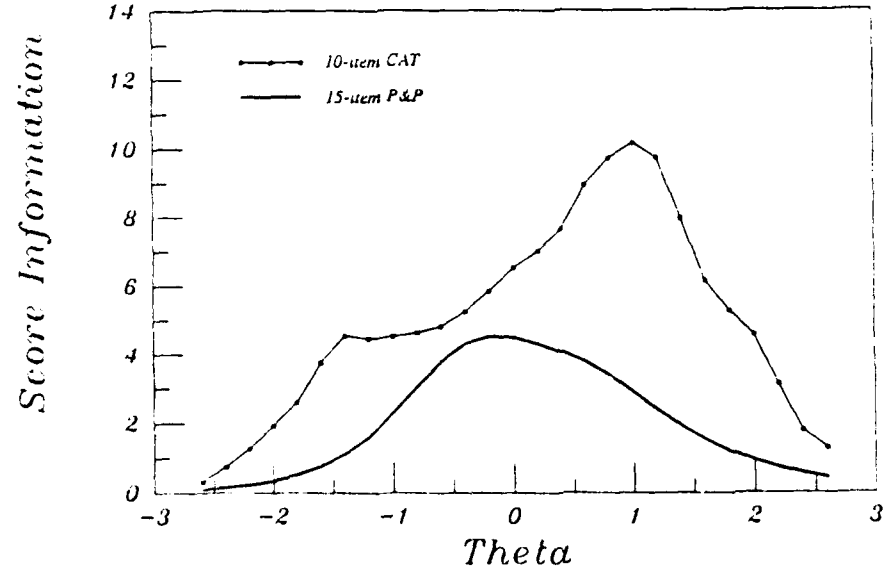


Figure 2: CAT-Form 2 and P&P-Form 9A



RESULTS AND DISCUSSION

Score Information Functions

Figures 1 and 2 present the comparisons between the conventional test and each of the two CAT forms. Note that the CAT tests are shorter (10 items) than the conventional test (15 items). The CAT-ASVAB score information functions indicate higher relative precision over the range of ability examined.

Test Reliabilities

Table 1 presents simulated test reliabilities for the adaptive and the conventional tests. The table shows that the reliabilities for both forms of the CAT are higher (.839 and .834) than the simulated test-retest reliability of the P&P (.759).

TABLE 1: Test Reliabilities

	CAT-Form 1	CAT-Form 2	P&P-Form 9A
$r^2\ (\theta, \hat{\theta})$	.839	.834	
$r\ (x_1, x_2)$			.759

where: r = Pearson correlation coefficient
θ = true ability
$\hat{\theta}$ = estimated ability
x_1, x_2 = test-retest number-right scores

CONCLUSIONS

These analyses indicate that the precision of the 10-item adaptive test is higher than the 15-item conventional test. This result holds across the levels of ability examined.

REFERENCES

Birnbaum, A. (1968). Some latent-trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Moreno, K. E. (1986). *Development of alternate forms for an adaptive test*. (NPRDC-TR in preparation). San Diego, CA: Navy Personnel Research and Development Center.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association, 70,* 351-356.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an adaptive item pool.* (AFHRL-TR-85-19). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sympson, J. B. & Hetter, R. D. (1985). *Controlling item exposure rates in computerized adaptive tests.* Paper presented at the 1985 Military Testing Association Conference, San Diego, CA.

Correlations of Computer-Adaptive Test Time Measures with Thetas and a Criterion Measure

Mark Y. Czarnolewski and Clessen J. Martin[1]
U.S. Army Research Institute
for the Behavioral and Social Sciences

Introduction

Computerization of the Armed Services Vocational Aptitude Battery (ASVAB) may offer an additional psychometric advantage to the usual advantages associated with adaptive testing. One advantage often associated with computer adaptive testing (CAT) for ASVAB over the current standard testing procedure is that CAT is interactive. In CAT, the computer presents a succeeding test item based on whether the preceding item was answered correctly. With each item pre-calibrated in terms of difficulty, discrimination and guessing one may quickly establish the ability of the examinee. Construct validity of CAT ASVAB has been established by factor analyses showing CAT ASVAB subtests loading on the same factors as their paper and pencil counterparts (Martin, Park and Boorom, 1986).

Another advantage of CAT ASVAB is that it allows for precise measurement of response time. Reaction time has long been one of the most popular variables used to investigate psychological processes in the information processing literature (Pachella, 1974), and increasing attention is being given to this variable as an individual difference variable in cognitive psychology. Time spent on each CAT ASVAB subtest may, thus, provide an important additional parameter.

The administrative advantages of CAT ASVAB has resulted in the Department of Defense planning to introduce CAT into the Military Entrance Processing Command (MEPCOM) early in calendar year 1989. Instead of the 10 separately-timed subtests on the conventional paper-and-pencil (P&P) ASVAB, CAT ASVAB has 11 subtests, with the paper and pencil Auto/Shop subtest separated into two adaptive tests in CAT ASVAB. With the exception of Numerical Operations and Coding Speed subtests, the remaining nine CAT ASVAB subtests are self-paced, with total time spent on each of the subtests recorded for each examinee. Time spent on each subtest was the focus of this research.

The purpose of this research was to (a) identify individual differences in the test times for each subtest of CAT ASVAB, (b) determine whether CAT ASVAB test times could be used to interpret individual differences in terms of the processes that test takers may be employing for each subtest and (c) determine whether CAT ASVAB times could provide incremental validity in predicting Advanced Individual Training (AIT) course performance.

---

[1]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. Acknowledgement is extended to Gregory Candell for constructing the data base and to Denise Boorom for data collection.

Method

Subjects

The analyses in this report are based on a total of 335 recruits stationed at either the Army Engineer School or the Defense Mapping School at Fort Belvoir, Virginia. All recruits were enrolled in Advanced Individual Training courses either in 52D (Power Generator Equipment Repairman) or in 81B (Construction Drafting).

Procedure

The CAT ASVAB test was administered on a Local Area Network (LAN) consisting of 8 Apple IIIs. Seven recruits were tested in each session. The remaining Apple III was used as a Test Administrator Station. This testing system was developed at the Navy Personnel Research and Development Center and is fully documented in Hardwick, Eastman, Cooper, and Vicino (1984).

Separate CAT ASVAB subtests were developed on item pools consisting of approximately 200 items. The item pools were developed specifically for CAT ASVAB and do not overlap with items on the existing P&P ASVAB. All items were pretested using applicants from MEPCOM. A total of approximately 2,000 responses were obtained for each item. The item parameters were obtained from a slightly modified program developed by Wood, Wingersky, and Lord (1976) called LOGIST 2b. This version of CAT ASVAB uses Owen's (1969, 1975) sequential Bayesian procedure to estimate ability via a three parameter model. All nine cognitive tests were fixed-length tests of 15 items each.

Results

Table 1 presents descriptive statistics of each subtest for theta and time. These data indicate that most of the theta estimates have a slightly negative skew, while most of the time data indicate a moderate to high positive skew. In short, the ability estimates are indicating more uniform distributions than the time data, with some skewness for theta at the low end of the ability distribution. For purposes of the study, observations more than 3 standard deviations from the mean of these untransformed time data were excluded. One subject was excluded for unusually small number correct (i.e. 3) for Number Operations. Subjects were also excluded if they had recorded times greater than 6 minutes for Number Operations or greater than 11 minutes for Coding Speed. Any times greater than these maximum time limits indicated unusual test administration or recording procedures.

A hierarchical regression procedure was employed to test for a qualitative difference in the relationship between theta and time for different theta levels. Time was the dependent measure and theta the independent variable. The theta estimate was forced first into the regression to test for a significant linear relationship betweeen theta and time, and the square of theta was forced into the regression second to determine if a nonlinear relationship existed beyond the linear relationship tested in the first step (Cohen & Cohen, 1975). Table 2 presents the results of the hierarchical regressions.

Table 2 shows larger Rs between theta and time when time has a

Table 1

Descriptive Statistics of each CAT Subtest for Theta and Time

	Ability Estimate (Theta)			Time on Test (Min.)		
CAT Subtest	Mean	S.D.	Skewness	Mean	S.D.	Skewness
General Science	.524	.574	-.206	6.1	1.6	1.4
Arithmetic Reasoning	.490	.715	-.146	15.9	6.3	1.6
Word Knowledge	.397	.609	-.106	4.9	1.6	1.6
Paragraph Comprehension	.320	.746	-.067	12.8	4.8	1.5
Number Operations[1]	38.8	9.1	-.664	4.3	.4	.6
Coding Speed[2]	48.5	13.3	-.189	8.6	.8	1.1
Auto Information	.257	.763	-.096	6.5	1.8	1.3
Shop Information	.201	.869	-.357	6.2	1.7	1.3
Math Knowledge	.532	.693	-.159	8.3	3.0	1.0
Mechanical Comprehension	.188	.697	-.402	11.1	3.1	.6
Electronics Information	.124	.827	.120	5.2	1.2	.8

[1]Number Correct out of 50 items within 6 minutes.
[2]Number Correct out of 84 items within 11 minutes.

Table 2

Hierarchical Regressions Testing for Qualitative Differences in the Relationship Between Theta and Time for Each CAT ASVAB Subtest

	Untransformed Time				Log Transformed Time			
	Linear Component		Quadratic Component		Linear Component		Quadratic Component	
	R	B	R	B	R	B	R	B
GS	.05	-.16*	.12	.15*	.04	-.14	.10	.12
AR	.22	.16*	.23	.09	.24	.20**	.25	.07
WK	.12	-.11	.12	-.01	.14	-.11	.14	-.05
PC	.10	.19**	.21	-.21***	.20	.32***	.32	-.28***
NO	.26	1.29***	.34	-1.56***	.27	1.34***	.36	-1.62***
CS	.30	-.38	.30	.08	.30	-.34***	.30	.04
AI	.19	.16*	.20	.07	.21	.21***	.22	.01
SI	.10	.11*	.19	-.16**	.14	.15**	.23	-.18***
MK	.51	.46***	.52	.08	.54	.54***	.54	-.00
MC	.50	.50***	.50	.04	.52	.52***	.52	-.00
EI	.06	.14	.24	-.25***	.11	.20***	.31	-.30***

Note. The R for the linear component is for the first step of the hierarchical regression, and the R under the quadratic component is the accumulated R after the second step. The Beta weights (B) for both components were computed at the second step.
The n's vary from 318 to 325. B = Standardized Beta.
Significant levels for B are starred: *$p < .05$ **$p < .01$ ***$p < .001$.
Significant levels for R's are: $R > .12$, $p < .05$; $R \geq .16$, $p < .01$; $R \geq .19$, $p < .001$.

logarithmic transformation than when time is not transformed. Table 2 also shows significant increases in the Rs for the quadratic component of the relationship between theta and time for four subtests: Paragraph Comprehension, Number Operations, Shop Information and Electronics Information. The Beta for the linear component for each these four tests is positive, while the quadratic component is negative. This indicates that, initially, the higher the ability estimate, the longer subjects take on the test. However, after a certain point on the ability continuum there is a shift, with higher ability subjects becoming increasingly faster. The quadratic component of General Science approaches significance, $p < .06$. The significant linear and nonlinear Betas for GS reflect an opposite pattern than found for the other four "nonlinear" ASVAB subtests. Higher ability for GS initially relates to faster times, while after a certain point on the ability continuum, times become slower. The nonlinear component for these five subtests is acting as a suppressor (Cohen & Cohen, 1975) as seen by the Betas for the linear components of these tests being significant (or more significant for NO) once the nonlinear component is in the regression.

Incremental validities were determined for the subtests in the GM and AFQT composites. The GM composite was chosen because it is the selector composite for this engineering MOS and the sample size was sufficient. The AFQT was chosen because it is a measure of trainability. Choosing both composites also allowed for testing the incremental validities of all subtests exhibiting nonlinear relationships between the theta estimate and time statistics. The criterion was the average score from ten tests, each representing a different course module.

Incremental validities were tested in two ways. First, separate hierarchical regressions were performed for each subtest in which the theta estimates were forced in first and the test's time parameters forced in afterwards. Second, separate regressions were run for the GM and AFQT composites. Each set of subtests, (i.e. for the GM and AFQT composites) were forced in first followed by their respective time parameters.

The Electronics Information and Auto Information subtests exhibited significant incremental validities for their respective time parameters. For Electronics Information the multiple R increased significantly, from $R = .39$, $p < .001$ for theta to $R = .46$, $p < .001$ when all three time parameters were in the regression. The partial r's for each time parameter, controlling for theta, were $pr = -.21$, $p < .001$ for the Time Variable, $pr = .14$, $p < .085$ for the linear component of the theta x Time interaction, and $pr = .18$, $p < .021$ for the quadratic component of the theta x Time interaction. Similarly, for Auto Information the R increased from $R = .33$ to $R = .38$, with both Time and the linear component of the theta x Time interaction having $pr = -.16$, $p < .04$.

For the AFQT subtests, Paragraph Comprehension experienced a significant increase in the multiple R from $R = .40$, $p < .001$ to $R = .47$, $p < .001$, with Time having a significant partial correlation, $pr = .21$, $p < .005$.

The second set of hierarchical regressions were performed as follows: For the GM composite, first, simultaneously force in the thetas for Math Knowledge, Electronic Information, Auto Information, Shop Information and the General Science subtests; second, simultaneously force in the times for each

subtest; third, simultaneously force in the linear component of the theta by time interaction; fourth, for the Electronic and Shop Information subtests, force in the quadratic component of the theta by time interaction. A similar model was employed for the subtests comprising the AFQT composite. For Number Operations a z-score transformation was computed based on the raw score of number correct divided by 2. The z-score represented a "theta" for this speeded subtest.

The multiple R for the thetas of the five subtests comprising GM equals .61, $F(5, 150) = 17.47$, $p < .001$. The significant subtests were MK, $pr = .47$, $p < .001$; AI, $pr = .20$, $p < .01$; and SI, $pr = .12$, $p < .07$. The time parameters that were significant or approached significance were the time variable for EI, $pr = -.15$, $p < .08$ and AI, $pr = -.13$, $p < .10$. The linear component for the Time by theta interaction was significant for SI, $pr = -.16$, $p < .05$, and the quadratic component for the time by theta interaction for EI approached significance, $pr = .14$, $p < .09$. The latter quadratic interaction had a Beta = .15, indicating a tendency for longer EI times having higher school averages. Regardless of the regression model employed, the time parameters as a set of predictors did not significantly improve prediction, although their partial r's were encouraging.

The multiple R for the thetas of the four subtests comprising AFQT equals .57, $F(4, 151) = 17.89$, $p < .001$. The significant subtests were AR, $pr = .31$, $p < .001$, PC, $pr = .13$, $p < .06$ and NO, $pr = .14$, $p < .09$. The time parameters that were or approached significance were the time variable for PC, $pr = -.24$, $p \leq .003$; WK, $pr = -.17$, $p \leq .03$. The linear component of Number Operations by Time interaction approached significance, $pr = -.13$, $p < .10$. As a group of predictors, the time variables significantly added 4.8% of explained variance to the prediction of school grades, F change equals 2.79, $p < .03$. Other models did not significantly add to explained variance.

Discussion

A pattern emerges when observing tests with significant linear components in Table 2. The Math Knowledge and Mechanical Comprehension tests, two tests with the strongest linear components, appear to represent structured tasks requiring highly defined skills. Word Knowledge, on the other hand, may represent an unstructured task requiring equal access to information regardless of ability level. Arithmetic Reasoning, whose linear component correlation falls between these high and low "structured" tasks, may be eliciting a structured approach for some test takers and an unstructured, equal access approach for other test takers.

For tests showing significant, negative nonlinear relations between time and theta (or number correct), one may suggest that the items identifying test takers at the lower ability levels do not elicit the processes or skills in the automatized or integrated fashion as those items at the upper ability levels. One explanation is that abilities differ in the novelty and automatization required. (Sternberg, 1984).

One may suggest, for example, that the fast times for low ability items in Paragraph Comprehension may be identifying those subjects who are not fully integrating the various lexical, syntactic and semantic components required for reading. Subjects in the middle of the ability continuum may be

integrating these components but have not automatized them, while subjects a high ability levels have successfully automatized these reading components. One may add the novelty factor of Sternberg's theory to this model describing Paragraph Comprehension performance to explain performance on the Electronics and Shop Information subtests. The significant positive nonlinear relationship between time and theta for GS may be reflecting items whose differences relate to the interaction between the continua of relative structure and automatization.

A taxonomy of ability measures which incorporates (a) this paper's hypothesized continuum of high versus low structure for ASVAB subtests and (b) a distinction between tests showing significant, negative nonlinear theta by time interactions versus those tests with only linear theta by time interactions may have diagnostic and predictive utility. Diagnostic utility would be seen by successful identification of those characteristics of these "nonlinear" tests that allow for faster processing times. Low and moderate ability subjects could be sensitized to those characteristics in order that they improve their performance on negative nonlinear tests. High ability subjects could be sensitized to transfer those skills they employ for solving complex items in negative nonlinear tests to those linear tests in which they solve similarly complex items, with complexity defined by theta. Diagnostic and predictive utility would be seen by increased validity coefficients that result from experimental intervention that teaches the transfer of those information processing skills identified by this taxonomy.

References

Cohen, J., & Cohen, P. (1975). Applied multiple regression/correlational analysis for the behavioral sciences. New York: Lawrence Erlbaum Associates.

Hardwick, S., Eastman, L., Cooper, R. and Vicino, F. (1984). Computerized Adaptive Testing (CAT): A User Manual (NPRDC Tech. Rep. 84-32). San Diego: Navy Personnel Research and Development Center.

Martin, C.J., Park, R.K., and Boorom (1986; April). Validating a computer adaptive testing system using structural analysis. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Chicago, Ill.

Owen, R.J. (1969). A Bayesian approach to tailored testing (Research Bulletin 69-92). Princeton, New Jersey: Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Pachella, R.G. (1974). The interpetation of reaction time in information processing research. In B.H. Kantowitz (Ed.) Human information processing: Tutorials in performance and cognition. Potomac: Md.: Erlbaum.

Sternberg, R.J. (1984). Facets of human intelligence. In J.R. Anderson and S.M. Kosslyn (Eds.) Tutorials in learning and memory: Essays in honor of Gordon Bower. New York: Freeman.

Wood, R.L., Wingersky, M.S., and Lord, F.M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum No. 76-6). Princeton, NJ: Educational Testing Service.

Using Focus Groups in Military Manpower Research[1]

Rebecca M. Pliske and Timothy W. Elig

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

Focus groups can provide military manpower and personnel researchers a useful method for collecting qualitative data. Qualitative data are non-numerical in nature and are not subject to statistical analysis. In this paper, we will briefly describe the focus group method and discuss its strengths and limitations. We will illustrate its usefulness by describing two current projects at the U.S. Army Research Institute for the Social and Behavioral Sciences (ARI) which have used this method.

Focus Group Method

Focus groups are composed of approximately 8 to 10 individuals who are brought together to discuss a particular topic. A moderator is present to keep the discussion "focussed" on the topic. The success of the focus group depends on the dynamics of the group. That is, participants should feel free to express their opinions and to react to the opinions expressed by others. The size of the focus group is critical. Small groups (fewer than 8 participants) are often unable to promote free discussion because the participants feel too vulnerable. On the other hand, some participants in large groups (more than 10 participants) may "hide" and never express their feelings. Furthermore, some large groups may get out of control because participants will initiate private conversations among themselves that are irrelevant to the topic of interest.

The moderator structures the focus group discussion by giving an introduction (called the script) that presents the topic for discussion. The moderator also provides some guidance on what is expected from participants and gives assurances of anonymity when appropriate. Participants may need such assurance when video and/or audio recording equipment is used to document the session. The moderator has a topic guide (i.e., an outline) that lists the subjects to be covered during the session. These subjects are often used as probes to stimulate additional discussion relevant to the topic of interest.

For a successful focus group, the moderator should be knowledgeable about small group dynamics and possess both verbal and nonverbal communication skills. He or she must encourage active participation of all group members while remaining unintrusive. It is often necessary to redirect the discussion when it digresses from the topic of interest and to control participants who want to dominate the discussion. The success of the focus group is largely dependent upon the moderator's ability to maintain control over the group while simultaneously stimulating free discussion.

---

[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.

Strengths and Limitations of the Focus Group Method

A discussion of the strengths and limitations of the focus group method must take into consideration the purpose for which the data is being collected. Calder (1977) distinguishes between the following three approaches to collecting qualitative data from focus groups: exploratory, clinical, and phenomenological. Each approach has a different purpose and is subject to different limitations.

The exploratory approach characterizes focus groups that are conducted to provide provisional information. This approach allows researchers to generate theoretical ideas or research hypotheses that can subsequently be verified with large scale, quantitative research. For example, researchers sometimes use focus groups to pilot test instruments designed for a large sample survey. Exploratory focus groups can provide researchers with valuable information with a limited investment of time and money. The weakness of this approach is the lack of generalizability of the results. When subsequent quantitative research with a statistically representative sample is never completed, exploratory focus groups can not be interpreted as providing conclusive information.

In contrast, the clinical approach refers to researchers who conduct focus group research as an end in itself. Focus groups provide them with the qualitative information they need to understand the issue in question. This approach assumes that the real causes of behavior are best understood through the clinical judgment of a qualified observer. The moderator may use a variety of qualitative techniques (e.g., projective tests) to uncover the underlying motivations of the participants. The success of a clinical focus group depends on the accuracy of the clinical judgment of the moderator. "To the extent that the process of clinical judgment fails, the clinical approach results in everyday knowledge which masquerades as scientific. Therefore, at its best, the clinical approach yields quasiscientific knowledge; at its worst, it yields phony scientific knowledge (Calder, p. 358)."

The purpose of the phenomenological approach to focus group research is to bridge the social gap between the researcher and the group of interest. For example, marketing researchers may want to "experience" groups of consumers discussing their product because they realize that their perception of the product may be quite different from the consumers' perceptions. In contrast to the exploratory and clinical approaches that attempt to obtain prescientific or quasiscientific knowledge, the purpose of the phenomenological approach is to obtain "every day" knowledge about the attitudes of the the group participants. No attempt is made to generalize this type of knowledge. The role of the moderator is somewhat different for this approach because he or she needs to be more actively involved in the discussion to share the experience of the other group members. This approach to focus groups is particularly useful to researchers who may be out of touch with the group of people they are studying.

Although Calder's distinction between the three approaches to collecting qualitative data in focus groups is useful in evaluating the strengths and weaknesses of the focus group methodology, in practice many focus groups involve a combination of the approaches. The focus groups conducted for two projects described below involved a combination of the exploratory and phenomenological approaches. Each of these projects will now be briefly described.

Example 1: Modeling the Army Choice (MTAC)

The objective of the MTAC project is to develop new quantitative instruments to measure the factors involved in the career decision making process of prospective Army recruits. The project was designed as a three phase effort. In the first phase, new instruments were developed and pilot tested. The second phase involves a nation-wide data collection to validate the new instruments. If the instruments prove to be predictive of enlistment behavior, then they will be adapted for use as a decision aid during the third phase of the project.

To obtain a better understanding of the career decision making process of young adults, five focus groups were conducted with 17-20 year old high school seniors and high school graduates during the first phase of the project. The method and results of these focus groups are described in detail in McTeigue, Kralj, Adelman, Zirk, and Wilson (1986). Participants for the focus groups were recruited by local interviewing/marketing firms in five metropolitan areas around the country. The following topics were covered during the focus group sessions: current occupational status of each participant, types of careers they have considered, factors that are important to them related to their careers, sources of information/influence about careers, determinants of participants' career choice, positive and negative perceptions of the Army as a place to work, sources of information about the Army, comparison of the Army to other services, comparison of civilian life with Army life, and participants' reactions to their contact with recruiters. The focus group sessions were recorded on both video and audio tapes.

Two focus groups were conducted during the first few months of the project. The primary purpose of these groups was for the research team to get in touch with our "consumers." However, we had also formulated some tentative theoretical models of the enlistment decision process for which we were seeking confirmatory evidence. Thus, these groups could be characterized as both exploratory and phenomenological. The results of the first two groups confirmed our expectations and also provided new insights into the career decision process of young adults.

Based on a thorough search of the research literature on decision making and the results of the first two focus groups, we constructed quantitative instruments based on an expectancy theory model. These instruments were administered to participants in the three additional focus groups prior to their open discussion of the topics listed above. Participants were also asked to provide feedback on the instruments they had completed.

The qualitative data collected in the focus groups for the MTAC project have proved to be quite useful. Observing young men and women discuss their own career decision making process provided our research team with valuable insights. We were able to obtain confirmation for our preliminary theoretical constructs. We also gained a better understanding of our consumer. For example, we were somewhat surprised at their perception of time. Two years (the minimum Army enlistment) is perceived as an extremely long commitment to these young people. The focus groups also provided us with an efficient means for pilot testing our quantitative instruments. Participants were able to give us feedback that we have used to revise the instruments.

Example 2: Hometown Alumni Recruiting Program (HARP)

The HARP project was conducted to assess the feasibility of developing a recruiter assistance program using Army veterans. A detailed summary of this project is provided by Wilson, Celeste, Pliske, Elig, and Ramsey (1986). The HARP concept grew out of the 1985 Army Experience Survey (AES) that was administered to recently separated Army veterans (Westat, 1986). Preliminary results from the AES indicated that Army veterans would be willing to assist local recruiters (Kimmel, Nogami, Elig, and Gade, 1986). The HARP project attempted to assess the utility of such assistance and to determine how to develop a program that would use veterans to assist recruiters.

The HARP project utilized both qualitative and quantitative approaches to obtain the necessary information. Qualitative data was collected from six focus groups conducted with experienced recruiters and from informal discussions with Army personnel involved in the management of the recruiting effort. In addition, subsequent quantitative analyses were conducted on the AES data to address issues raised in the focus groups. Quantitative analyses were also completed on AES data to develop projections of participation rates and profiles of potential HARP volunteers.

The focus group approach employed in the HARP project was primarily exploratory (although not in a theoretical sense). We had preliminary concepts about how HARP could be structured and we wanted to get the reactions of experienced recruiters to these concepts. The focus groups were also somewhat phenomenological in nature because we realized that our perceptions of what would be a useful recruiter aid program may be very different from the recruiters' perception of the program. In other words, we needed to experience their reactions to HARP.

The focus groups were conducted at the Army recruiting school at Ft. Benjamin Harrison with recruiters who were attending advanced recruiter training courses (e.g., the station commander's course). The focus group discussions were recorded on audio tape for subsequent study. The recruiters were asked to discuss the many issues involved in the development of HARP. For example, they discussed their reactions to the use of recent veterans as volunteer recruiter aids, what type of veteran would make a good volunteer, who should select the veterans, and what the volunteer should do to aid the recruiter.

The focus group data collected for the HARP project provided useful information for the feasibility study. For example, recruiters were unanimous in their concern that HARP would undoubtedly add to their administrative burden. They did not want another program that gave them additional milestones and quotas to meet (e.g., a requirement that they contact each veteran in their area within 30 days of their release from the Army). Some of the focus group members were able to overcome their resistance to the introduction of a new program and were able to "brainstorm" new variations of the HARP. For example, recruiters had many suggestions about how some type of an Army alumni association might be beneficial to recruiters.

The HARP project also demonstrates how qualitative and quantitative methods can compliment one another. Many of the concerns expressed by recruiters could be addressed with data collected in the AES. For example, recruiters expressed their concern that the only veterans who would volunteer for HARP would be undesirable (e.g. "rocks"). Analyses of the AES data indicate that there is a relationship between AFQT category (and education level) and expressed interest in volunteering to help recruiters. Individuals from lower AFQT categories (and with less than a high school diploma) express greater interest in volunteering than the more intelligent (higher AFQT) and better educated veterans. However, the projections made based on the quantitative analyses indicated that there will be sufficient numbers of volunteers to allow for fairly extensive screening. That is, even if it was determined to be desirable that only volunteers from upper AFQT categories were selected, there will still be sufficient numbers of veterans volunteering to aid Army recruiters.

Summary and Conclusions

In this paper we have presented the focus group method and illustrated its usefulness in military manpower research settings by describing two recent ARI projects. In both projects, the qualitative data collected in the focus group allowed us to explore new concepts and to obtain a better understanding of the complexity of the issues involved. Focus groups can often provide a wealth of data in a cost effective and timely manner. However, researchers must keep in mind the lack of generalizability of exploratory focus group findings. These findings have to be validated using representative samples and quantitative measures before any firm conclusions can be made.

References

Calder, B.J. (1977) Focus groups and the nature of qualitative marketing research. Journal of Marketing Research, 14, 353-364.

Kimmel,M.J., Nogami, G.Y., Elig, T.W., and Gade, P.A. (1986) The one-term soldier: A valuable Army resource. Soldier Support Journal, 13(5), 4-7.

McTeigue, R.J., Kralj, M.M., Adelman, L., Zirk, D.A., and Wilson, M. (1986) Predecisional processes involved in the enlistment decision. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Westat, Inc. (1986) The 1985 Army Experience Survey: Data sourcebook and user's manual (Research Note 86-01). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wilson, M., Celeste, J., Pliske, R., Elig, E., and Ramsey, V. (1986) Exploring the feasibility of the Hometown Alumni Recruiting Program (HARP). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

THE USE OF FACTORIAL SURVEYS IN MILITARY PERSONNEL AND MANPOWER RESEARCH[1]

Michael J Wilson

Westat, Inc.
1650 Research Blvd.
Rockville, MD 20850

Introduction

This paper provides an introduction to a rigorous method of survey data collection that could prove particularly useful in the area of military personnel and manpower research--the factorial survey.[2] To date, the factorial survey has received its greatest elaboration in the area of market research where it has been used to study consumer purchase decisions.[3]

The methodology of the factorial survey is especially relevant for the study of military manpower and personnel issues as it effectively solves two problems existing in this research area. These are:

- The natural collinearity of the social world, and
- The artificiality of survey response formats.

Solutions to these problems are not gained without cost, however. To obtain the rigor available from a factorial survey, the researcher must impose strict design constraints on data collection before entering the field. The factorial survey, as a result, is a technology not appropriate for all social surveys.

In the next section, the salient characteristics of the factorial survey are discussed. This discussion shows how the factorial survey solves the twin problems of collinearity and artificiality. In the section following, practical difficulties affecting the implementation of a full factorial survey are raised. After this, some illustrative solutions to problems are presented for consideration.

The Full Factorial Survey: Advantages

As commercially used in the modeling of consumer decision making, factorial surveys are designed to elicit preference behavior. That is, respondents are presented a set of comparable consumer goods and asked to rank these from most to least preferred. Using the resulting preference rankings, analysts are then able to determine which characteristics or attributes of the various goods in the

[1]This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-85-C-0476. All statements expressed in this paper are those of the author and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

[2] The term, "factorial survey" is appropriated from Rossi and Nock (1982). As will become obvious, the purpose of the techniques discussed here are to facilitate less than a full factorial presentation of stimuli.

[3] For consumer and market researchers, purchase choices between competing commodities are made on the basis of the tradeoffs a prospective buyer is willing to make between the attributes of the alternatives offered. That is, each alternative has associated with it a variety of attributes such as price, performance, attractiveness, and such. The consumer takes these factors into account and decides, for example, whether he or she can trade off a greater price for better performance. In the context of military personnel and manpower research, military enlistment can be conceptualized as a "purchase" of a service having many attributes such as physical challenge, the opportunity for training, enlistment bonuses, etc.

response set most effected consumer preference.[4]

Presented in this way, the factorial survey appears a rather simple technique for gathering information regarding the structure of individual preferences. Actually, the technique is somewhat more complex. The set of "comparable consumer goods" presented for ranking is very carefully assembled. In fact, it is assembled according to principles of experimental design. An example should clarify.

Suppose the goods set out for preference ranking were enlistment contracts. While there are many characteristics or attributes which can distinguish one enlistment contract from another (e.g., term of enlistment, MOS, bonuses, educational benefits, etc.), let us assume that only two attributes differ in the contracts to be ranked--term of service and bonus. At the risk of being overly simplistic, assume further that these attributes have only two levels--three years and six years and none and $5,000 for term and bonus, respectively. With this information in hand, the researcher can now construct a response set for ranking.

The full factorial response set requires four contract profiles having the following attribute configurations:

- three year term - no bonus,
- three year term - $5,000 bonus,
- six year term - no bonus, and
- six year term - $5,000 bonus.

With the collection of preference rankings on this response set, the researcher gains a powerful tool for drawing inferences regarding the affect of term and bonus on the relative attractiveness of enlistment contracts--the data.

Collinearity of the Social World. One of the difficulties associated with the study of real world enlistment decision making (e.g., studies that investigate, for example, the effects of an increase in college fund benefits on enlistment decisions) is the correlation existing between prospect characteristics and the provisions of enlistment contracts offered for consideration. Prospects having lower educational attainment or lower AFQT scores, for example, will never be eligible for certain bonuses. This natural collinearity between prospect and contract characteristics confounds the estimation of the effect on enlistment decisions of, say, of increasing educational benefits.

The full factorial survey avoids this difficulty by constructing a response set where all attributes of the enlistment contracts are orthogonal. That is, the experimental design used for constructing the set of enlistment contracts assures that each attribute of the contract is uncorrelated with every other attribute. In this way, unconfounded estimates of the effects of attributes on enlistment decision making can be obtained.

Artificial Response Format. The response formats most often adopted in social surveys investigating enlistment decisions are also problematic. Frequently, respondents are asked to rate (one-by-one) the importance of various attributes of an enlistment contract (e.g., term of enlistment, size of bonus, etc.) in their enlistment decision. The difficulty here is that the response format does not reflect the environment in which decisions are actually made. Rather than considering alternative courses of actions in isolation, individuals evaluate the ensemble of attributes composing alternative and choose in terms of the total packages. Attribute-by-

[4] The particular statistical techniques used by market researchers for the analysis of preference data include multidimensional scaling, conjoint analysis, logistic regression, and cluster analysis. Readers interested in a discussion of these techniques in the marketing research environment are referred to Green and Wind, 1973.

attribute importance ratings do not effectively capture this environment.

The factorial survey does replicate in some measure the circumstances occurring during decision making. Respondents are faced with a variety of enlistment packages that vary on a number of attributes. They are then asked to rank, using any rules of preference they wish, the packages from most to least preferred.

This concatenation of experimental principles with the social survey has some obvious advantages. Full factorial surveys also pose some rather formidable problems.

The Full Factorial Survey: Respondent Fatigue and the Need for Alternative Experimental Designs

As a concrete illustration of the difficulties associated with full factorial surveys, consider the following. Prospects (respondents) are presented with a full factorial crossing of enlistment packages where the following five attributes of military enlistment are considered: (1) enlistment bonuses, (2) money for college tuition, (3) the opportunity for skill training, (4) job security, and (5) the personal challenge of military service. Assume further that each of these attributes of military service are said to have three levels (e.g., bonuses as well as skill training vary from none to some to maximal). In this circumstance, construction of a full factorial response set requires the presentation of 3x3x3x3x3 or 243 separate enlistment packages to respondents for rating. Figure 1 illustrates this circumstance in the form a table shell for listing all possible enlistment packages.

If one wished to be even more realistic, the decision process could be complicated a bit more. Enlistment decisions are generally made in the context of other viable options such as civilian employment or college enrollment. Pursuing the logic of full factorial presentation, if these options also had five salient attributes, each with three levels, respondents would be required to preference rank 243^3 or 14,348,907 triplets of potential career choices.

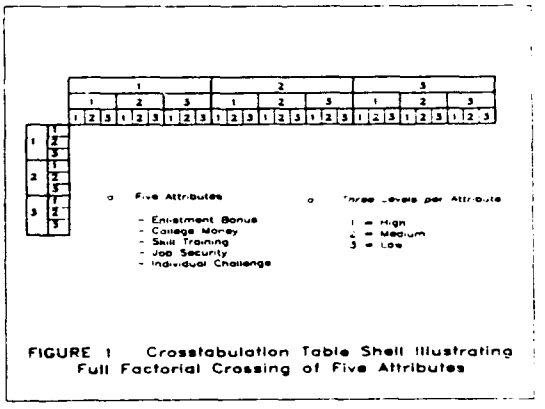


FIGURE 1 Crosstabulation Table Shell Illustrating Full Factorial Crossing of Five Attributes

Obviously, any proliferation of attributes leads quickly to the construction of response sets vastly exceeding the capacity of the respondent's rating ability. It is just not practical to expect respondents to rank millions of choices in order of preference. This is the motivation for introducing less than full factorial experimental designs for respondent ranking tasks.

A method and rationale for reducing the size of response sets can be developed if the information yield from a full factorial design is explored. Consider only the case of preference ranking the 243 enlistment packages identified above. If a full factorial ranking is pursued, information is gained sufficient to allow the estimation of:

- 5 main effect parameters,
- 10 first order interaction parameters,
- 10 second order interaction parameters,
- 5 third order interaction parameters, and
- 1 fourth order interaction parameter,

or a total of 31 parameters.

A significant reduction in respondent burden can be realized if, for example, it is not necessary to estimate all interaction parameters (i.e., some are assumed equal to zero). The place of assumptions regarding respondent fatigue and parameter estimation in the choice of experimental designs are discussed in the next section.

Issues in the Selection of an Optimal Experimental Design

As the (above) hypothetical example of ranking enlistment packages demonstrates, a full factorial crossing of all attributes and levels can quickly lead to the construction of very large response sets. In order to limit their size, less than full factorial experimental designs can be used for set generation. The determination as to which designs are appropriate is made by the answers to questions such as the following:

1. What type of model does the researcher wish to apply?
   a. Main effects only
   b. Main effects plus selected interactions
2. How many attributes does the researcher wish to vary in each response set?
   a. All attributes
   b. A subset of attributes

Each of these questions raises practical considerations that affect the choice of an optimal experimental design. For example, if a main-effects only model is hypothesized to underlie preferences, then a highly fractionated design may be used. Such designs would require respondents to rank only a small subset of the 243 possible enlistment profiles. The decision regarding the number of attributes to vary in the ranking task has critical design implications as well. It may be believed, for example, that respondents are cognitively unable to rank profiles with all five attributes varying simultaneously. In this case a design must be used that varies only a subset of attributes in any single ranking task. Through a consideration of design responses to the above questions, the following discussion provides an indication of the range of ways in which response sets can be constructed.

A Priori Analysis Decisions. One of the first questions that must be addressed concerns the underlying model adopted for analyzing enlistment package preferences. The most parsimonious response sets (in the sense of requiring the smallest number of rankings) are obtained when it is decided that only main effects need to be estimated. For example, if only main effects are to be estimated, orthogonal arrays (Addelman, 1962) could be used to determine the minimal response set required. Continuing the example, an orthogonal array can be constructed which requires respondents to rank only 16 enlistment packages but still allows the unconfounded estimation of all main effects. The savings in response burden in this case is considerable. Figure 2 presents just such a plan.

BONUS	COLLEGE	SKILL	SECURITY	CHALLENGE
L	L	L	L	L
L	M	M	H	M
L	H	H	M	M
L	M	M	M	H
M	L	M	M	M
M	M	L	M	H
M	H	M	H	L
M	M	H	L	M
H	L	H	H	H
H	M	M	L	M
H	H	L	M	M
H	M	M	M	L
M	L	M	M	M
M	M	H	M	L
M	H	M	L	H
M	M	L	H	M

L = LOW M = MEDIUM H = HIGH

FIGURE 2 AN ORTHOGONAL ARRAY ALLOWING ESTIMATION OF MAIN EFFECTS

If, on the other hand, the researcher is unwilling to posit a main-effects only model but instead wishes to estimate selected interactions, fractional factorial designs (Cochran and Cox, 1957) may prove useful for constructing response sets. The trade-off encountered in introducing

interaction terms, however, is in the size of the response set required. Considerably more information is required to estimate the interaction parameters. A fractional factorial design that allows estimation of all main and first order interaction effects requires the ranking of nine separate response sets each with nine profiles. Figure 3 presents the response sets for such an experimental design.

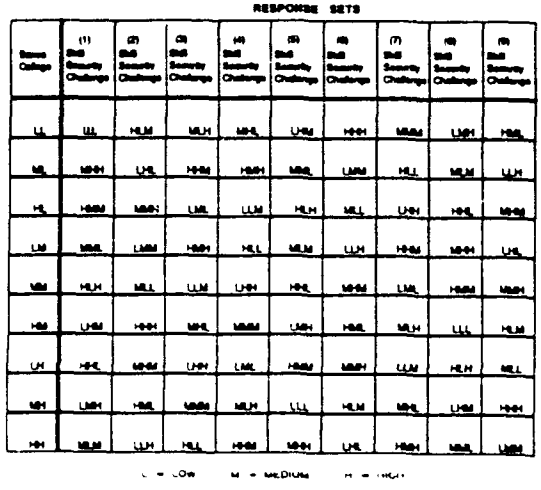


FIGURE 3 A FRACTIONAL FACTORIAL DESIGN ALLOWING MAIN EFFECT AND FIRST ORDER INTERACTION ESTIMATION

Clearly, then, as the number of parameters to be estimated increases, so does response burden. It is unlikely that researchers may often be content with models estimating only main effects so fractional factorial designs will often prove useful. However, it is quite reasonable to assume that in some cases not all first order interactions are of interest. For example, in considering the enlistment decision, it could be argued that no interaction is expected between job security and personal challenge attributes. If this decision were made, then some reduction in response burden could be realized by using a fractional design that allows estimation of only nine (or fewer) of the ten possible first order interactions.

A Priori Cognitive Decisions. Question 2 (above) deals with a somewhat different problem than the first. The issue is not respondent fatigue or the size of the response set, it is the cognitive ability of respondents to rank enlistment profiles when all attributes levels vary simultaneously.

In this case, the researcher may decide that only three or four attributes can be allowed to vary during any ranking task. Here the designs discussed previously are really of little help. Even if one could design (using orthogonal arrays) small sets of three-attribute profiles, the respondent would be faced with making a total of ${}_5C_3$ times 3 or 105 rankings. Clearly, it would be helpful to devise a method to reduce response burden in this kind of situation as well.

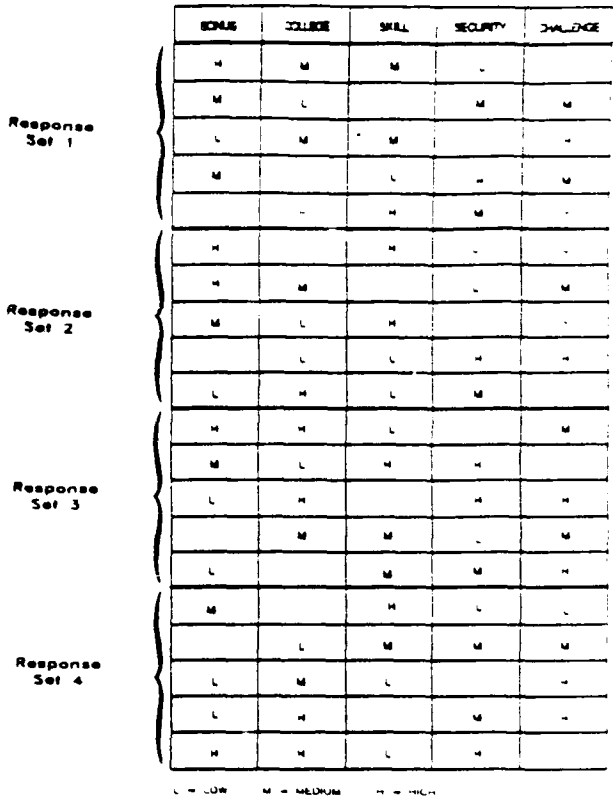


FIGURE 4 A PARTIALLY BALANCED INCOMPLETE BLOCK DESIGN WITH FOUR ATTRIBUTES VARIED

Designs useful for developing response sets in just such circumstances are partially balanced incomplete blocks (PBIB). Such designs can be used to construct response sets where each profile ranked has only, for example, four attributes that vary (Box, *et al.*, 1978). Sets of five profiles are

termed blocks. Each block forms a separate response set. Figure 4 presents a PBIB design that could be used in the present example in which respondents are asked to rate enlistment profiles where only four attributes are varied.

In this particular PBIB design, the five blocks (response sets) are presented to respondents as separate rating tasks. This results in a total burden of 20 preference ratings. This represents a savings in response burden over that which would be required if an orthogonal array design were implemented.

Conclusions

Using hypothetical hypothetical examples drawn from the area of enlistment decision making, this paper has illustrated some of the major issues confronting what has been termed her the factorial survey. As a methodology for rigorously studying choice behavior in the context of military manpower and personnel issues, the factorial survey is evaluated as superior to several existing methods. The potential benefits from this concatenation of experimental design and social survey technologies are considerable. This technology clearly deals with issues of:

- Real world collinearity, and
- Response format artificiality

This is not a methodology acquired without expense, however. Much of the discussion has been concerned with issues of respondent fatigue or cognition. These require the researcher to decide *a priori* the dimensions to be explored in any particular survey Specifically, the researcher using a factorial survey must specify, before entering the field, important survey characteristics such as:

- the number of salient attributes considered,
- the number of levels for each attribute,
- whether main effects only or main effects and selected interactions will be modeled,
- the number of attributes that can be simultaneously varied during ranking, and
- the effective upper limit allowable for respondent burden.

If investigators are able (and willing) to specify their research problems in this much detail, the factorial survey can yield impressive results.

REFERENCES

Addelman, S. (1962). Orthogonal Main-Effect Plans Asymmetrical Factorial Experiments, **Technometrics**, 4, pp. 21-46.

Box, G.E.P., Hunter, W.G., and Hunter (1978). **Statistics for Experimenters**. New York, NY: John Wiley and Sons.

Cochran, W.G. and Cox, G.M. (1957). **Experimental Designs**. New York, NY: John Wiley and Sons.

Green, P.E. and Wind, Y. (1973). **Multiattribute Decisions in Marketing: A Measurement Approach**. Hinsdale, Il: The Dryden Press.

Rossi, P.H. and Nock, S.L. (1982). **Measuring Social Judgments**. Beverly Hills, CA: Sage Publications.

APPLICATIONS OF CONSUMER DECISION MODELS TO MANPOWER RESEARCH

Alan P. Romanczuk
Michael J Wilson

Westat, Inc.
1650 Research Blvd.
Rockville, Md 20850

Introduction

This paper introduces two theories or models of decision making--Fishbein and Ajzen's (1975) theory of reasoned action and Coombs' (1964) unfolding theory of preferential choice--to the arena of military manpower and personnel research. Each of these models has been widely and successfully applied by market researchers in studies of consumer purchase behavior (Lutz, 1975; Bagozzi, 1983; Fiedler, 1972; Davidson, 1973). It is argued here that these models hold considerable promise for military manpower and personnel research, as well, when the objective is to understand individual decision making behavior. In particular, these models can prove especially helpful for the identification and modeling of major factors influ- ıcing, for example, enlistment, attrition, and retention decisions. As motivation for this discussion, we will describe these theories in terms of a military enlistment decision.

These theories are offered as a selection from models currently used in market research. As will be seen, each approaches the modeling of enlistment decision from a different vantage point. Since the purpose of this paper is only to introduce new enlistment decision analysis paradigms for consideration, no evaluations will be offered with regard to the relative merits of each *vis a vis* the other. This is a topic better addressed after the accumulation of empiricai findings.

Decision Models

Before entering into a discussion of the theories of reasoned action and preferential choice, it will be helpful to specify what we are terming here a decision theory. By a decision theory, we reference a modeling of the psychological processes involved during decision making. When an enlistment decision model, for example, is discussed, we are referring to a model which has as its analytic components (i.e., variables) the mental processes and concepts leading to a decision.

This conception of a decision model is different than that commonly used in current enlistment decision research. The focus of most models reported in the literature is on the effects of exogenous factors on enlistment decisions. Economists, for example, have investigated the effects of market factors such as unemployment on aggregate enlistment rates (Dale and Gilroy, 1984) while other social scientists have brought attention to normative influences such as patriotism (Burk and Faris, 1982) on the enlistment decision. Our focus, obviously, is different in that we stress mental processes rather than external influences. Both the theory of reasoned action and the unfolding theory of preferential choice, therefore, are offered as perspectives useful for extending the agenda of current enlistment decision research.

This extension of the scope of enlistment decision research, we believe, will enhance the insights gained by previous and ongoing enlistment research. By focusing on the mental processes and tradeoffs involved in an enlistment decision, a more "textured" understanding may be gained regarding how external factors affect the decision process.

We now turn to a consideration of, first, Fishbein and Ajzen's theory of reasoned action then Coombs' unfolding theory of preferential choice. Fortunately, each is adequately summarized in a visual format. Following these discussions, a few remarks will be offered regarding the place

of these theories/models in the overall scheme of enlistment research.[1]

The Theory of Reasoned Action

The goal of the theory of reasoned action is to understand and predict the behaviors of individuals. As a central tenet, it assumes that individuals construct their decision behavior in a systematic manner. Additionally, individuals are believed to process and evaluate information regarding choice alternatives in an orderly fashion (i.e., the processing of information is not idiosyncratic). As a result, the theory of reasoned action posits the hypothesis that individuals make choices among multiattribute alternatives according to a prespecified series of mental rules.

Figure 1 provides a simplified illustration of the main factors that influence behavior according to this theory. Behavioral beliefs concern the type of outcome that the individual expects from a given behavior. Normative beliefs represent what the individual believes "significant others" think he or she should do regarding the behavior. Behavioral and normative beliefs directly impact attitudes regarding the behavior and perceived subjective norms, respectively. For any given situation and individual, these attitudes and subjective norms vary in relative importance. These attitudes and subjective norms directly determine intentions regarding the behavioral act. Intentions, in turn, determine the ultimate behavior itself.

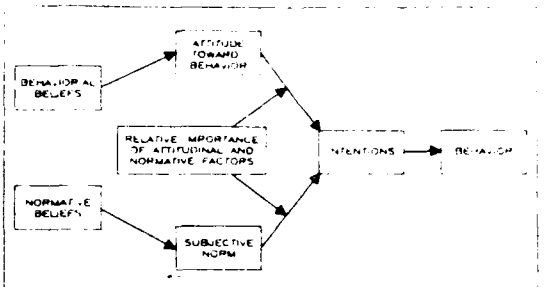


FIGURE 1 Determining Factors of Behavior

Implementing the Theory of Reasoned Action. Operationally, researchers employing this model must first establish the set of attributes that are salient (i.e., the chance to learn a skill or earn money for college) to individuals considering, for example, military enlistment. It is important here to distinguish between what may be salient to the prospect population and what is salient to manpower and personnel planners. The two may not coincide. Civilian/military wage ratios, for example, may be salient to policy makers but have no effects upon an enlistment decision.

In establishing the set of salient attributes (either through a study of previous research findings or by independently polling prospects), researchers should guard against the tendency to include a long list of attributes. Although an individual may hold a large number of beliefs about enlistment attributes, it appears that only a relatively small number of beliefs will serve to determine his or her attitude toward enlistment. Research on attention span, apprehension, and information processing suggest that an individual is capable of attending to or processing 5 to 9 items of information at a time (e.g., G.A. Miller, 1956; Mandler, 1967).

Once a set of salient attributes is established, modeling may begin. Using the expectancy-value approach subsumed under

[1] No attempt is made here to trace the development and elaboration of either the theory of reasoned action or the unfolding theory of preferential choice. Within the confines of this paper we will address, rather, the present formulation (and use) of these theories. (For descriptions of the historical and methodological developments leading to the present theories, see Ajzen and Fishbein, 1980; Fishbein & Ajzen, 1975; Rosenberg, 1956; Coombs, 1964; Green and Wind, 1973; Green and Rao, 1972). Suffice to say that these perspectives on decision making have been "in the field" and working for market researchers for well over a decade.

this model, behavioral beliefs are measured on two dimensions: (1) belief probability (e.g., "Obtaining useful job training while in the military is: *likely...unlikely*"), and (2) belief evaluation (e.g., "Joining the military to receive job training is a *good...bad* idea"). That is, both the likelihood that enlistment will lead to an outcome (e.g., job training) and the desirability of the outcome are measured.

Belief probabilities and evaluations are then combined using the following rule:

$$\Sigma b_i e_i$$

where b corresponds to the probability that military enlistment will lead to a specified outcome and e corresponds to the respondent's evaluation of that outcome. These products are then summed across all salient outcomes to form a prediction of an individual's attitude toward enlistment.

The other set of beliefs to be measured are normative beliefs. Normative beliefs are also measured on two dimensions: (1) salient referent normative beliefs (e.g., "My father thinks it would be a *good...bad* idea to enlist"), and (2) motivation to comply (e.g., "I *do...do not* care what my father thinks"). It is important in the measurement of normative beliefs that all significant others (i.e., friends, classmates, teachers, etc.) are identified. Once measurements are obtained on normative beliefs they are combined as were behavioral beliefs to form a prediction of subjective norms used in the enlistment decision.

The theoretical expectation that the attitude and subjective norm toward enlistment are predicted by behavioral and normative beliefs must be evaluated. This is most commonly accomplished by correlating belief measures with direct indicators of attitudes and subjective norms.[2] The informal rule-of-thumb adopted by researchers using this model is that correlations should exceed .60 to constitute any measure of validation for the model (Ajzen and Fishbein, 1980). In cases where a well-designed instrument is used this poses little difficulty as correlations regularly exceed .80.

The linkages between attitudes and subjective norms to enlistment intentions are most often made using a structural equation model such as multiple regression.[3] This step in the modeling strategy determines the relative influence of attitudes and subjective norms on enlistment intentions.[4] This information, in turn, increases the researcher's understanding of the determinants of behavioral intention. Finally, the relationship between intention and behavior is modeled once the referent time span used for modeling has lapsed. In the present example, if a six month horizon was used for studying enlistment decisions, a measure of enlistment behavior would be gathered in order to validate the predictions made on the basis of enlistment intentions.

We will now move to a consideration of the unfolding theory of preferential choice.

The Unfolding Theory of Preferential Choice

Coombs' unfolding theory of preferential choice approaches the modeling

---

[2] Attitudes and subjective norms are frequently measured using semantic differentials. A partial measurement of the attitude toward enlisting, therefore, would be the response to the statement, "My enlisting in the military would be *smart...dumb*."

[3] In recent years researchers have increasingly used covariance structure techniques as they allow the explicit estimation of the measurement and structural models simultaneously (Bagozzi, 1984).

[4] Intention may be measured using a format such as: "It is *very likely...highly unlikely* that I will enlist in the military within the next six months.

of decision making using a perspective exactly opposite from that adopted by Fishbein and Ajzen. Rather than composing the decision process from its constituent parts, unfolding theory takes as its point of departure decisions themselves.[5] Using decisions as raw data, unfolding theory attempts to work backward from the end point in the process and infer the latent constituent parts and the manner in which they were combined to reach a decision. This requires some explanation.

The raw data most often used in the modeling phase of an unfolding analysis are preference orderings. Respondents are presented with a collection of items and asked to rank them from most to least preferred. The collection can be of anything that is describable in terms of a series of attributes (e.g., automobiles can be described in terms of attributes such as price, option packages, gas mileage, performance, etc.). Assume for the moment that respondents were asked to perform a preference ranking of enlistment packages varying in only one attribute--*the amount* of physical challenge required during the term of enlistment.

Using preference rankings of enlistment packages, unfolding theory begins with the recognition that individuals express preferences when presented with alternative courses of action. For example, a particular individual may prefer the enlistment package requiring only a minimal physical challenge. From this ideal point, the theory predicts that increases and decreases in physical challenge are monotonically less preferred the further they deviate from a minimal level. That is, the utility function posited for individuals is U shaped.

The theory proceeds on the basis of this assumption to articulate the manner in which an underlying (latent) utility scale can be constructed from multiple preference rankings. Figure 2 illustrates a hypothetical example where two individuals' preference ranking are unfolded to form a utility scale for physical challenge.

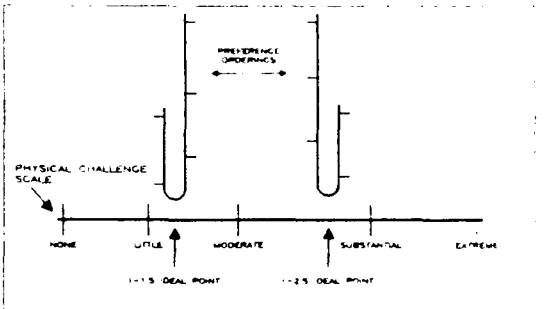


FIGURE 2 UNFOLDING OF TWO PREFERENCE RANKINGS ON SINGLE UNDERLYING DIMENSION

In this figure, Individual 1's ideal point is little physical challenge. In decreasing order of preference, this individual then ranks moderate challenge, no challenge, substantial challenge, and finally extreme challenge. Individual 2 provided a different preference ranking. Most preferred was substantial physical challenge followed by moderate challenge, extreme challenge, little challenge, and no challenge. We can see that as the preference rankings are unfolded at their ideal points, they form a consistent underlying scale.

While this hypothetical example only reinforces the obvious and yields no new information, it does highlight the fact that the theory of preferential choice can be considered, on one level, an algorithm for scale construction. Though the unidimensional example may seem trivial, consider a multidimensional case where five attributes of an enlistment package are varied for the preference ranking task (e.g., enlistment bonuses, money for college, skill training, job security, and physical challenge). In this circumstance the extraction of underlying scales becomes akin to factor analysis. That is, with five attributes varied, it is possible to define five scales. However, it is also possible that analysis will determine that only three scales are needed to adequately represent the preference ratings. Though unlikely, it is

---

5 The use of the word "decision" here is potentially misleading. In actuality, information is collected on the "preferences" of individuals. This point will become clear as we proceed.

conceivable that money for college and enlistment bonuses are unfolded onto one scale as are job security and skill training).

It is in the multidimensional case that the "working backward" from decisions to inferences is most apparent. From preferential rankings, unfolding theory is able to identify the salient dimensions or attributes which explain the preferences. In this way, the unfolding theory is a more exploratory perspective than the theory of reasoned action. It acknowledges that decisions are made, but remains initially uncommitted with regard to the salient dimensions that motivate decisions.

Implementing the Unfolding Theory of Preferential Choice. The major characteristics of a study of enlistment decisions from the vantage point of unfolding theory should already be clear. The basic data required are preference rankings. These can be gathered in a number of ways.

If a survey is administered one-on-one, respondents can be given a set of cards which describe various enlistment packages and asked to place them in order from most to least preferred. Paper and pencil surveys can present a list of packages to respondents and request that they number enlistment packages in order of preference. Telephone surveys present difficulties in that respondents cannot be expected to efficiently rank a series of packages. In this case, respondents can be presented with several distinct ranking tasks where only two packages are ranked at a time. From such paired comparison data, it is possible to reconstruct a complete profile of rankings for all packages (David, 1963; Gulliksen and Tucker, 1961).

The problem of eliciting preference rankings in telephone interviews raises the more generic implementation issue of respondent burden. Consider, for example the enlistment decision. If there were five attributes of packages one wished to study for their impact on decisions and each had three levels, the total potential list of packages that could be ranked is 3x3x3x3x3 or 243. This is obviously is too large a task. Market researchers have responded to just such a circumstance by adopting experimental design which eliminate the total number of packages which must be ranked in order to extract the needed information (Green, 1974).

CONCLUSIONS

This paper has provided an overview of two decision models we believe could be of use to military personnel and manpower researchers. Both models are explicitly psychological in that their respective foci are the mental processes leading to a decision. This perspective is not currently represented to any great degree in military decision research. For this reason, such models can be useful in extending the research program of military researchers.

As was indicated in the introduction, no evaluation of the relative merits of each model has been provided. Two characteristics of these models, however, should be explicitly mentioned.

- Parsimony
  - no exogenous variables used
  - decision variables only
- Each theory is a measurement model
  - variables clearly identified
  - analysis flows from theory

These characteristics are important in that they prescribe a fixed domain for the use of these models. That domain is the decision making process of individuals. For this reason, exogenous variables are not appropriate. This clear delineation of theoretical territory, so to speak, yields very parsimonious models. In addition, each theory explicitly contains a measurement model. The theory of reasoned action

presents a measurement model whereby decisions are composed from their constituent parts. The unfolding theory of preferential choice, on the other hand, decomposes decision behavior and identifies the salient attributes which explain the decisions made. While each theory approaches measurement from a different direction, both contain a methodology for determining the appropriateness of variable inclusion.

Potentially, such individual decision making models could aid in the construction of cross-level theories that would link aggregate- with individual-level enlistment decision theories.

REFERENCES

Ajzen, I. and Fishbein, M. (1980). **Understanding Attitudes and Predicting Social Behavior**. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bagozzi, R.P. (1984). A Field Investigation of Causal Relations Among Cognitions, Affect, Intentions, and Behavior, **Journal of Marketing Research**, Vol. 19, pp. 562-584.

Burke, J.S. and Faris, J.H. (1982). **The Persistence and Importance of Patriotism in the All-Volunteer Force.** (USAREC Study Report 82-6). Fort Sheridan, Il: Program Analysis and Evaluation Division.

Coombs, C.H. (1964). **A Theory of Data**. New York, NY: John Wiley and Sons.

Dale, C. and Gilroy, C (1984). Determinants of Enlistments: A Macroeconomic Time-Series View, **Armed Forces and Society**, Vol. 10, pp. 192-210.

David, H.A. (1963). **The Method of Paired Comparisons**. New York, NY: Hafner Publishing.

Davidson, J.D. (1973). Forecasting Traffic on STOL, **Operations Research Quarterly**, Vol. 24, pp. 561-569.

Fiedler, F.E. (1982). Personality, Motivational Systems, and Behavior of High and Low LPC Persons, **Human Relations**, Vol. 25, pp. 391-412.

Fishbein, M. and Ajzen, I. (1975). **Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research**. Reading, MA: Addison-Wesley.

Green, P.E. and Wind, Y. (1973). **Multiattribute Decisions in Marketing: A Measurement Approach**. Hinsdale, Il: The Dryden Press.

Green, P.E. and Rao, V.R. (1972). **Applied Multidimensional Scaling**. Hinsdale, Il: The Dryden Press.

Green, P.E. (1974). On the Design of Choice Experiments Involving Multifactor Alternatives, **Journal of Consumer Research**, Vol. 1, September, pp. 61-68.

Gulliksen, H. and Tucker, L.R. (1961). A General Procedure for Obtaining Paired Comparisons from Multiple Rank Orders, **Psychometrika, 26**, pp. 173-183.

Lutz, R.J. (1975). Changing Brand Attitudes through Modification of Cognitive Structure, **Journal of Consumer Research**, Vol. 1, pp. 49-59.

Mandler, G. (1967)."Verbal Learning." In T.M. Newcomb (ed.), **New Directions in Psychology**, Vol. 3. NY: Holt, pp. 1-50.

Miller, G.A. (1956). "The Magical Number Seven; Plus or Minus Two: Some Limits on Our Capacity for Processing Information." **Psychological Review**, 63, pp. 81-97.

Rosenberg, M.J. (1956). Cognitive Structure and Attitudinal Affect, **Journal of Abnormal and Social Psychology**, Vol. 53, pp. 367-372.

Using Confirmatory Factor Analysis to Aid in Assessing Task Performance[1]

Jeffrey J. McHenry
American Institutes for Research
Washington, DC

James H. Harris
Human Resources Research Organization
Alexandria, VA

Scott M. Oppler
American Institutes for Research
Washington, DC

In their landmark 1959 paper, Campbell and Fiske urged psychologists to adopt a multitrait-multimethod approach to the measurement of psychological constructs. Over the past 25 years, psychologists have applied Campbell and Fiske's ideas to a host of assessment problems.

The Campbell and Fiske paper had a profound impact on the design of the U.S. Army Research Institute's Project A. The goal of Project A is to validate the Armed Services Vocational Aptitude Battery (ASVAB) and a set of new, experimental predictor tests. Through the first four years of Project A, we have devoted much of our time and resources to the development of reliable, valid measures of job performance. The development efforts were guided by our theory of job performance, which holds that job performance is multidimensional. There is no single attribute, outcome, or factor that can be pointed to and labeled as "job performance" (Campbell & Harris, 1985; Hanser, Arabian & Wise, 1985). Consequently, one of the critical activities in performance measurement is to describe the basic factors that comprise performance. To ensure that these factors were measured adequately, four different types of job performance measures were developed: hands-on job sample tests, multiple-choice knowledge tests, performance rating scales, and administrative measures.

In a large-scale study of those measures, almost 5000 first-tour enlisted personnel in nine Army Military Occupational Specialties (MOS) participated in a one and one-half day job performance assessment last summer and fall. Their data were used to help build a model of first-tour enlistee job performance (Wise, Campbell, McHenry & Hanser, 1986).

In developing this model, one of the first things we noticed was that scores on the hands-on and written job knowledge tests were fairly highly correlated, as were scores from the rating scales and administrative measures. However, the hands-on and written tests were only moderately correlated with the performance ratings and administrative measures, suggesting that these different measurement methods were tapping different portions of the job performance space. The hands-on and written knowledge tests were measuring "can do" or maximal performance, while the rating scales and administrative

---

[1]This research was funded by the Army Research Institute Contract No. MDA-903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions of the U.S. Army Research Institute or the Department of the Army.

measures were assessing "will do" or typical performance. Within the "can do" performance domain, two performance constructs were identified. The first, Core Technical Proficiency, was comprised of those performance components that were specific to a particular job (e.g., "typing correspondence" for an administrative specialist, "driving a tank" for a tank crewman, etc.). The second construct, General Soldiering Proficiency, was defined by common soldier tasks (e.g., navigation, first aid, operating an M16). In addition to these two "can do" constructs, three "will do" constructs were also identified: Effort and Leadership; Personal Discipline; and Physical Fitness and Military Bearing.

One of the most important implications from the Wise et al. study is that researchers must be aware of possible confounds between trait and method when they use a multitrait-multimethod approach to assessment. However, in the Wise et al. study, the performance ratings were not designed to measure the same traits as the hands-on and written knowledge tests. The performance rating scales were designed to measure broad dimensions of job performance, and had been developed using the critical incident technique (Flanagan, 1954). The hands-on and written tests were designed to measure performance of critical tasks. The purpose of this paper is to see if similar results are obtained when task-specific performance rating scales are used instead of rating scales developed from critical incidents.

Method

Subjects

Subjects were first-tour enlisted soldiers drawn from the following nine MOS:

- Infantryman (11B) (N = 613)
- Cannon Crewman (13B) (N = 535)
- Armor (Tank) Crewman (19E) (N = 410)
- Radio Teletype Operator (31C) (N = 280)
- Light Wheel Vehicle Mechanic (63B) (N = 477)
- Motor Transport Operator (Truck Driver) (64C) (N = 527)
- Administrative Specialist (71L) (N = 344)
- Medical Specialist (91A) (N = 410)
- Military Police (95B) (N = 588)

Measures

The following three sets of measures were administered to each subject:

- Hands-on performance tests on approximately 15 critical tasks. These tasks were carefully sampled from the domain of important tasks for each job. Each hands-on test consisted of a number of critical steps, with each step scored GO or NO GO. The number of steps within a task varied from as few as six to as many as 62. The hands-on task score was the percent of steps scored GO.
- Written job knowledge tests consisting of three to 15 questions on each of the critical tasks. The score on each task was the percent of questions answered correctly.
- Supervisor and peer ratings of performance on each of the critical tasks. Each rater rated his/her assigned subject's performance on each task in terms of how well the subject

performed the task compared to other soldiers. On average, subjects were rated by two supervisors and three peers. Mean supervisor and mean peer ratings were computed for each task. These two mean ratings were then averaged to compute the final task rating.

Results

Model of Task Performance

Campbell (in preparation) has described a model of first-tour soldier task performance that was derived using the data from the subjects in this study. Briefly, the intercorrelations among the within-method task scores were examined to identify similarities across methods and across MOS. On this basis, five task factors were identified:

- Core Technical. Included tasks that were specific to the MOS (e.g., "typing correspondence" for an administrative specialist, "driving a tank" for a tank crewman, etc.).
- Communication. Included tasks related to operating a radio set.
- Vehicle Operation and Maintenance. Included tasks involving driving a vehicle and performing simple operator maintenance.
- General Soldiering. Included tasks that are critical to field and combat performance, such as weapons operation and maintenance, navigation, etc.
- Safety and Survival. Included tasks related to safety and first aid, including procedures for coping with nuclear/biological/chemical (NBC) conditions.

Each of the critical tasks was assigned to one of the five task factors. As Table 1 shows, some of the factors were not assessed for some of the MOS. For example, for Administrative Specialist (71L), there were no tasks for two of the factors: Communication, and Vehicle Operation and Maintenance. For Infantryman (11B) and Motor Transport Operator (64C), the table indicates that there was no Core Technical task factor. This is because Communication, General Soldiering, and Safety and Survival are the core technical part of the 11B job, and Vehicle Operation and Maintenance is the core technical part of the 64C job.

Table 1

Measurement of Task Factors by MOS

Task Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Core Technical		X	X	X	X		X	X	X
Communication	X	X	X	X					X
Vehicles				X		X			
General Soldiering	X	X	X	X	X	X	X	X	X
Safety/Survival	X	X	X	X	X	X	X	X	X

Analyses

The objective of this study was to test whether the observed

correlations among the hands-on and written knowledge tests and task ratings were consistent with the Campbell task factor model. Confirmatory factor analysis (Joreskog & Sorbom, 1981) was used to conduct this test.

To perform a confirmatory factor analysis, one must first specify a set of latent constructs that explains the relationships among a set of observed variables. In the present study, two sets of latent constructs were hypothesized. The first consisted of the task factors identified by Campbell. The second included three method factors, representing the three measurement methods that were used to assess subjects' performance.

Each task score was allowed to "load" on one task factor and on one method factor. For example, we allowed the hands-on task score for "typing correspondence" for 71L to load the Core Technical task factor and the Hands-On method factor; its loadings on the remaining factors were constrained to zero.

We also specified the relationships among the underlying factors. We specified that the three method factors were uncorrelated with each other and with any of the task factors. However, we allowed the task factors to be correlated.

The confirmatory factor analysis program, LISREL, then derived the non-zero loadings of the tasks on the task and method factors and the correlations between the task factors. These loadings and correlations were derived to be as consistent as possible with the observed correlations among the task scores.

Finally, LISREL computed a chi-square index to describe the level of agreement between the observed correlations and the factor loading and correlations that it has derived. Essentially, LISREL does this by working backwards and estimating the correlations from the factor loadings and correlations, then comparing these estimated correlations to the observed correlations. A large and significant chi-square value indicates that the observed and estimated correlations differ.

The portion of Table 2 labeled "With Task Ratings" shows results from the present study. The table shows that the observed and estimated correlations differed significantly for all nine MOS.

Table 2

Fit between the Task Factor Model and the Observed Correlations

	With Task Ratings			Without Task Ratings			Change		
MOS	Chi^2	df	p	Chi^2	df	p	Chi^2	df	p
11B	632.6	492	.00	182.6	206	.88	450.0	286	.00
13B	3250.7	1218	.00	788.2	521	.00	2462.5	697	.00
19E	1033.5	696	.00	232.4	293	.99	801.1	403	.00
31C	1372.5	935	.00	439.8	395	.06	1335.7	540	.00
63B	1300.5	942	.00	440.3	402	.09	860.2	540	.00
64C	791.5	492	.00	234.9	206	.08	556.6	286	.00
71L	950.7	492	.00	225.2	206	.17	725.5	286	.00
91A	1910.1	942	.00	719.7	402	.00	1190.4	540	.00
95B	1359.0	813	.00	414.5	344	.01	944.5	469	.00

We felt that there were two possible reasons for this result. Our first hypothesis was that the model was not appropriate, and that a different set of task factors would do a better job of explaining the observed correlations among task scores. Our second hypothesis was that the model was working quite well for the hands-on and job knowledge tests, but was not appropriate for the task ratings because the task ratings were not measuring "can do" performance. We chose to investigate this second hypothesis.

Marsh and Hocevar (1983) have suggested a method for testing such hypotheses using LISREL. To implement their suggestion, we re-ran LISREL without the task ratings data (and dropping the ratings method factor). According to Marsh and Hocevar, one can compare the chi-square and degrees of freedom from the new analyses with the chi-square and degrees of freedom from the original analyses to determine whether the model fit the data better after the ratings data were dropped. The portion of Table 2 labeled "Change" shows that the improvement in fit was significant for all nine MOS. The portion labeled "Without Task Ratings" shows that the Campbell model was consistent with the observed correlations for seven of the nine MOS.

Discussion

The results in Table 2 indicate that the factor structure of the task rating scales is different from that of the hands-on and written job knowledge tests. Other analyses (not reported in this paper) indicated that the performance construct most highly correlated with the task rating scales was the Effort and Leadership "will do" performance construct.

The data point to the need to consider the relationship between measurement methods and traits when employing multitrait-multimethod techniques to assess individual differences. Even though measures drawn from two methods have the same name (e.g., "driving a tank"), it is no guarantee that they measure the same underlying construct. Researchers must be guided by theory and previous research in deciding when it is appropriate to expect that measures from different methods will be useful in analyzing a given construct.

Within the field of performance measurement, for example, Hunter (1983) has shown that the relationship between cognitive abilities and supervisory performance ratings is different from the relationship between cognitive abilities and hands-on or written knowledge tests. Hunter has developed a theory to account for the relationships among different performance measures. His work, the Wise et al. (1986) research, and this research all suggest that one should not expect a one-to-one correspondence between performance ratings and other measures of job performance.

Other results from Project A promise to shed additional light on the constructs underlying different performance measures. For example, preliminary results of Project A validity analyses (Campbell, 1986) indicate that cognitive ability tests are much more highly correlated with the "can do" performance constructs (i.e., with scores from the hands-on and written knowledge tests) than with the "will do" performance constructs (i.e., with performance ratings and administrative measures). On the other hand the Assessment of Background and Life Experiences (ABLE) (Hough, Barge & Kamp, in

press), a temperament/biodata questionnaire, was a much better predictor of "will do" performance than "can do" performance. In fact, the validity of ABLE scales often exceeded the validity of ASVAB scales for predicting performance ratings (Campbell, 1986).

Finally, the present study demonstrates the usefulness of confirmatory factor analysis for testing theories about the latent variables underlying a set of observed scores. Most commonly, researchers use confirmatory factor analysis programs such as LISREL to obtain statistical tests of the agreement between their theories and a set of observed data (Joreskog & Sorbom, 1981). In this study, we also used LISREL to test two competing theories (Marsh & Hocevar, 1983). As these results demonstrate, LISREL provides a powerful tool for improving the quality of our theories and the conclusions that we draw from our data.

References

Campbell, C. H. (in preparation). Developing basic criterion scores for hands-on tests, job knowledge tests, and task rating scales (ARI-TR-___). Alexandria, VA: U.S. Army Research Institute.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Campbell, J. P. (1986, August). Project A: When the textbook goes operational. Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, DC.

Campbell, J. P., & Harris, J. H. (1985, August). Criterion reduction and combination via a participative decision making panel. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

Hanser, L. M., Arabian, J. M., & Wise, L. L. (1985). Multidimensional performance measurement. Proceedings of the 27th Annual Conference of the Military Testing Association. San Diego: Military Testing Association.

Hough, L. M., Barge, B. N., & Kamp, J. D. (in press). Non-cognitive measures: Pilot testing. In N. G. Peterson (Ed.), Development and field test of the Trial Battery for Project A (ARI-TR-___). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisory ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, NJ: Lawrence Erlbaum Associates.

Joreskog, K. G., & Sorbom, D. (1981). LISREL VI user's guide. Mooresville, IN: Scientific Software.

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 20, 231-248.

Wise, L. L., Campbell, J. P., McHenry, J. J., & Hanser, L. M. (1986, August). A latent structure model of job performance factors. Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, DC.

ESTIMATION OF SIMULTANEOUS STRUCTURAL EQUATION MODELS:

APPLICATIONS IN LISREL

David K. Horne and Curtis L. Gilroy*

Social scientists are often faced with the problem of estimating the structural parameters of simultaneous equation models. Estimation of structural equations has been simplified with the introduction of LISREL, but the parameter estimates may be highly sensitive to the specification of the error structure and to the relationships between the endogenous variables. In LISREL, the models are specified as a result of defining the coefficient and error matrices (Joreskog 1982). However, the relationship between the matrix formats and the estimation techniques are somewhat obscure.

This paper discusses possible specifications between endogenous variables, between equation disturbances, and the implications for the appropriate estimators. These comments are formulated in the LISREL notation framework (Joreskog and Sorbom 1984), but are applicable to any estimation procedure. Four general specifications are discussed here: (1) recursive equations, (2) seemingly unrelated equations, (3) simultaneous equations with uncorrelated disturbances, and (4) simultaneous equations with correlated disturbances.

The structural model in LISREL can be defined in simple form as

$$1. \quad Y = BY + GX + E,$$

where Y is a vector of endogenous variables, X is a matrix of exogenous variables, and E is defined as a vector of disturbances. This example assumes no latent variables, but these can easily be incorporated into the model with the addition of the measurement models. The notation is slightly clumsy without Greek characters, but B and G denote the coefficient matrices, with the small letters (b,g) referring to elements of the matrix. For the purposes of simplification the

---

*The authors are Economist and Chief Economist of the Manpower Personnel and Policy Research Group, U.S. Army Research Institute. The paper reflects the views of the authors and does not necessarily represent those of the U.S. Army or the Department of Defense.

above system may be assumed to represent the following two-equation model:

2. $y_1=b_{12}y_2 + g_{11}x_1 + g_{12}x_2 + e_1$

3. $y_2=b_{21}y_1 + g_{23}x_3 + g_{24}x_4 + e_2$

The Recursive Model

The simplest system is a recursive one. In that case the B matrix is lower-triangular, which can be represented in the two-equation model as:

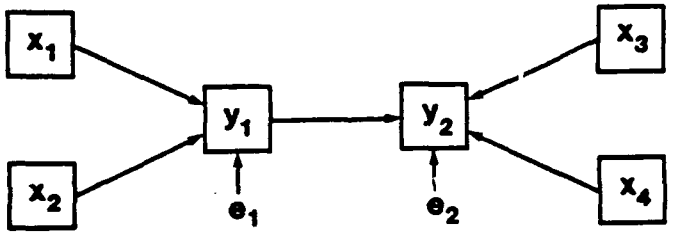
4. $y_1 = g_{11}x_1 + g_{12}x_2 + e_1$

5. $y_2=b_{21}y_1 + g_{23}x_3 + g_{24}x_4 + e_2$

In this formulation y_2 is omitted from the first equation. Y_1 is predetermined in the second equation. The ordinary least squares estimator will yield unbiased and efficient parameter estimates for each equation individually if the errors have zero mean and are independent and identically distributed (as a scalar identity matrix) within each equation. If the errors do not have identical variance within each equation the error covariance matrix for each equation will no longer be diagonal. In this case each equation may be estimated using a generalized least squares (GLS) technique. The GLS estimator will be the best linear unbiased estimator given the full disturbance covariance matrix.

The path diagram of the recursive model is illustrated below. The arrow from y_1 to y_2 implies that y_1 influences y_2, but the effect does not work in the opposite direction.

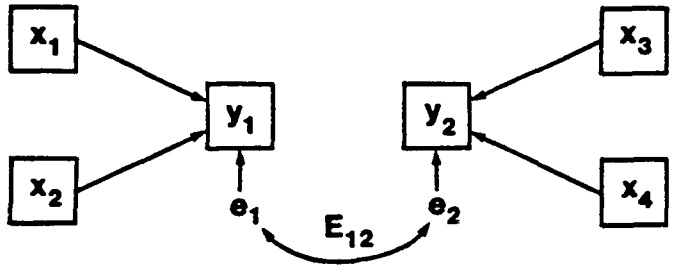
Figure 1: Recursive Model



The Seemingly Unrelated Model

The seemingly unrelated model is derived on the assumption that the B matrix is zero (no simultaneity), but the disturbance terms are correlated between equations. In terms of equations 2 and 3, e_1 is correlated with e_2. This may be the case if unmeasured or omitted variables are reflected in the error term of each equation. Information contained in the covariance matrix of the disturbances should increase the efficiency of the parameter estimates. This "seemingly unrelated regression" (SUR) is implied in LISREL by allowing a symmetric disturbance matrix in place of the diagonal residual matrix. The SUR approach treats the system in equation 1 as a single equation, and applies the GLS method to the entire system (Zellner, 1962). The path diagram that reflects the SUR specification is given below.

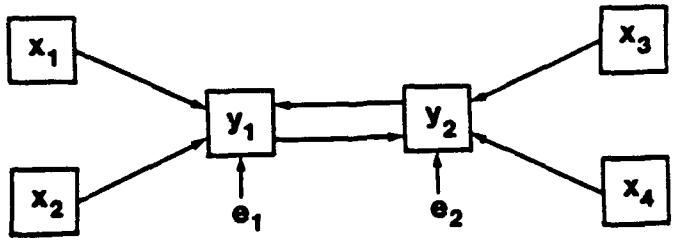
Figure 2: Seemingly Unrelated System



The Simultaneous Model with Uncorrelated Errors

A simultaneous equation model is implied by a symmetric B matrix, as illustrated in equations 2 and 3. It is assumed here that the error matrix E is diagonal. The path diagram for this system is given in figure 3. Ordinary least squares does not generate unbiased or consistent estimates of the structural parameters because the endogenous regressors are correlated with the equation errors. If the equations are just identified, unique structural parameters can be derived from the reduced form parameter estimates. However, consistent estimators are available for overidentified equations.

Figure 3: Symmetric System, Uncorrelated Disturbances



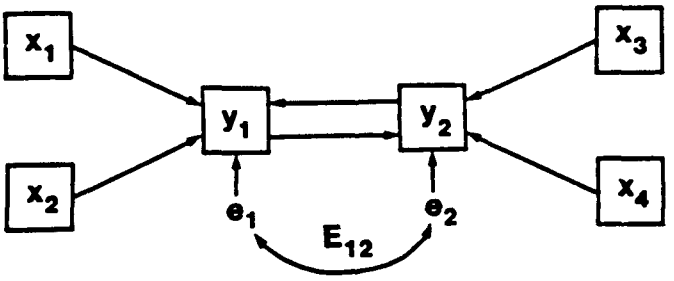
A variety of single-equation estimators can be used to generate unbiased and consistent estimates of the structural parameters of individual equations within the linear system if the particular equations of interest are overidentified. The estimators are generally classified as instrumental variable methods, because they involve substituting an "instrument" for the endogenous regressors. An instrument should be correlated with the endogenous regressor but uncorrelated with the error term in the probability limit. In most cases suitable instruments cannot be found. However, instruments can be created. Two-stage least squares (2SLS) is one method of creating an instrument. The endogenous regressor, y_2, may be regressed against all the exogenous variables, creating a fitted value $\hat{y}_2$. This instrument is a linear combination of the exogenous variables and should therefore be uncorrelated with the disturbance terms. The 2SLS can also be derived as a GLS estimator (Maddala, 1977; Judge et al., 1982).

Maximum likelihood (ML) estimation also yields unbiased estimates for individual equations within a simultaneous system. In the limited-information maximum likelihood (LIML) method, the likelihood function for the entire system is maximized under the assumption of normally distributed errors, subject to the restrictions on only the equation being considered.

The Simultaneous Model with Correlated Errors

In the final model, both simultaneity (in the sense of endogenous regressors) and correlation between disturbance terms is allowed. This path diagram below demonstrates this system, consistent with a LISREL specification.

Figure 4: Simultaneous System with Correlated Disturbances



This model is a generalization of the previously considered simultaneous system. If a model is truly simultaneous, the endogenous regressors in the equations will ensure that the errors are indeed correlated across equations. The single-equation estimators such as 2SLS and LIML can still be used to obtain consistent structural parameter estimates. However, these estimators are asymptotically inefficient. The information contained in the off-diagonal elements of the

disturbance covariance matrix is not utilized by the single-equation estimators. The endogenous variables in the system but not in the estimated equation are also excluded from consideration.

A least squares estimator of the structural parameters of the entire system can be derived through a GLS approach similar to that described for the set of seemingly unrelated regressions (see Judge et al., 1982 or Maddala, 1977). Because elements of the disturbance covariance matrix are unknown, the procedure is generally executed in three stages. First, the equation-specific residuals must be estimated. The resulting residual covariance matrix is used in the GLS estimation of the entire system. When the equation residuals are estimated by 2SLS the approach is referred to as 3SLS (Zellner and Theil, 1962). Further iterations of the covariance matrix may also be incorporated into the estimator. The LISREL approach is not technically a 3SLS; the covariance matrix (which is essentially used to weight the least squares estimates) is derived directly from the sample variance-covariance matrix.

The full-information maximum likelihood (FIML) method is an alternative way of estimating the full simultaneous system. All the restrictions implied by the system, such as zero restrictions on the coefficients, are incorporated into the FIML estimator. If there are no restrictions on the covariance matrix of the residuals, the asymptotic distribution of the FIML estimator is the same as that for the 3SLS estimator.

The GLS and ML estimators in LISREL are implicitly full system estimators. The error covariance matrices are by default symmetric, which incorporates the information in the off-diagonal elements of the disturbance matrix. The GLS estimator in LISREL is a variation of the system GLS estimator used in the 3SLS described above, although the disturbance covariance matrix in LISREL is not generated by 2SLS (Joreskog and Goldberger, 1972). The ML estimator is FIML. A LIML may be approximated by a diagonal error matrix, but the restrictions on the all equations are included in the likelihood function. Specification of a diagonal disturbance matrix using GLS will approximate the seemingly unrelated regression system.

System estimators are preferred to single-equation estimators because of the asymptotic efficiency of the system estimators. The small sample properties of these estimators are not well-defined, although Monte Carlo studies seem to indicate that the asymptotic properties are a good guide to the small sample properties (Judge et al., 1982, pp. 387-388). However, the full system, full information estimators are highly sensitive to model misspecification. Structural parameter estimates may be biased in all equations if any single equation in the system is misspecified. A single equation estimator such as 2SLS may yield less efficient

estimates than the system estimators. However, only those equations which are misspecified will be inconsistently estimated (Hausman, 1978, discusses the consequences of misspecification and provides a specification test). The LISREL user should be careful when specifying a system of equations, understanding the nature of the estimators and the consequences of using single equation versus full system estimators.

Hausman (1978) "Specification Tests in Econometrics," Econometrica 46, 1251-1271.

Joreskog, Karl G. (1982) "The LISREL Approach to Causal Model-Building in the Social Sciences,' in Systems Under Indirect Observation: Causality, Structure, Prediction," ed. K.G. Joreskog and H. Wold. New York: North-Holland (81-99).

Joreskog, Karl G. and Arthur S. Goldberger (1972) "Factor Analysis by Generalized Least Squares,'Psychometrika 37, 243-269.

Joreskog, Karl G. and Dag Sorbom (1984) LISREL User's Guide. Mooresville, Indiana: Scientific Software.

Judge, George G., R. Carter Hill, William Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee (1982) Introduction to the Theory and Practice of Econometrics. New York: Wiley.

Maddala, .G. S. (1977) Econometrics. New York: McGraw-Hill.

Zellner, A. (1962) "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias" Journal of Econometrics 4, 285-294.

Zellner, A. and H. Theil (1962) "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations." Econometrica 30, 54-78.

AN EVENT HISTORY ANALYSIS OF DEP CONTRACT LOSSES[1]

Jeanna F. Celeste
Michael J Wilson

Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850

Background

This paper reports on the application of an analytic technique relatively new to the military manpower research environment, i.e., event history analysis, that was used to examine the Army's Delayed Entry Program (DEP). The DEP is the Army's primary vehicle for recruiting young people into enlisted service. The DEP is essentially a reservation system in which potential recruits may select the time of entry, specific occupational training, and other enlistment options and bonuses as much as 12 months prior to entering active duty. Enlistment in the DEP constitutes a legally binding agreement between the contractee and the military service, the ultimate objective of which is the contractee's entry into active duty. DEP is used extensively by the Army, as well as the other branches of military service, as recruiting for the DEP has numerous advantages over recruiting applicants for immediate shipment. For example, use of the DEP helps to evenly distribute the recruiting mission across changes in the economy and seasonal fluctuations. Setting recruiting missions through a DEP facilitates internal planning for recruit training and manpower requirements (Morey, 1983) and provides flexibility and "lead-time" to adjust mission objectives based upon changes in the recruiting environment. Maintenance of a DEP pool has been shown to enhance recruiting productivity through recruit networking (Freeman, 1980; Hanssens and Levien, 1980), and to reduce training attrition (Schumacher, 1981) by serving as a screening and socializing mechanism for new recruits.

Due to its time-dependent structure, the DEP is subject to contract attrition, i.e., some DEP enlistees fail to report for active duty. Contract attrition from the DEP has several effects on the Army's recruiting mission. First, DEP losses have a disproportionate impact on the yield of desirable high quality recruits. Second, the ability to smooth the flow of enlistments across economic changes and fluctuations in the business cycle is substantially reduced. Finally, resources dedicated to DEP attritees represents some loss of efficiency; recruiting effectiveness is diminished since efforts expended to initially recruit and follow-up DEP attritees could have been allocated to increase recruiting yield in other areas.

The DEP research project was undertaken with the purpose of extending the knowledge of the DEP contracting dynamics in a manner consistent with the goal of improving DEP management. Conceptually, the DEP was studied as a case flow process (see Figure 1). Using this model, the DEP was found to be a complex process involving multiple decision points in which enlistees can exit the system, recycle to an earlier stage in the contracting process,[2] or progress to the next decision point in the contracting process. DEP begins upon contract signing (assuming that the enlistee does not enter active duty immediately upon qualification) and ends when the enlistee enters active duty. Army regulations govern the length of time that different classes (based upon educational status) of enlistees may remain in the DEP.

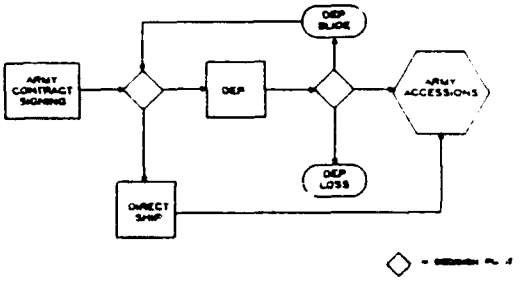


Figure 1. The DEP Contracting Process

Overview of Event History Analysis

In its most generic form, event history analysis deals with the modeling of data that have as their principal characteristic the length of time occurring between some well-defined starting point and an event. An event, in this framework, is a qualitative change that occurs at a specific point in time. In medical and mortality studies the event is often death; in industrial studies the event can be the failure of

[1]This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No MDA903-81-C-0227. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

[2]For example, DEP slide can occur any time after initial contract signing up to the time of entering active duty. Thus, DEP slide requires a new cycle of contract negotiation. For purposes of studying the DEP contracting process, a DEP slide event was defined as a renegotiated DEP enlistment contract in which one or both of the following items were changed at the enlistee's request: the originally contracted military job; the projected active duty date.

a mechanical component. In studying the DEP, the event of interest was failure to accede into active duty. The requirement of a well-specified starting point is important for constructing the event history. For medical studies the event history often covers the period from the onset of treatment to death; industrial studies begin with the initiation of stress tests and end when components fail. For this study of the DEP, the event history covered the period from contract signing to scheduled accession.

One typical characteristic of event histories which creates major difficulties for more traditional statistical approaches, such as multiple regression, is the censoring frequently occurring in event histories. Censoring refers to the condition in which information is lacking regarding either the exact beginning of the history (left censoring) or the end of the history (right censoring). For example, in the testing of industrial components for failure, it may be the case that some of the components stressed did not fail within the period of testing. All that is known is that these components functioned satisfactorily throughout the test period. Many statistical techniques would have to discard these observations as values are missing on the time-to-failure. Event history analysis, on the other hand, is fully able to retain censored cases and utilize the information they contain. In fact, it is largely due to the inadequacies of many statistical techniques in the face of censoring that has led to the development of event history techniques.

During the last fifteen years, for example, many innovative approaches have been developed to accommodate this peculiarity of event history data (Tuma, 1976; Tuma and Hannan, 1978; Tuma, Hannan, and Groenveld, 1979; Kalbfleisch and Prentice, 1980; Anderson, et al., 1980). These methodological advances have come out of a diversity of disciplines including the social and behavioral sciences, demography, biostatistics, and engineering. As a result, there is no single method of event history analysis but rather a collection of related methods that sometimes compete but more often complement one another.

Allison (1984) has outlined several major dimensions distinguishing different approaches to the analysis of event history data as follows:

Distributional versus regression methods. Much of the early work on event history analysis can be described as the study of the distribution of time until an event occurs or the time between events. More recently, the focus has shifted to regression models in which the occurrence of an event depends on a linear function of explanatory variables.

Repeated versus nonrepeated events. Biostatisticians and engineers have tended to emphasize methods for analyzing single, nonrepeatable events (e.g., death, the failure of industrial components). Social scientists, on the other hand, have interests in the study of events like job changes and marriages which can occur many times over the lifetime of an individual.

Single versus multiple kinds of events. In some instances, it may be expedient to treat all the events in an analysis in exactly the same way. For example, in a study of job terminations it may not be necessary to distinguish one termination from another. For other purposes, it may be important to identify voluntary and involuntary job terminations.

Parametric versus nonparametric methods. Biostatisticians have tended to favor nonparametric methods which make few if any assumptions about the distribution of event times. Engineers and social scientists, on the other hand, have gravitated toward models which assume that the time until an event or the times between events come from very specific distributional families, the most common being the exponential, Weibull, and Gompertz distributions. A major bridge between these two approaches is the proportional hazards model of Cox (1972), which can be described as semi-parametric or partially parametric. It is parametric insofar as it specifies a regression model with a specific functional form; it is nonparametric insofar as it does not specify the exact form of the distribution of event times. In this sense, it is roughly analogous to linear models that do not specify any distributional form for the error term.

Discrete versus continuous time. Methods that assume that the timing of event occurrence is measured exactly are known as "continuous-time" methods. When event timing is measured in more general terms (e.g., prior to or following some date or specific occurrence such as marriage, birth of first child) or in large units of time--months, years--it is more appropriate to use discrete-time methods (also known as grouped-data methods).

Use of Proportional Hazards Model to Examine DEP Contracting Process

The particular event history method chosen to examine the DEP contracting process was Cox's (1972) proportional hazards model.[5]

[5] It is impossible, within the confines of this forum, to pursue a discussion of the proportional hazards model proper. Two central features of this model, however, must be at least briefly mentioned. The first is the hazard function $\lambda(t)$. This function estimates (in the context of the DEP) the probability that during a given time period an individual will become a loss. Second, the survival function, $S(t)$, provides an estimate that an individual will remain in the DEP (i.e., not become a loss) during period t. The method of estimation used for calculating the probabilities reported in this paper is the method of partial likelihood.

A number of features of the proportional hazards technique make this a particularly appropriate model to use for the modeling of the DEP. First, the model is explicitly dynamic. Loss and accession probabilities are considered a function of time as well as other factors. The inclusion of a temporal dimension in statistical modeling conforms to our conceptualization of the DEP as a dynamic time-dependent process. An additional characteristic of the proportional hazards technique that makes it attractive for DEP modeling is its ability to use incomplete (i.e., censored) information. This was an especially important feature for this research effort because contracts signed during FY83 did not have a full year (the maximum authorized duration of DEP contract tenure) to fully "mature." Contract accession and loss information was only available through the end of FY83. This meant that contracts signed near the end of the fiscal year were not likely to have become either accessions or losses[4] at the time of analysis (i.e., contracts written with projected active duty dates falling in FY84).

Traditional regression techniques would have to exclude these cases from the analysis, but a proportional hazards model can utilize this incomplete information in its estimation of the survival function.[5] Rather than excluding the information that a case has become neither an accession or a loss through a particular month of tenure in the DEP, this information is used in the estimation of the survival functions.[6]

Proportional hazards analysis also provides statistical estimates and graphic output that are readily interpretable. For example, if one hundred persons were in their fifth month of DEP tenure and ten of these people became losses during the month, the estimated survival function for month five [S(5)] (assuming uncensored data) would be 0.9.[7] That is, persons in the fifth month of DEP tenure have a ninety percent chance of not becoming a loss during the month (alternately, they have a ten percent chance of becoming a loss). The straightforward nature of interpreting the statistical findings from proportional hazards models is often enhanced by presentation of results in a graphical, rather than tabular, form. Figure 2 shows the survival functions for the total of DEP contracts written during FY81-FY83 over their tenure in the DEP pool from month one through month twelve. This figure clearly shows that the probability of survival in any month is easily read by referring to the vertical axis.

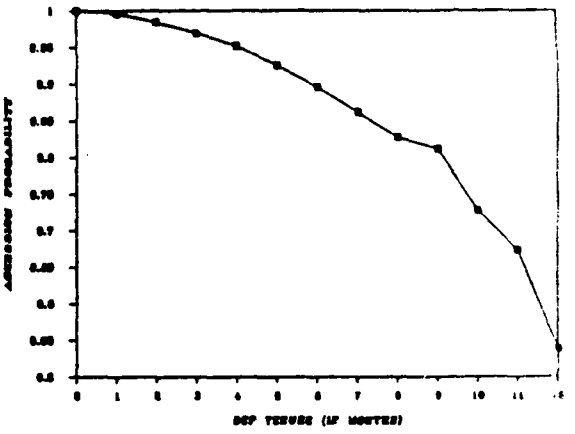


Figure 2. Accession Probabilities by DEP Length

Proportional hazards models, then, provide a useful and interpretable method for analyzing the DEP. In this research, proportional hazards models were used primarily as descriptive rather than analytic techniques. Though it was expected that DEP length would have an effect on accession probability, the precise form of the relationship was not known. More specifically, the effects of DEP tenure on different groups of contractees was unclear. Proportional hazards modeling yielded this information.

Findings

Proportional hazards analysis extended the insights gained during earlier cross-tabular and correlational descriptive investigations of the DEP contracting process. These investigations identified a number of variables significantly associated with contracting outcomes. Proportional hazards modeling explored the interrelations between potential predictor variables, length of DEP tenure, and the probabilities of DEP loss/accession.

Figure 2 plots the survival function across DEP tenure for the entire DEP pool. This figure prompts two comments regarding features generally observed in all the survival function plots constructed. First, survivability decreases as DEP tenure advances. The longer an individual's tenure in the DEP, the greater the likelihood of becoming a loss during the immediate

---

[4] Cases with incomplete information regarding final disposition are termed right censored.

[5] Indeed, these models were developed expressly to deal with censored data.

[6] An additional form of censoring occurring in the DEP is accession. An individual entering active duty during month five was not at risk of becoming a loss during the entire fifth month. Proportional hazards estimation is well able to accommodate this form of censoring.

[7] This example assumes that none of the one hundred persons became accessions during the fifth month and no covariates were used in the estimation equation.

month. Second, survival probabilities generally decline sharply following the ninth month of DEP tenure. In most all the figures presented, the tenth, eleventh, and twelfth months constitute especially critical periods of DEP participation for the total DEP pool. Figure 2, then, clearly demonstrates the dynamic characteristics of DEP tenure and this observation is verified by subsequent subgroup analyses.

Figures 3 and 4 display the survival functions for slide and gender subgroups respectively. In both cases there exists a sharp distinction between subgroups with respect to survivability. Non-sliders consistently have high survival probabilities. Even during their twelfth month of DEP tenure, two of three non-sliders will become accessions [S(12)=.67]. No such optimism can be expressed for sliders as a group, however. Figure 3 convincingly shows the impact of slide events on the likelihood of becoming a loss. The decay in survivability rapidly increases for sliders over time [S(3)=.89, S(6)=.72, S(9)=.53, and S(12)=.15).

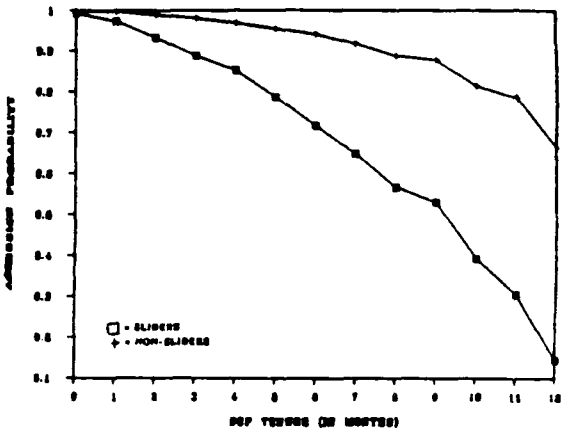


Figure 3. Accession Probabilities by Slide Group

A similar, though not as dramatic, comparison can be drawn between males and females (see Figure 4). Females are at greater relative risk than males over long tenures in the DEP. Up to the third month of tenure, survival functions for these two groups are similar (S(3)=.97 for males and .95 for females). Beyond this month, however, the survivability of females quickly declines. By the sixth month in the DEP, the probability of female contractees' surviving to accession is .79 and in the ninth month the estimate is .58 (compared to .92 and .83, respectively, for males).

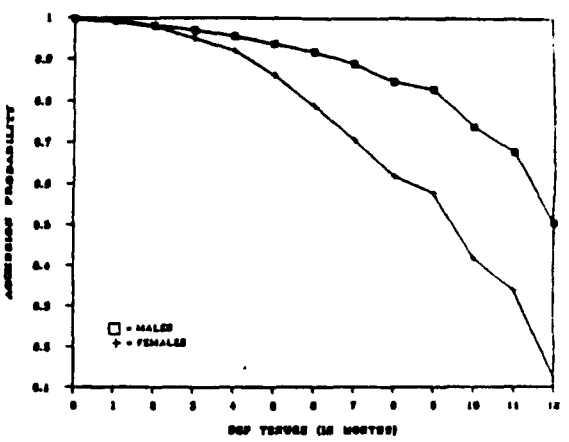


Figure 4. Accession Probabilities by Gender

Figure 5 displays the survival rates for three different educational groups, high school seniors, high school graduates (and above), and non-high school graduates. High school seniors demonstrated the highest survivability over time in the DEP. After nine months' DEP length, their survival rates decreased markedly. Graduates and non-high school graduates showed considerably less ability to survive long tenure in the DEP. Three to four months appeared to be the optimal range of time for these two groups. After that point, the rate of loss grew rapidly among non-graduates and even more so among graduates.

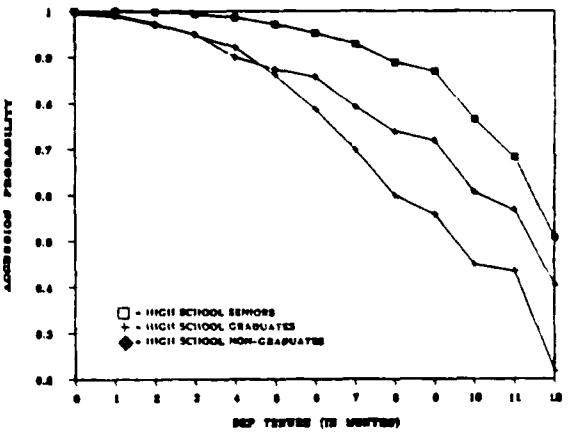


Figure 5. Accession Probabilities by Education Group

In partial explanation of these findings, high school seniors tend to contract for the DEP the summer before or early in their senior year in high school. Training dates are set for the

following summer. Seniors entering active duty at that time have been in the DEP approximately nine months. Senior contractees who do not enter active duty at that time are likely to become losses. Being a full-time student reduces one's susceptibility to becoming a loss due to the long waiting period in the DEP.

Non-high school graduates are permitted a maximum of six months in the DEP (USAREC Regulation No. 601-50). The average contracted length in DEP is shorter, at three months. Non-high school graduates signing up for longer than average DEP lengths face a greater risk of loss. While permitted up to 12 months tenure in DEP, on the average, high school graduates tend to contract for only three months' delay. After three to four months in DEP, the contract survival rate declines. This contractee group behaves more like the non-high school graduates than like the seniors over time in DEP.

Figure 6 plots the survival functions for groups distinguished by their projected active duty date (PADD). Figure 6 shows that there is a significant difference in the probability of survival during DEP tenure depending on the quarter in which one is contracted to enter the Army. The last two quarters of the year display the highest survival functions while the first two the lowest. This finding suggests that accession probability is dependent on time in the DEP and contracted time of accession. This bivariate relation was generally not sustained during multivariate analyses (only non-graduates in the lower mental category were affected by their PADD).

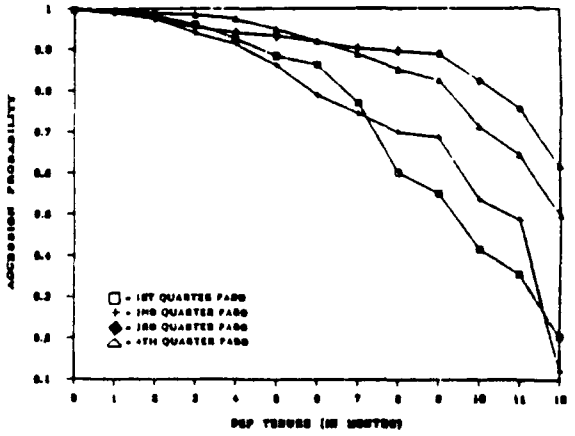


Figure 6. Accession Probabilities by PADD Group

Figure 7 provides an indication why the bivariate and multivariate results are at odds with each other. This graph shows that yearly accessions occur in cycles. Graduate and non-graduate accessions peak generally during the fall and winter months (first and second quarters) while senior accessions take place predominately in the summer. If these differential cycles are viewed in concert with survival functions plotted by education group (Figure 7), it becomes evident that the influence of PADD on survival probabilities is an artifact due to the accession timing of the educational groups.

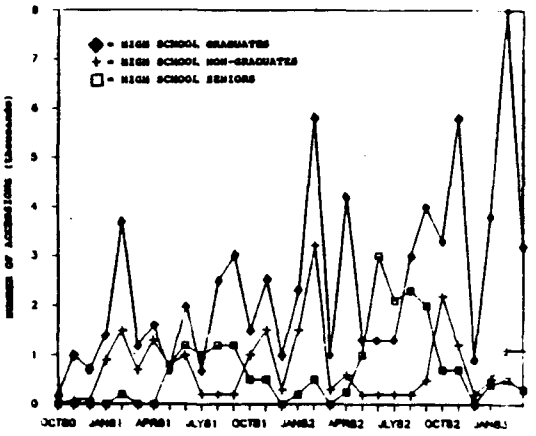


Figure 7. Male I-IIIA Army Accessions by Educational Group: October 1980 Through March 1983*

*Accession totals only cover persons contracting in the period October 1980 through March 1983

Figures 2 through 6 graphically presented the results of survival modeling. To recap, the major findings included the following:

- The probability of becoming a loss increases with DEP tenure;
- Sliders have a much lower survival rate over time than non-sliders;
- Females have lower survival rates than males; and
- Educational groups display differential loss probabilities over time in the DEP.

Discussion

Event history analysis techniques are especially well-suited to performing analyses of military manpower processes. The present application employing a proportional hazards model to examine the DEP contracting process enabled evaluation of the effects of different lengths of tenure in the DEP on different subgroups of contractees. The proportional

hazards model enabled the identification of critical time periods for specific subgroups of enlistees that suggest a need for Army intervention to preserve contract viability.[8]

Characteristics of event history analysis suggest its potential for broader applicability within the military environment; its ability to utilize censored data make it ideal for providing interim status reports on recruiting success, training attrition, and reenlistment rates; and its abillty to examine changes as a function of time fits well within the dynamic military manpower environment. Event history techniques may prove to be of even greater utility in future research efforts by expanding the basic techniques described here to develop causal models of event occurrences based upon multilinear functions of explanatory variables.

REFERENCES

Allison, P.D. (1984). Event History Analysis: Regression for Longitudinal Event Data. Series Number 07-046: Quantitative Applications in the Social Sciences. Beverly Hills: Sage Publications.

Allison, P.D. (1982). "Discrete-time methods for the analysis of event histories," in S. Leinhardt (ed.) Sociological Methodology 1982. San Francisco: Jossey-Bass, pp. 61-98.

Anderson, S., Auguier, A., Hauck, W.W., Oakes, D., Vandaele, W., and Weisberg, H.I. (1980). "Survival Analysis" Statistical Methods for Comparative Studies: Techniques for Bias Reduction. New York: John Wiley & Sons, pp. 199-234.

Cox, D.R. (1972). "Regression models and life tables." Journal of the Royal Statistical Society, Series B34, pp. 187-202.

Freeman, D.L. (1980). CNRC production upgrade management program (PUMP): A concept paper. Arlington, VA: U.S. Navy Recruiting Command.

Hanssens, D.M., and Levien, H.A. (1980). An econometric study of the effectiveness of U.S. Navy recruiting. Los Angeles: Graduate School of Management, University of California at Los Angeles.

Kalbfleisch, J.D., and Prentice, R.L. (1980). The Statistical Analysis of Failure Time Data. New York: John Wiley.

Manganaris, A.G., and Schmitz, E.J. (1985). Impact of Delayed Entry Program Participation on First-Term Attrition. Alexandria, VA: U.S. Army Research Institute.

Morey, R.C. (1983). Management of the armed services delayed entry pools: A comparison of recruiting philosophies and issues. Durham, NC: Duke University, January.

Schumacher, (Captain). (1981). The delayed entry program: A costing analysis or 'What is the optimum size of the DEP?', briefing papers. Air Force Recruiting Command.

Tuma, N.B. (1976). "Rewards, resources and the rate of mobility: A nonstationary multivariate stochastic model." American Sociological Review, 41, pp. 338-360.

Tuma, N.B., and Hannan, M.T. (1978). "Approaches to the censoring problem in analysis of event histories," in K.F. Schuessler (ed.) Sociological Methodology 1979. San Francisco: Jossey-Bass.

Tuma, N.B., Hannan, M.T., and Groenveld, D.L. (1979). "Dynamic analysis of event histories" American Journal of Sociology, 84, pp. 820-854.

U.S. Department of the Army (Headquarters U.S. Army Recruiting Command). Delayed Entry Program, Regulation No. 601-50.

[8] Such decisions, of course, need to be made in view of the tradeoffs between DEP loss and attrition during training. Manganaris and Schmitz (1985) found that the probability of training attrition declines with longer periods of DEP tenure.

Assessing the Accuracy of the AFOQT Quick Score Procedure

Toni G. Wegner and Lawrence O. Short

Air Force Human Resources Laboratory

The Air Force currently uses a major aptitude test, the Air Force Officer Qualifying Test (AFOQT), to assist in selecting officer candidates from those individuals applying to a commissioning program. In its present form the AFOQT is composed of 16 subtests which are variously combined into five composites: Pilot, Navigator-Technical, Academic Aptitude, Verbal, and Quantitative. Both composites and subtests are shown in Table 1. Scores on the AFOQT are reported in terms of percentiles - one for each composite. The percentile scores are converted from raw composite scores based on the number of correctly answered items in each composite. The Verbal and Quantitative composites are used primarily for selection decisions, and the Pilot and Navigator-Technical composites are used for classification into undergraduate pilot and navigator training, respectively. The AFOQT is used as part of the selection process for both Air Force Reserve Officer Training Corps (AFROTC) and Officer Training School (OTS). The current version of the AFOQT, Form 0, has been in operational use since 1981. It is the 15th version of the AFOQT.

Table 1. Construction of AFOQT Form 0 Composites

	AFOQT Composites				
Subtests	Pilot	Navigator-Technical	Academic Aptitude	Verbal	Quantitative
Verbal Analogies	X		X	X	
Arithmetic Reasoning		X	X		X
Reading Comprehension			X	X	
Data Interpretation		X	X		X
Word Knowledge			X	X	
Math Knowledge		X	X		X
Mechanical Comprehension	X	X			
Electrical Maze	X	X			
Scale Reading	X	X			
Instrument Comprehension	X				
Block Counting	X	X			
Table Reading	X	X			
Aviation Information	X				
Rotated Blocks		X			
General Science		X			
Hidden Figures		X			

A problem with the use of the AFOQT was the delay (often one to two weeks) between the time an applicant took the test and had his or her scores reported back to the recruiter. As such, a quick and accurate estimate of an examinee's scores was needed to help eliminate delays in applicant processing. Such measures were developed as reported by Rogers (1985) and called Officer Screening Composites (OSCs). The purpose of the OSCs, then, was primarily to provide a tool to assist recruiters in identifying candidates likely to succeed on the AFOQT.

There are five OSCs, one corresponding to each composite. Each OSC is made up of a representative subset of items contained in the complete composite. The OSCs for the Pilot, Navigator-Technical, Academic Aptitude, Verbal, and Quantitative composites consist of 40, 60, 40, 20, and 20 items, respectively. A hand scoring key is available to score only the items used in the OSCs.

Items to compose the OSCs were selected principally by inspection of item correlation statistics across a group of 37,409 AFOQT examinees. OSC scores are derived as the raw score sum of the correctly answered items in the OSC. Tables were developed to determine full composite expected percentile scores based on OSC raw scores and to define the 90% confidence intervals around the expected composite percentile scores. Rogers reports the following conclusions regarding the use of the OSCs:

> 1. Recruiters can place a high degree of confidence in the prescreening procedure. On any single composite, the expected AFOQT-0 percentile score will fall within the score interval provided in the conversion table for at least 90% of the applicants whose tests are scored using the corresponding OSC.
>
> 2. Recruiters can use the OSC to rank-order applicants from highest to lowest predicted AFOQT-0 scores. The results provide the opportunity for applicant resources to be managed more effectively; recruiters can expedite the processing of high-ranking applicants who are most likely to meet Air Force aptitude entry requirements.
>
> 3. Although OSC scores were originally designed to aid recruiters who process applicants for OTS, they can also be used effectively by test administrators at Reserve Officer Training Corps (ROTC) detachments (1985, p. 8).

Despite the potential value of the OSCs, feedback from Air Force testing officials to the Air Force Human Resources Laboratory earlier this year indicated that the OSCs were being little used in the field. Possible reasons for this lack of use included concerns about the accuracy of expected percentile scores and concerns that confidence interval ranges were too large to be of practical use. For example, the expected percentile score range at the 50th percentile on the Pilot composite extends from 25 to 74. As a result, this research was designed to: assess the accuracy of expected scores and confidence intervals for

existing OSC conversion tables; compare the accuracy of alternative methods for computing expected scores and confidence intervals for the existing OSCs; and if applicable, recommend new procedures for development and presentation of OSC conversion tables. Due to space limitations, this paper will address only the first issue.

Method

A random sample of 22,000 examinees who had tested on AFOQT Form 0 was used in analyses designed to test the accuracy of the expected percentile scores and confidence intervals in the existing OSC conversion tables. The sample was drawn from a file containing all persons who had taken AFOQT Form 0 and who had tested at other-than AFROTC testing sites. This included over 90,000 people with test dates between September 1981 and December 1985, most of whom intended to apply for OTS. This sample was selected because recruiters who process candidates for OTS are the users most likely to benefit from a quick score procedure.

Two types of analyses were conducted to examine the accuracy of the current tables. First, each person's expected percentile score (based on the OSC) and actual composite percentile score were computed for each of the five composites. At each OSC raw score point, the percentage of examinees in the validation sample whose actual composite percentile scores fell within the designated expected percentile score range (based on a 90% confidence interval) was computed. This information was summarized for intervals of about 20 percentiles and across each composite.

The second analysis examined the accuracy of the expected percentile scores using a method designed to maximize the utility of the OSCs. The way the OSC tables are currently used, recruiters find an expected composite percentile score that corresponds to the OSC raw score that has been derived through hand scoring. They are then able to use the confidence interval of the expected value at that score point to estimate the degree of accuracy of that expected score. For the recruiters' purposes, however, the usefulness of being able to predict an examinee's expected percentile score lies in being able to accurately predict whether a score will fall above or below a specified value. For example, recruiters may want merely to identify those examinees who will exceed the percentile minimum scores on the Verbal and Quantitative composites (15 and 10, respectively), or, in better recruiting times, they may want to identify examinees who score above the 50th percentile on these composites.

To assess how accurately the existing OSCs can be used to identify examinees above and below selected percentile cutoffs, hit/miss tables were generated for each composite at intervals of five percentile points. For a specific OSC raw score and a specific actual composite percentile score, the hit/miss table identified the number of examinees correctly identified above and below the selected cutoff (hits), and the

number of examinees whose OSC score incorrectly identified them as above or below the cutoff (misses). Experimenting with the OSC raw scores that were chosen to correspond to the specific actual composite percentile scores made it possible to vary whether hits would be maximized or the different types of errors would be minimized.

Results and Discussion

Table 2 presents the results of the analysis of the accuracy of the confidence intervals for each composite collapsed across all OSC score points. The finding that at least 90 percent of the examinees in the validation sample fell within the 90 percent confidence intervals for each composite offers evidence that the confidence intervals were indeed calculated correctly and are accurate. Furthermore, looking at these numbers at 20 percentile intervals for each composite reveals that this is true not only for the composites as a whole, but across the range of scores. The results across 20 percentile intervals for the Academic Aptitude composite, which are representative of the results of the other composites, are presented in Table 3. The percentage of scores falling outside the confidence intervals is fairly evenly divided above and below the intervals, with a slight tendency for out-of-range scores at lower percentiles to fall more often above the interval and at higher percentiles to fall below the interval.

Table 2. Accuracy of the OSC Expected Score Confidence Intervals (CI) Across Composites

Composite	Percent Below CI	Percent Within CI	Percent Above CI
Pilot	3.3	91.3	5.3
Nav-Tech	3.4	90.6	6.1
Academic Aptitude	4.1	91.5	4.4
Verbal	4.5	91.9	3.6
Quantitative	3.4	92.1	4.4

The fact that the confidence intervals are accurate is of little value, however, if recruiters hesitate to use them. As mentioned earlier, the size of the confidence intervals is so large at some points that recruiters may have little faith in the expected percentile scores derived from the OSCs. By examining the utility of the expected scores independent of their confidence intervals, the second set of analyses sought to determine whether the OSCs can be used with practical value.

Table 4 shows the results of the hit/miss analyses for selected actual composite percentile scores for each of the composites. For the results presented here, the OSC raw scores corresponding to actual percentile scores were selected with the goal of minimizing the error rate of

identifying examinees as acceptable (based on the OSC raw score) when their actual scores did not exceed the selected cutoff (i.e., the -+ column). This is a likely criterion in times of a good recruiting market when it is more important to save money not processing unqualified applicants than to risk the potential loss of qualified applicants. The "hits" and "misses" columns in Table 4 provide the overall success and error rates, respectively. The overall high rate of correctly classifying examinees above and below a cutoff indicates not only that the expected scores are accurate, but that they can have significant practical utility for recruiters who use the OSCs for this purpose.

Table 3. Accuracy of the Academic Aptitude OSC Expected Score Confidence Intervals (CI) at 20 Percentile Intervals

Percentile Interval	Percent Below CI	Percent Within CI	Percent Above CI
0-20	3.4	91.1	5.5
21-40	3.9	91.8	4.3
41-60	4.0	92.0	3.9
61-80	4.8	90.6	4.6
81-100	4.7	92.2	3.0

Table 4. Accuracy of Predicting Selected Actual Percentile Scores for Each Composite

Composite	Actual Percentile Score	OSC Raw Score	OSC Expected Score	Percentage of Examinees[1] ++	--	+-	-+	Hits	Misses
Pilot	25	20	28	69.8	19.4	5.5	5.3	89.2	10.8
	50	26	53	36.8	48.0	9.3	5.9	84.8	15.2
Nav-Tech	25	33	27	67.7	24.0	4.9	3.3	91.7	8.3
	50	41	52	37.4	52.1	6.2	4.2	89.5	10.5
Academic	25	19	26	67.7	25.2	4.3	2.8	92.9	7.1
Aptitude	50	26	51	39.9	52.1	4.4	3.6	92.0	8.0
Verbal	15	6	17	86.0	8.8	2.4	2.9	94.8	5.2
	25	9	30	72.6	20.0	4.4	3.0	92.6	7.4
	50	14	53	44.8	46.6	5.4	3.1	91.4	8.6
Quantitative	10	6	14	82.9	10.0	4.5	2.7	92.8	7.2
	25	9	28	61.9	28.3	5.7	4.2	90.2	9.8
	50	13	54	33.8	56.7	5.8	3.6	90.6	9.4

Conclusions and Recommendations

The analyses presented here confirm the accuracy of both the expected scores and confidence intervals for the existing OSC conversion tables. Furthermore, the use of the tables by recruiters to predict scores above and below a selected cutoff appears to be more straightforward and have more practical utility than the current use of the tables with confidence intervals. It is recommended that the existing tables continue to be used, but that the format be changed to allow recruiters to use them for the purpose of identifying examinees above and below selected cutoffs. The tables have been shown to be accurate for this purpose, and their use in this way will allow recruiters to achieve the intended goal of eliminating processing delays for potentially qualified applicants.

Reference

Rogers, D. L. (1985). Screening composites for Air Force officers (AFHRL-TP-85-2). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

[1]The symbol ++ is the percentage of examinees correctly identified by the OSC raw score as above the actual percentile score, -- is the percentage correctly identified as below the score, +- is the percentage identified below the score that are actually above the score, and -+ is the percentage identified above the score that are actually below the score; hits is the sum of ++ and --, and misses is the sum of +- and -+; fluctuations in the numbers presented here are due to rounding.

The Effects of Reading Burden on SQT Performance

Clay V. Brittain

Paul R. Vaughan

U.S. Army Training Support Center

This study was concerned with the reading requirements imposed by paper and pencil multiple-choice tests used to assess facets of job competence. The problem was addressed here in reference to Skill Qualification Tests (SQTs) used by the Army to test enlisted soldiers. The SQTs are a part of the Army Individual Training Evaluation Program which is designed to evaluate enlisted personnel on common military skills and their job knowledge and performance. SQTs provide an objective and comparative measurement of Military Occupational Specialty (MOS) competence for soldiers in the same MOS and skill level. There are about 850 SQTs and they are scored in such a way as to yield scores ranging from 0 to 100.

The two questions of major interest were: (1) To what extent do the tests measure reading ability? (2) What are the implications for test validity?

The reading demands imposed by SQTs were assessed in two different ways: (1) through the application of readability formulas, and (2) through the judgments of experts from the civilian testing community who critiqued the tests.

Readability Formulas Used

SQT readability was estimated through the application of three formulas: McLaughlin's (1969) SMOG formula, the Flesch (1948) formula as adapted by Kincaid, (Kincaid, Fishburne, Rogers and Chissom, 1975) and the FORCAST formula, a formula developed by Caylor, Sticht, Fox and Ford (1973) of the Human Resources Research Organization (HumRRO).

Table 1 shows readability estimates for 16 SQTs. The SMOG and the FORCAST formulas give very similar estimates of the RGL of the tests. The Flesch-Kincaid formula gave consistently lower estimates of test RGLs. However, when these tests are rank-ordered in terms of RGL, the different formulas give very similar results. If the sixteen tests are ranked from lowest RGL to highest RGL based upon the SMOG formula then ranked based upon the Flesch-Kincaid formula there are only modest shifts. The rank-order correlation is .85.

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Training Support Center or the Department of the Army.

TABLE 1

Test Readability (RGL) as Estimated by Different Formulas

	Readability Formula		
Speciality (MOS)	SMOG	FORCAST	Flesch-Kincaid
Infantryman	8.3	8.6	6.5
Indirect Fire Infantryman	9.2	9.1	7.8
Lance Crewmember	10.1	9.7	8.4
HAWK Fire Control Crewmember	10.6	10.5	8.6
Field Artillery Radar Crewmember	10.0	9.4	7.5
HAWK Launcher/ Mechanical Repairman	10.4	10.1	7.8
Tactical Communications System Operator/ Mechanic	10.1	10.0	8.4
Aircraft Power-train Repairman	8.4	9.4	7.4
Personnel Actions Specialist	11.3	9.7	9.2
Photolithographer	9.3	9.2	6.2
Petroleum Lab Specialist	10.4	10.1	8.2
Air Traffic Control Tower Operator	12.1	11.2	10.9
Hospital Food Service Specialist	8.9	8.8	6.7
Correctional Specialist	9.5	9.8	8.2
Image Interpreter	11.9	10.0	9.1
Interrogator	10.9	10.8	8.2

SQT Readability Gap

In an earlier report the General Accounting Office (GAO, 1977) expressed concern about a "readability gap" (i.e., a gap between the reading levels of military personnel and the difficulty of the materials which they encounter in their careers). We computed SQT readability gap, which was defined as the reading grade level (RGL) of the SQT minus the mean reading score, expressed as RGL, in the target soldier population.

Test readability was estimated through application of the formulas described above. Reading ability scores for soldiers were derived from the General Technical Aptitude score (or GT) of the Armed Services Vocational Aptitude Battery (ASVAB).

In a study conducted by the Air Force Human Resources Laboratory (AFHRL), (Mathews, Valentine and Sellman, 1978), a number of different reading tests were administered to recruits who had also taken the ASVAB, and various ASVAB scores were correlated with reading test scores. In this analysis, GT scores

were found to be a good predictor of reading test scores and a conversion table was developed for translating GT scores to RGL scores. From this conversion table an average RGL for soldiers in any MOS can be derived and a "readability" gap computed for each SQT shown in Table 1.

Figure 1

Readability Gap: Test Readability Relative To Average Reading Level In Target Population

A - Span Based Upon SMOG

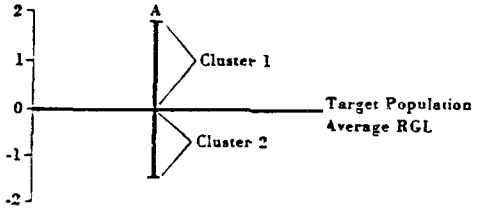


Figure 1 portrays readability gap graphically. The horizontal line in the middle of figure 1 represents the mean RGL of the soldier population in an MOS. The vertical line A represents the range of SMOG estimates of test readability relative to population mean reading ability. Our hypothesis is that the higher the readability of the SQT relative to the mean reading ability of the soldier population tested, the higher the correlation of the soldiers' SQT scores with reading scores. In testing this hypothesis we divided the 16 SQTs into two groups.

TABLE 2

Readability Gap and Correlation Between Test Scores and Reading Scores

Specialty	Correlation Cluster 1	Correlation Cluster 2
Air Defense Mechanic		.64
Interrogator		.53
Petroleum Specialist		.49
Infantryman *Air Defense Crewmember*	.46	.46
Communications Operator		.42
Indirect Fire Infantryman		.41
Aircraft Repairer		.37
Image Interpreter	.32	
Air Traffic Controller	.31	
Photolithographer Personnel Actions Specialist	.29	.29
Lance Crewmember		.28
Food Service Specialist Field Artillery Crewmember	.27	.27
Correction Specialist	.22	

Tests in cluster 1 have SMOG estimated readability levels higher than the mean reading ability of the target population. Tests in cluster 2 have estimated readability levels below the mean reading ability of the target population. Our hypothesis predicts higher correlations between the soldiers' test scores and reading scores for tests in cluster 1 than for tests in cluster 2.

It is clear from an inspection of Table 2 that the correlations are not in line with the prediction. The findings do not support the hypothesis.

Judgmental Assessments of SQTs as Measures of Reading Ability

Over the past several years many of the Army's Skill Qualification Tests have been reviewed by experts from the civilian testing community. Based upon these reviews, we have identified tests of two types; i.e., "reference-based" and "memory-based" tests. It is useful to comment further about reference-based tests.

A reference-based test is one in which the examinee is allowed to use certain references; e.g., field manuals (or extracts from manuals), checklists, computational formulas. The purpose in using such tests is to make the test more faithful to the job. On many technical tasks soldiers in certain specialities are permitted to use, or even required to use, references. In an effort to make the tests more job-like, soldiers have access to the manuals (or extracts therefrom) in taking the tests.

In contrast to reference-based tests, there are other SQTs which make little use of references. These are in specialities in which the tasks performed by soldiers depend more strongly on signs, nomenclature, procedures, applications, etcetera, which are stored in memory. SQTs in such specialities have been labeled here as memory-based tests.

In the judgement of test experts who critiqued the tests, many reference-based SQTs are largely tests of reading ability. This leads to the prediction of higher correlations of reading scores with scores on reference-based tests than with scores on memory-based tests. The correlation coefficients presented in Table 3 bear upon this prediction.

TABLE 3

Correlations Between SQT scores and Reading Scores:

Reference-Based	Memory-Based
.64	
.53	
	.49
	.46
	.41
.38	
.37	
.31	
.29	.29
.27	.27
	.22

These are correlations between reading scores and scores on seven referenced-based SQTs which were judged to be largely tests of reading ability and between reading scores and scores on six memory-based SQTs. These correlations fail to demonstrate any consistent difference between the two groups of tests.

Discussion

The major conclusion to which these results point is that neither readability formulas nor the judgement of experts is a dependable indicator of the extent to which these multiple-choice job tests are measures of reading ability.

In retrospect, the failure of readability formulas in this regard seems reasonable. To say that a multiple-choice test is largely a test of reading ability is to say that an examinee who is ignorant of the content domain of the test can utilize information on the printed page to answer questions correctly. In SQTs this information might be of two types: (1) information in manuals or extracts from manuals to which the soldier can refer in taking a reference-based test and (2) information provided unintentionally in the test items, e.g. length of answer options, subtleties of wording, grammatical inconsistencies, etc. The ability to use information of the second type is generally subsumed under the term "test wiseness." Readability formulas are simply not sensitive to either type of information. However, one would expect that the expert reviewer would be sensitive to the presence of these two features that make a test largely a reading test.

We believe that our test experts were sensitive to internal cues and could discern whether the item could be answered simply by consulting the reference aid. We think that what the reviewers were not in a position to consider was the behavior of soldiers in taking the test. For example, the expert sees that the correct answer is cued by length, by grammatical consistency, by position, etc., but young soldiers taking the test may not be attuned to such cues. The expert reviewer may be able to answer questions on the test by consulting the reference aid, but soldiers may answer the questions from memory, without consulting the reference. So that a test which is a reference-based reading test for the reviewer may or may not be reference-based for the examinees. Let us illustrate the point in relation to two of the tests we looked at. One was for a technical specialty heavily weighted with trouble-shooting tasks. This test made extensive use of references from which test questions could be answered. The correlations between the test scores and the reading scores of examinees was relatively high, that is .64, which supported the judgement of the reviewer that it was largely a reading test.

The other test was for food service specialists. This test also made heavy use of references from which questions could be answered. The correlation here between SQT scores and reading scores was only about .25. We conjecture that the differences between the two correlations was in the manner in which the two groups of soldiers took the respective tests.

The degree of complexity of the technical MOS in the first case here induces soldiers to make extensive use of manuals. The tasks are too demanding for memory alone. Making use of manuals on the job, the soldiers tend to make use of manual extracts in taking the tests. This is to say that they probably took the test as a reference-based test. But our conjecture is that taking the test was quite a different matter for the food service specialist. The questions could be answered by consulting the manual extracts provided by the test, so that the test could have been taken as a reference-based test. But it probably was not.

What we are suggesting, in effect, is that the behavior of the two groups of soldiers taking the test tended to parallel their behavior in doing their jobs. How this impinges on test validity is a question that merits a great deal more study.

REFERENCES

Caylor, J. S., Sticht, T. G., Fox, L. C., & Ford, P. J. Methodologies for determining reading requirements of military occupational specialities. Technical Report 73-5. Presidio of Monterey, CA: HumRRO Division No. 3, March 1973

Flesch, R. A, A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221-233.

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

Mathews, J., Valentine, L., and Sellman, W. Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB). AFHRL-TR-78-82. Brooks Air Force Base, TX: Air Force Human Resources Laboratory, December 1978.

McLaughlin, G. H. SMOG grading: A new readability formula. Journal of Reading, 1969, 12, 639-646.

U.S. General Accounting Office, A need to address illiteracy problems in the military services, FPCD-77-13, March 31, 1977

Recall Versus Recognition in a Mathematics Test

Stephen J. Elliott
RAAF Psychology Service,
Department of Defence (Air Force Office),
Canberra ACT 2600, Australia.

The RAAF Psychology Service employs a selection and classification test battery, known as the Groundstaff test battery, in the recruitment of civilian job applicants to the 90 or so non-commisioned job-types, or 'musterings', in the RAAF. The Groundstaff battery has been essentially the same for 20 years, as the aptitudes measured still show good validity against training performance criteria. In 1980, the requirement to convert units of measurement for some items led to a review of the psychometric features of the tests. As part of this review, multiple choice versions of several tests that previously had been in a free-response format were produced.

The advantages associated with completely multiple choice answer format tests are substantial. Firstly, ambiguity in the test items can be better eliminated, since the options given also help clarify the meaning of the item, and borderline acceptable answers can be excluded. Secondly, the multiple choice items are quicker and easier to manually mark. Finally, test answer sheets can be produced in a Optical Mark Reader (OMR) format, with very substantial manpower savings in terms of test marking and coding of data for computer analysis. The RAAF Psychology Service intends to introduce OMR technology from 1987, and thus aptitude tests will need to be in multiple choice format, where psychometrically feasible.

Despite the above administrative advantages, some concerns were expressed within the RAAF Psychology Service that a multiple choice format, which involves a process akin to recognition of the correct answer to an item, might tax differing aptitude resources than the free response format, which is more akin to recall. The following paper describes an experiment that addressed this issue with one of the Groundstaff tests (a mathematics test, MA2), which was converted to multiple choice format.

METHOD

Subjects and Procedure

Subjects were 182 males and 20 female enlistees commencing RAAF Recruit Training Courses, held at No. 1 Recruit Training Unit, RAAF Base Edinburgh, South Australia, in 1985.

Trainees were randomly assigned to one of two groups by the use of a table of random numbers:

a. Experimental Group (N=100), administered the multiple choice format version of test MA2;

b. Control Group (N=102), administered the free-response version of MA2.

94 of the experimental group and 91 of the control group had previously been tested on the free-response format MA2, as part of the selection phase, which, on average, preceded enlistment by about 6 months (mean: 8.9 months; median: 5.5 months). Enlistees were tested under identical group testing conditions to those prevailing at selection.

Test Development

MA2 is a 28 item test of mathematics knowledge that would be within the scope of a 15-year old Australian secondary student. It samples arithmetic calculations and reasoning, geometry, trigonometry, algebra and logarithms. 12 minutes is given for completion of the test.

Development of the multiple choice format of MA2 proceded upon the assumption that the best multiple choice distractors would be those wrong answers most commonly nominated by testees. Responses that were clearly based on misinterpretation of the question or that were simply variants of other responses were not chosen as distractors. It was believed that this development process would result in a multiple choice test that best paralleled the free-response format test, since testees would still have to generate answers and check them against the likely-sounding multiple choice options.

Hence, wrong answers to each item on the free-format MA2 were tallied, until a tally of 50 responses had been made for each of 4 wrong answers, that fitted the criteria stated above, for that item. These 4 wrong answers, along with the correct answer, were randomly assigned to options A through E for that item.

Analyses

The control and experimental groups were compared on:

a. overall mean score on MA2, using t-tests;

b. change in mean scores on MA2, as a result of retesting, using t-tests;

c. correlation between scores at selection and recruit training (ie test-retest reliability coefficients);

d. Kuder-Richardson Formula 20 inter-item reliability coefficients.

e. item difficulty levels for those reaching each item, using 2x2 Chi square tables;

f. item-total point-biserial correlation coefficients.

RESULTS AND DISCUSSION

Table 1 compares sample statistics for control and experimental groups. Subjects attempting the free-response MA2 scored significantly lower ($t=2.175$, $p<.05$) than did those attempting the multiple choice test. It was also observed that experimental subjects were more likely to complete the test (see Figure 1), and that those completing the test did so, on average, several minutes more quickly, than the control subjects. Thus, it seems reasonable to conclude that the multiple choice test is easier than the equivalent free-response MA2.

Table 1 Comparison of scores on test MA2 by group

Group	N	Mean	S.D.	t-value
Control	102	11.73	5.50	
				2.175*
Experimental	100	13.44	5.62	

* $p<.05$

Table 2 compares initial and retest performances on MA2 as well as the test-retest correlations, for control and experimental groups. The control group's performance does not change significantly as a result of retesting ($t=-.08$), whereas the experimental group's score rises significantly on the second testing ($t=4.03$, $p<.01$). Test-retest correlation coefficients are high for both groups ($r=0.90$ and $r=0.87$, respectively), indicating good test-retest reliability for the free-response form of MA2 and also good alternate form reliability for the multiple choice test. KR20 inter-item reliability coefficients confirm this picture, with coefficients of 0.882 and 0.869 for control and experimental groups respectively.

Table 2 Comparison of changes in MA2 performance as a result of retesting

		Testing					
		Initial		Retest		Test-retest	
Group	N	Mean	S.D.	Mean	S.D.	t-value	r
Control	91	12.66	5.35	12.64	4.95	-.08	0.90**
Experimental	94	12.89	5.30	14.10	5.18	4.03**	0.87**

** p<.01

Item difficulty levels were compared for the two tests, based on those testees who either attempted or skipped the item. Those who failed to reach an item were not used to calculate item difficulty levels, since they might well have passed the item had they reached it. The average item difficulty level for the free-format test was P=0.48, whilst that for the multiple choice test was P=0.52, showing the multiple choice test to be marginally easier. Individual comparisons of item difficulty by Chi square analysis found only two items to be significantly changed (they became easier) by conversion to multiple choice format. These items were of the form:

'What fraction is .... of .... ?', and

'What are the factors of ..... ?'.

It would seem that offering reasonable multiple choice options to these two items helped to dissipate confusion relating to the meanings of those items, since many wrong interpretations of the meaning of the question were excluded by the choice of distractors. The 2 multiple choice items retained acceptable item difficulty and correlated slightly higher with test total score, giving no reason to believe that their psychometric qualities were harmed by conversion to multiple choice. There was minimal difference between the different format tests for item-total correlations (mean rpbi=0.48, for the free-response test, and mean rpbi=0.46, for the multiple choice test).

One difference that emerged between the two tests is clearly shown in figure 1. It can be seen that multiple choice test takers did in fact attempt more questions than did the free-response group, a fact which seems to account for much of the difference in scores between the two test formats.

CONCLUSIONS

Some tentative conclusions can be made from the small-scale study described above.

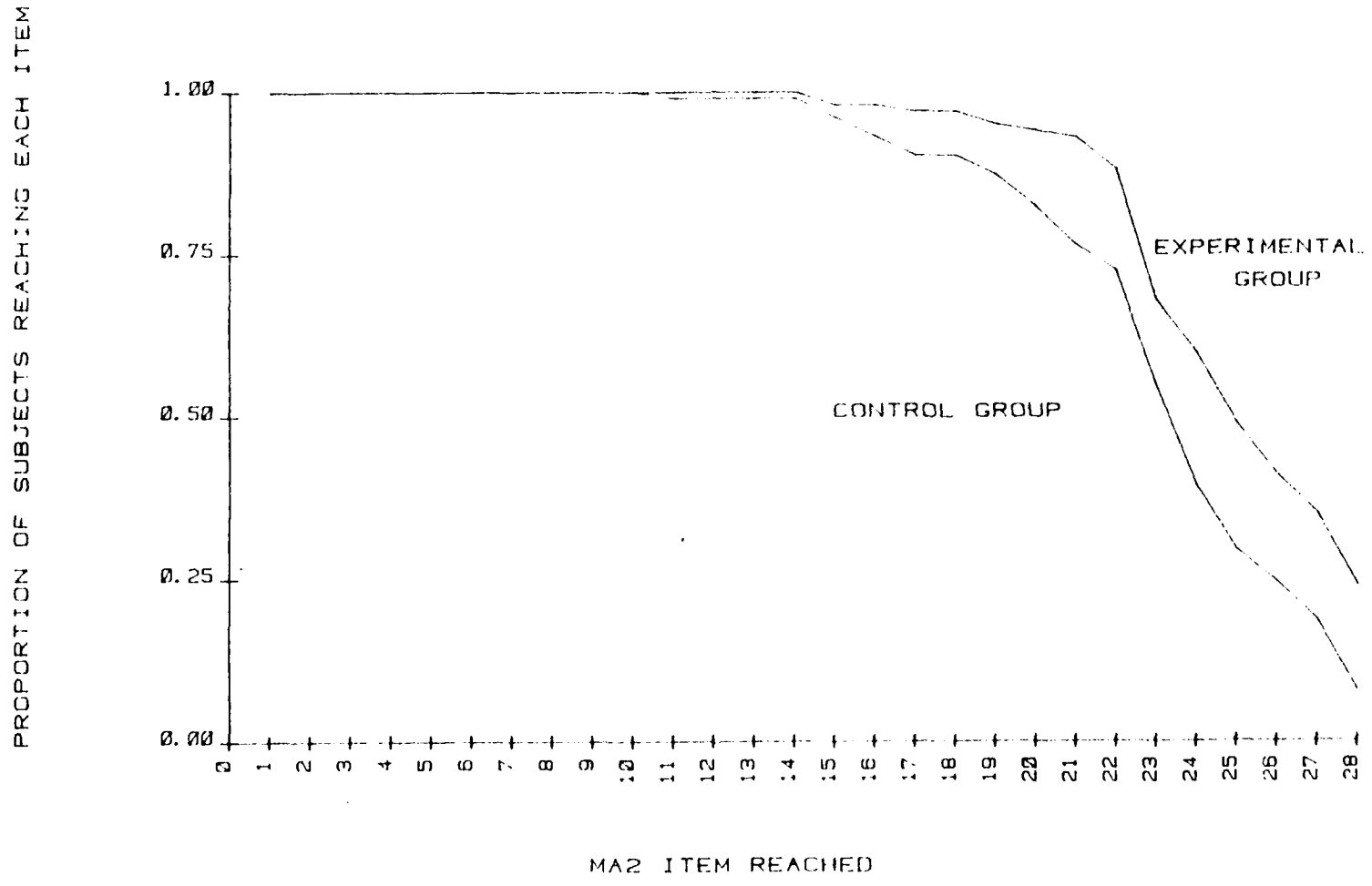
In terms of internal consistency and construct validity, there was negligable difference between the free-response and multiple choice format tests MA2, thus it can be concluded that they are indeed measuring the same aptitudes.

Respondents on the multiple choice mathematics test scored one or two points higher on the multiple choice test, apparently because the optional answer format enabled items to be processed more quickly, but for the most part no more accurately, than did the free-response format.

The free-response format test showed no discernible practise effect due to prior administration of the test at selection. This suggests that those applicants who present for retesting in the selction process and who score better on tests, may have improved their scores by virtue of a genuine increment in aptitude or attainment, rather than because of test familiarity. The issue of test familiarity in retesting is particularly important to the RAAF Psychology Service, and more research is planned to better understand the effects of test familiarity on different forms of testing.

Finally, the approach for used to develop multiple choice distractors seems to have worked very satisfactorily. This author has not seen the technique described in the test construction literature, however, despite - or perhaps because of - its simple logic.

Figure 1 PROPORTIONS OF SUBJECTS REACHING EACH ITEM OF MA2



ENHANCING VALIDITY AND RELIABILITY

IN

PERFORMANCE ORIENTED TRAINING ASSESSMENT

CAPTAIN K. W. VAIL
Command Training Development Centre
Combat Training Centre, CFB Gagetown
New Brunswick, Canada

BACKGROUND

The Canadian Army employs a performance oriented approach in addressing training requirements. The central concept involved in this approach purports that training activities should be reflective of actual job task performance requirements. Actualization of the performance oriented approach is accomplished within the five phases of the Army Systems Approach to Training (ASAT) concept - Analysis, Design, Conduct, Evaluation, and Validation. Each of these five phases has specific associated activities which are reflective of the involvement of varied levels of authority and responsibility involving National Defence Headquarters (NDHQ), Mobile Command (FMC), and individual base training Schools and Agencies. The major control documents corresponding to these three levels are Occupation Specifications (OS's) and Occupation Speciality Specifications (OSS's), Course Training Standards (CTS's), and Course Training Plans (CTP's).

A review of various training programs conducted at the Combat Training Centre (CTC) over the past four years, through formal and informal evaluation activity, revealed the existence of a number of specific deficiencies in all five ASAT phases that were contributing in some way to the weakening of performance assessment validity and reliability. Specific deficiencies identified included:

a. inadequate/imprecise job task performance lists;

b. poorly defined Performance Objectives (PO's) and Enabling Objectives (EO's);

c. practical and theoretical assessment instruments, (Performance Checks (PC's), Enabling Checks (EC's) and written test/examinations) which misaligned with job performance level requirements;

d. poorly defined/non-existent test plans and assessment plans; and

e. personnel with inadequate training and experience in the preparation of control documentation, curriculum design, and program evaluation/validation.

Thus discrepancies encompassed individual and institutional perceptions, personnel training and program development and implementation.

AIM

The aim of this paper is to provide an overview of how the Canadian Army Command Training Development Centre (CTDC) is addressing validity and reliability deficiencies in performance oriented training assessment instrumentation through a formal training package.

TRAINING DESIGN PACKAGE

The CTDC Training design Workshop is designed to train School Standards personnel in preparing training documentation including assessment instrumentation. The package requires approximately seventy (70) hours to complete. The emphasis is practical and the design features call for the practical application of acquired abilities in addressing every day curriculum concerns. Required student participation includes, individual readings, individual and group exercises, and individual and group interaction with each other and the Course Manager (CM).

The package requires individuals to work their way through the ASAT phases from both a theoretical and practical perspective. Personnel are taught to interpret Specifications for the purpose of preparing precise job task performances, to organize the established job tasks requiring training into meaningful job performance related objectives, and to prepare appropriate assessment instrumentation to reflect the objective performance requirement. Other CTS and CTP entities are also addressed and overall the package promotes the enhancement of validity and reliability in the design of performance assessment instrumentation.

JOB ANALYSIS - TASK LIST DEFINITION

The importance of this aspect of curriculum definition cannot be overemphasized. It is only through the careful application of analysis procedures that the job can be adequately defined and training requirements established. The quality of this activity has a direct impact on the development of assessment instrumentation which will ensure the varied job components are validly and reliably measured.

The package stresses the extraction of specific job performance tasks from OSS's. This often presents a challenging task as most Specifications are quite vague and contain a mixture of job, duty and task/knowledge statements. Typical vague statements include, 'perform the duties of a company commander' and 'act as a platoon second-in-command'. Obviously these statements have to be interpreted/defined in terms of specific job performance tasks. The type of question that should be asked is - what are the tasks that are performed when an individual performs the duties of a company commander or acts as a

second-in-command? Specific acceptable tasks could include - 'plan a deliberate attack', 'plan a road move', conduct an advance', etc. Figure 1 illustrates the relationships that should exist in a Job-Duty-Task relationship.

SAMPLE JOB DUTY-TASK RELATIONSHIP

JOB	DUTY	TASK
Infantry Officer	Command a company	Plan a deliberate attack Plan a road move

Figure 1

A second existing problem area involves the format employed in expressing knowledge requirements in Specifications. A typical statement employed is 'possess a knowledge of ------'. The question that should be asked in this instance is - what does possessing the knowledge enable the individual to do? The knowledge translation process is illustrated in Figure 2.

KNOWLEDGE TRANSLATION PROCESS

SPECIFICATION STATEMENT	TRANSLATION QUESTION	JOB PERFORMANCE TASK
Possess a knowledge of algebraic functions	So the individual can?	Calculate gun muzzle velocity.

Figure 2

The definition of a comprehensive set of job performance task statements is the first step in preparing scalars and developing objectives. Failure to define specific tasks descriptive of complete job performances inevitably leads to the formulation of abstract objectives which do not adequately describe required job performances. This in turn impacts on the validity and reliability of the performance assessment instrumentation that will be developed during the Design phase.

STRUCTURING PERFORMANCE AND ENABLING OBJECTIVES

The ASAT concept describes all training activities in a three part objective format. Each objective consists of a Performance statement which states in clear and precise terms the performance required, a Condition statement which defines the parameter under which the performance must be addressed and a Standard statement which describes the quality of overall performance required to indicate job ready applicability. Educational taxonomies, employed in a practical manner, provide the basis for behavioural objective definition. Performance statements, descriptive of the overall job related performance, are expressed at the highest relevant level in accordance with

appropriate domain structures as expressed in the taxonomies of Bloom, Krathwohl and Simpson (Gronlund, 1978). Similarly Standards statements also employ the various taxonomies and are expressive of a variety of equivalent or lower category behavioural performances.

Deficiencies in these two entities are often the result of confusion between describing the job and describing training. While Condition and Standard entities should be shaped by the actual job they are sometimes abstract from the job for various reasons. One of these is due to confusion between the concepts of 'standards' and 'assessment'. As pointed out by Donofrio (1984) a typical Standard statement often appearing in many objectives, particularly at the Enabling level, is - 'must obtain 60% on a written exam'. This is not a Standard of performance statement, rather it is an assessment statement reflecting the acceptable percent required for achievement of an examination. In regards to Conditions, confusion often exists between conditions that affect the job and conditions that address instruction. For example, - 'given a classroom' is not a condition that impacts on a performance. It is a statement that reflects an instructional concern reference the requirement for a suitable area to address the objective.

Conditions and Standards that do not relate to the overall Performance being described provide weak objectives which in turn provide for weakening the validity and reliability of the overall design including the assessment instrumentation. The training package addresses these deficiencies from both a theoretical and practical perspective. Particularly stessed is the unity of the objective in defining a total job performance entity. Figure 3 (next page) presents a sample objective that illustrates the unity concept.

ASSESSMENT INSTRUMENTATION

Practical and theoretical assessment instruments are a vital part of any training program. They provide valuable information to a variety of audiences which includes students, training staff, managers and employers. They are intended to be reflective of practical job performances objectives. When they are not, deficiencies are usually related to misalignments between the intent of the objective performance and the actual performance requirement of the assessment instrument(s).

In PC's, EC's, tests and examinations, the level of involvement is critical. The use of educational taxonomies (Gronlund, 1985) stressed during task identification and objective writing is also promoted in the development of assessment instrumentation. As a result personnel are made aware of the importance of maintaining continuity during assessment instrumentation development, thus enhancing the validity and reliability of developed entities.

SAMPLE PERFORMANCE OBJECTIVE (PO)

1. PERFORMANCE - Perform CH136 non-critical emergency procedures.

2. CONDITIONS - Given:

   a. Kiowa CH136 Operational Checklist (C-12-136-000/MC-000); and

   b. assistance of crew member, if applicable.

3. STANDARD - The pilot must:

   a. address simulated non-critical emergencies in the following systems:

      (1) fuel,

      (2) electrical,

      (3) hydraulic,

      (4) mini-tat, and

      (5) target marking;

   b. accurately assess the situation and respond appropriately in a timely fashion employing correct procedures per Kiowa Operational Checklist;

   c. maintain safe aircraft operating parameters per Kiowa CH136 Aircraft Operating Instructions (C-12-136-000/MB-000);

   d. declare emergency, outlining the nature and progress, using appropriate communication net(s) to an appropriate agency which could include:

      (1) ATC,

      (2) command post (CP), or

      (3) another aircraft;

   e. apply sound airmanship in handling all emergency situations; and

   f. explain and verbally substantiate all actions involving non-critical emergency situation recognition and response IAW the appropriate AOI and CFP 421(1) Kiowa, Volume 1, Manual of Flying Training.

Figure 3

The use of 'test plans' is stressed in developing written test instruments. These plans address the questions of content emphasis and level of involvement requirements. Personnel are taught to address validity and reliability by preparing various types of test questions, by calculating the ease and differentiation indexes for applicable question types and by considering the level of learning requirement as a function of the job task statement within the objective context. The text 'Constructing Achievement Tests' (Gronlund, 1982) is used extensively during this portion of training.

The development of 'checklists' is also stressed in preparing practical performance assessment instruments. Emphasis is placed on the alignment between the job requirement and the 'checklist' in terms of content, procedures and levels of involvement. CFP 9000, Volume 5, Evaluation, provides the basis for the design of checklists.

ASSESSMENT PLANS

At the present time these are not addressed by the training package. The intent of addressing this area will be to further enhance validity and reliability by removing the confusion mentioned previously between 'standards' and 'assessment'. Personnel will be required to prepare an assessment plan which clearly defines what entities are to be assessed, how they are to be assessed, the relative importance of these entities, the approach/method to be employed in assessing them and how to display the total assessment in a meaningful manner that relates to job performance requirements. This should address the current deficiency involving reducing/relating performance to abstract percentages.

CONCLUSION

The importance of effective and efficient training programs designed to maximize student learning while minimizing performance gaps between the training program and actual job requirements are highly desireable from any employers point of view. The CTDC Training Design Workshop provides the Canadian Army with an effective tool useful in preparing personnel to address the need for gap reduction through the preparation of quality assessment instrumentation.

To date the package has been used in training CTC School personnel and civilian contractors responsible for designing training for the Low Level Air Defence (LLAD) program. The results are encouraging and feedback indicates users are being provided with the process understanding and practical experience required to promote the development of high quality training documentation including valid and reliable assessment instrumentation.

REFERENCES

Donofrio, Capt. R.M. (1984, Fall). The Incomplete Performance Objective. Canadian Forces Training Development Quarterly, No. 14, pp 31-40.

Gronlund, Norman E. (1985). Stating Objectives for Classroom Instruction (3rd ed.). New York: MacMillan Publishing Co.

Gronlund, Norman E. (1982). Constructing Achievement Tests (3rd ed.). Englewood Cliffs: Prentice Hall.

Department of National Defence (1978). Canadian Forces Manual of Individual Training, Volume 5 - Evaluation. Ottawa, Ontario: Author.

VALIDATION OF TRAINING -

A PERFORMANCE ORIENTED APPROACH

Major Robert M. Donofrio
Captain Mark W. Thomson

Command Training Development Centre
Combat Training Centre
Canadian Forces Base Gagetown
New Brunswick, Canada

INTRODUCTION

The Canadian Forces Individual Training System (CFITS) provides the structure for the management of individual training in the Canadian Armed Forces (CF). Adherence to the governing principles and processes of the system ensures quantity and quality control of individual training (Canadian Forces Administrative Orders 9-47).

The CFITS is governed by the performance oriented training concept which calls for training the correct number of personnel in only those tasks they are required to perform in their employment. Inherent in this interactive systems approach to training is the need to obtain feedback to verify that the graduates of training have received the correct training in individual skills and knowledge for them to perform their job. In the CFITS, this feedback process is called validation.

AIM

The aim of this paper is to describe a practical performance oriented approach employed by the Command Training Development Centre (CTDC) to validate individual training in the Army.

THE CFITS

In controlling the quantity and quality of individual training, the CFITS assigns responsibility for various activities to different levels of authority. The CFITS interacts with other systems - such as compensation, career, employment, policy and doctrine - at the National Defence Headquarters (NDHQ) level. Responsibilities for other activities are assigned to either Command or Unit levels.

Quantity control is achieved by defining the manning requirements to support the CF operational roles, identifying the personnel to be trained and scheduling training based on predetermined priorities and optimal use of resources. Thus, the right number of trained personnel are available in a timely manner to allow the CF to carry out its operational missions.

QUALITY CONTROL

The quality control function is concerned with the management of what a person needs to learn and how he/she learns it. Quality control is ensured by the application of a logical, interacting series of processes between the identification of tasks to be performed and the provision of trained persons to do those tasks. The five processes or phases of the CFITS quality control are analysis, design, conduct, evaluation and validation (Canadian Forces Publication 9000(1) Part 2, 1978; Donofrio, 1985).

The aim of the Analysis process is threefold in nature:

a. describe military occupations;

b. define training requirements related to military occupations; and

c. provide directions to the training agency in meeting those requirements.

The responsibility for describing the occupations rests at the highest level of authority (NDHQ). The descriptions are based on thorough job analysis and are presented in occupation and occupation specialty documents. The definition of training requirements and the provision of directions to training agencies are produced at the Command level by a board comprised of senior performers, subject matter experts, and training agency and command representatives. The requirements are based on the specifications, are stated as performance objectives and are included in training standards documentation along with the directions to the training agency.

The design, conduct and evaluation phases are the responsibility of the training agency. During the design process, the training agency selects and organizes learning activities to satisfy the training standards identified during the analysis phase. A training plan is produced outlining the instructional strategy and providing specific instruction and direction to trainers. The conduct phase consists of the implementation of the training plan and student learning. The evaluation phase, basically an internal quality control element, aims at verifying the effectiveness (how well the graduates are able to perform to the standards) and efficiency (accomplish training with least expenditure of resources) of training.

The validation process closes the training loop. It aims at confirming the occupation description, verifying if the training requirements have been well defined, determining if the graduates are able to perform operationally and recommending changes to streamline the training. The responsibility for conducting validation rests with the Command which defined the training requirements during the analysis phase (Canadian Forces Publications 9000(6), 1979).

VALIDATION PROCESS

In order to meet the aims of the validation process, the following questions need to be answered:

a. have all the operational tasks been considered for training;

b. are the standards board decisions regarding task training/no training appropriate;

c. are the standards well detailed and do they correspond to the actual operational performance;

d. are the aim, scope and nature of training appropriate;

e. what tasks are deficient in training;

f. is there duplication or unnecessary training;

g. are the graduates employed in positions calling for the training; and

h. are the directions to the training agency appropriate and sufficient.

To answer these questions, feedback from graduates, their immediate supervisors and the commanding officers is obtained. Commanding Officers are canvassed on the more global aspects of training, such as the accuracy of the course aim, the correctness of the scope of training, the appropriateness of the nature of the training, as well as the number of personnel requiring this training to ensure success in operational missions. They are also asked to make recommendations and suggestions to improve the content or format of training.

The graduates and their supervisors provide more specific task and personnel history data. The personal data includes length and type of employment, appointments before and after training, previous training, length of service and other personal history information. The basis for the task data is a list of tasks, drawn from occupation specifications (analysis phase) and training documentation, which the graduates are expected to perform on the job. The following type of information is obtained from the graduates and their supervisors for each task statement:

a. is the task performed;

b. difficulty of task;

c. importance of task;

d. level of skill and knowledge required to perform the task;

e. frequency of performance;

f. whether the task is best learned on course or through on-job training;

g. whether the training adequately prepared the graduate to perform in operational employment; and

h. previous training on the task.

VALIDATION PROCEDURES

The validation activities follow a one year cycle. The courses to be validated are identified during the spring, the data collection instruments are developed during the fall, the data is collected during the January-April period, and finally the data is analyzed and the report prepared during the April-July timeframe. This cycle allows the validation team to visit units during periods of relative stability, after the summer postings and fall work-up and before spring collective training activities.

The validation procedures and activities are presented in detail by Thomson (1985) as he estimates the average effort to complete one validation to be approximately 350 person/hours. The three major phases of the validation process - development, administration, and report - correspond to the three parts of the validation cycle.

During the development phase, the validation plan is prepared, the course graduates and units are identified, units are advised of the upcoming validation, course documentation is collected and reviewed, the task statements are collated, data collection instruments are verified, the paperwork is prepared, and on-site visits planned.

The administration phase encompasses data collection via questionnaires, on-site visits, and follow-up interviews, as well as organization of the collected data. The final report phase includes data interpretation, preparation of the report, and follow-up activities such as clarifying elements of the reports and helping with implementation of the recommendations.

The interpretation of data combines the use of computerized statistical analysis of task statement responses from graduates and supervisors with the open-ended comments of Commanding Officers to arrive at a comprehensive view of the effectiveness of training in meeting operational needs. The report comments on the adequacy of the control documentation, identifies deficient or duplicate training, and presents specific recommendations for addressing shortcomings in the management or quality aspects of training.

CONCLUSION

The CFITS is designed to be an inherently flexible training system able to respond to changing needs, new roles and new equipment. The system is dynamic, providing feedback links among its various processes. The validation process provides feedback on the adequacy of training in satisfying operational reqirements as it closes the training loop; it also serves as a link to other systems active in the CF.

The validation process confirms occupation descriptions and training requirements, verifies adequacy of training in meeting operational performance and provides a vehicle by which training can be improved and streamlined. The validation approach used by CTDC focuses on the application of tasks learned during training to the operational world. The CTDC has used this performance based validation strategy effectively in validating Army individual training; refinements to the process promise even greater effectiveness.

References

Canadian Forces Administrative Orders 9-47 (1977), Individual Training Policy - Regular Force.

Canadian Forces Publications (1979), Canadian Forces Manual of Individual Training; Volume 1, General; Part Two Description. Department of National Defence.

Canadian Forces Publications (1979), Canadian Forces Manual of Individual Training; Volume 6, Validation of Individual Training. Department of National Defence.

Donofrio, Maj R.M. (1985, Fall). SAT - The big Picture. Canadian Forces Training Development Quarterly, No 18, pp 17-34.

Thomson, Capt M.W. (1985, Fall). The Implications of Conducting Validation at CTC. Canadian Forces Training Development Quarterly, No 18, pp 55-60.

Military Values: Structure and Measurement

Arthur C. F. Gilbert, Ph.D., Trueman R. Tremble, Jr., Ph.D.,
Gary M. Brosvic, Ph.D.,[1] and Guy L. Siebold, Ph.D.

U.S. Army Research Institute for the Behavioral
and Social Sciences[2], Alexandria, Virginia 22333

Concern about values in the military arises partly from the fact that values provide a basis for integrating the military with the larger American society and the individual members of the military with their unit and service. Internalization of values consistent with military goals enables more specific operating norms and rules which serve as guides for performance. This internalization of values leads to the behavioral consistency of individuals which is required for overall effectiveness in the diverse missions and conditions of different military operations.

Top Army leaders have acknowledged the importance of "values" by declaring them the Army theme for 1986. The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) supported this theme by a questionnaire survey of the values of Army personnel. The survey was designed to assess the importance of core American and core soldier values. Core American values are those national values reflected by the U.S. Constitution. Core soldier values are those common to soldiers in all good armies and are articulated in Army Field Manual 100-1, The Army. The purpose of this paper is to present the results of a preliminary analysis of the data obtained in the survey. It will describe the structure of the values rated in the survey and describe how different groups of soldiers differed in their endorsement of these values.

Method

A questionnaire was administered to soldiers who used a 7-point scale, anchored from "not at all important" (1) to "extremely important" (7), to rate the personal importance of 50 value statements. The statements were drawn from a number of sources to represent core American values, core soldier values, and other key values. Respondents included 2,114 leaders (pay grades of E5 or above), 2,143 soldiers in units (pay grade of E4 or less), 692 trainees, (soldiers about to complete advanced individual entry training), and 683 new recruits (soldiers who were just beginning training).

The structure of the values was examined by analyzing the responses of leaders because these leaders are the products, as well as the managers and deliverers, of the training which reinforces Army values. The matrix of correlations among the responses of this group to the value statements was factor analyzed by the principal components method using the highest row entry as the communality estimate. The selected factors were then rotated using the varimax method.

---

[1]Now at Glassboro State College, Glassboro, NJ.

[2]The opinions expressed are those of the authors and do not necessarily reflect the position or policy of the U.S. Army Research Institute or of the Department of the Army.

A scale corresponding to each factor was then developed. Items were selected for scale development on the basis of factor loadings and content. A selected item was assigned only to the scale on which it had the highest loading except in one instance where an item was used in two scales. Responses to the items assigned to a factor were summed to compute the corresponding scale score. Total scores on each scale were computed for the leader group, and item-scale correlations were computed and inspected. Split-half reliabilities of the scale were computed using the odd-even approach, and these reliabilities were then corrected for scale length using the Spearman-Brown prophecy formula.

Scale scores for each of the remaining groups of respondents (unit soldiers, trainees, and recruits) were then computed. Step-wise discriminant function analysis and univariate analysis of variance were used to assess the efficacy of the four scales in differentiating among leaders, soldiers in units, trainees, and recruits.

Results

Six factors having eigenvalues greater than unity were obtained, and these accounted for 84.8 percent of the variance. The eigenvalues of these factors were such that the first four factors appeared to provide the best solution. Before rotation, these four factors accounted for 49.7 percent, 14.6 percent, 6.7 percent, and 5.5 percent of the variance, respectively, or a total of 76.6 percent of the variance. After the four factors were rotated by the varimax method, they accounted for 23.4 percent, 21.1 percent, 19.6 percent, and 12.5 percent of the variance, respectively.

In Table 1, the factor loadings of the items loading highest on the four factors and used in subsequent scale constructions are presented. Inspection confirms that core soldier and core American formed part of the structure of the 50 values. The first factor appears to define common soldier values. Items loading heavily on this factor described values closely associated with the official Army values as expressed in Army Field Manual 100-1 and included such values as loyalty to the United States, loyalty to the the Army, willingness to risk life in defense of country, the Army, and being disciplined and courageous in battle. Therefore, this factor will be referred to as "soldier values" because of the loadings of these items.

The second and third factors expanded the structure beyond core American and core soldier values. The items loading highly on the second factor had to do with fair treatment of soldiers, a fair military judicial system, Army's concern for the soldier's welfare, and fast evacuation and good medical care for the wounded. Such items together suggest that this factor captured "fair treatment values." The third factor seems to refer to "life quality values." Items with high loadings on this factor dealt with the importance of a comfortable life, social recognition, wealth and luxury, an exciting life, and happiness.

The items with high loadings on the fourth factor concerned the importance of such American values as freedom of speech, the Constitution of the United States, and voting in elections. Given these items, the fourth factor is labeled "constitutional values".

Table 1

Factor Loadings of Statements Used in Scale Construction on the Four Factors after Rotation

Statement	Factor			
	I	II	III	IV
Loyalty to the United States	(.55)	--	--	.32
Loyalty to the United States Army	(.74)	--	--	--
Loyalty to your unit or organization	(.66)	--	--	--
Fellow soldiers before your own welfare	(.58)	--	--	--
Dedication to serving the United States	(.65)	--	--	--
Commitment to working as member of a team	(.53)	.41	--	--
Dedication to learning job and doing it well	.49	(.50)	--	--
Being disciplined and courageous in battle	(.62)	--	--	--
Standing up for what you believe is right	--	(.46)	--	--
Family security	--	(.37)	--	--
Freedom	--	(.37)	--	.31
Equality	--	(.42)	--	--
A world of beauty	--	--	(.57)	--
International friendship and goodwill	--	.34	(.50)	--
A comfortable life	--	--	(.69)	--
Happiness	--	.36	(.57)	--
Self respect	--	(.49)	--	--
True friendship	--	--	(.43)	--
Social recognition	--	--	(.65)	--
An exciting life	--	--	(.61)	--
The Constitution of the United States	.37	--	--	(.54)
Freedom of religion	--	--	--	(.61)
Freedom of speech	--	--	--	(.68)
Freedom of the press	--	--	--	(.63)
The right of the people to keep and bear arms	--	--	--	(.33)
Being able to vote	--	--	--	(.51)
Responsibility to defend country	--	--	--	(.41)
The Army	(.68)	--	--	--
Army concern for soldiers' well-being	.32	(.53)	--	--
A military justice system which is fair	--	(.56)	--	--
Fast evaluation and good medical care	--	(.55)	--	--
Treating all soldiers fairly	--	(.62)	--	--
Excellent military bearing and appearance	(.53)	--	.36	--
Building and maintaining physical fitness	(.41)	--	(.31)	--
Economic security	--	.32	(.48)	--
Wealth and luxury	--	--	(.62)	--
Living close to friends and relatives	--	--	(.44)	--
Being able to rest/go home when your job is done.	--	.36	(.53)	--
Being able to relax and enjoy yourself	--	.36	(.60)	--

[1]Only factor loadings of .30 or greater are shown.
NOTE: Parentheses indicate that items were used to construct a scale corresponding to that factor.

Table 2

Means and Standard Deviations of Scale Scores for each of the Four Groups and Univariate Tests of Significance

Value		Group					F-ratios
		Leaders	Soldiers	Trainees	Recruits	Total	
	N	2,114	2,143	692	683	5,632	
Soldier	M	51.0	43.1	49.2	50.6	47.7	297.2**
	SD	7.5	11.1	8.9	7.7	9.9	
Fair	M	54.9	52.9	55.6	55.8	54.3	66.5**
Treatment	SD	5.4	7.4	5.1	4.6	6.2	
Life	M	56.8	61.4	63.8	63.9	60.3	137.3**
Quality	SD	11.3	10.6	9.5	9.0	10.9	
Constitutional	M	35.4	33.9	35.7	35.4	34.8	27.3
	SD	6.0	6.8	5.6	5.8	6.3	

NOTE: The possible score ranges for the Soldier Values Scale and for the Fair Treatment Scale were 0 to 70; the possible range for the Life Quality Scale was 0 to 78 and for the Constitutional Values scale it was 0 to 42.

**Significant at the .01 level.

Table 3

Group Classification (Percent)

		Predicted			
Actual	N	Leader	Soldier	Trainee	Recruit
Leader	2,114	77.0	22.9	0.0	0.1
Soldier	2,143	31.6	68.2	0.0	0.2
Trainee	692	49.7	50.1	0.0	0.1
Recruit	683	57.0	43.0	0.0	0.0

Percent of Cases Correctly Classified = 54.9

The items specified in Table 1 were used to develop a scale corresponding to each factor according to the procedures described earlier. Item-scale correlations ranged from .51 to .79 across the four scales. The split-half reliabilities for the four scales, computed by the odd-even item method, were .82 for the soldier values scale, .86 for the fair treatment values scale, .91 for the life quality values scale, and .63 for the constitutional values scale. When these reliabilities were corrected for scale length using the Spearman-Brown prophecy formula, the resulting reliabilities were .90, .86, .91, and .77, respectively, for the four scales.

The means of the four groups of subjects on each scale are shown in Table 2. Inspection of these means reveals that soldiers in all four groups ascribed great importance to American and soldier values. The differences among the four groups were significant beyond the .01 level on each scale. The means of leaders were higher than the means for unit soldiers on the soldier values scale, fair treatment values scale, and constitutional values scale, but unit soldiers had a higher group mean on the life quality values scale. The means of recruits and trainees fell between the means of leaders and soldiers on the soldier values scale and were in that order. The mean scores of recruits were highest on the fair treatment scale and the life quality scale, with trainees having the next highest mean on these two scales. Trainees had the highest mean on the Constitutional values scale, and the mean of recruits was the same as the mean for leaders on that scale.

In Table 3, the classification of the four groups using the discriminant analysis technique is shown. This classification, which is based on all four scales, yields 54.9 percent correct classification for the four groups. Inspection of this table reveals that trainees and recruits were frequently classified erroneously as either leaders or as soldiers in units. When trainees and recruits were removed from the analysis, the percent of correct classification became 72.5 for leaders and soldiers in units.

Discussion

The results are quite important in that four clearly interpretable factors were derived and in that the differences among soldiers and leaders fit well with organizational socialization theory. Army leaders, as institutional leaders, have the responsibility for training and reinforcing values in their soldiers. It is therefore fitting that their expressed ratings load strongly on the soldier values scale. Similarly, it is typical that new recruits and trainees strongly support the general values of the institution of which they have recently become a member. Unit soldiers, on the other hand, while still giving solid support for soldier values, have enough time in the institution to experience its contradictions in the concrete situations of daily life. Hence they view the discrepancy between the ideals and the reality of unit life as an indication, in part, that the general institutional values are of lesser import in the wider scheme of things than they originally countenanced and, in part, that their leaders are demonstrating lesser commitment to institutional values. Day-to-day matters take on greater significance to these unit soldiers; thus, life quality scale scores for them are relatively higher.

It should be kept in mind that these data were obtained using a cross-sectional approach, and further research should involve a longitudinal approach to verify the results reported in this paper. Future research needs

to focus on the extent to which these values are endorsed by members of eif-ferent types of units and in different military occupational specialties. Longitudinal research is needed to assess change of values over time in units and in the Army as an institution. The latter is important to determine the effect of Army organizational changes on the values held by Army personnel. The causal mechanisms which create support for specific values need clearer delineation. While a socialization theory explanation is posited for the results described in this paper, additional theories need to be considered. Research needs to be directed at determining the relationship between value patterns, cohesion, discipline, and unit readiness.

The issue of values is critical to the military from a practical point of view. Expected battlefield opponents are likely to be very highly motivated from religious or political values and belief systems. Our forces must also be strongly motivated based on their values and beliefs to sustain them in combat. Future battle doctrine suggests that our land forces are likely to be fighting in small, widely dispersed units. Strong values will be needed to support the unit members during those periods of isolation. Finally, strong value congruence between soldiers and their leaders should facilitate communication and coordination during combat and training for combat. The value structure presented in this paper forms a foundation for carrying on future research or taking actions to promote positive values in the military.

Moskos' Institutional-Occupational Model: Reliability and Validity of One Operationalization [1]

Gary M. Brosvic [2] and Trueman R. Tremble, Jr.

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

In 1976, Moskos advanced his contention that the military is moving from an institutional (INS) to an occupational (OCC) model. In the INS model, the military is a complete way of life in which military members internalize a purpose (mission accomplishment) transcending individual self-interests. This ethos allows the organization to maintain legitimate and primary behavioral controls over its members both on and off-duty. In return, the military takes care of its own. In contrast, self-interest in the OCC model is ascendant, and the military is viewed in the context of the marketplace where compensation is driven by individual skills and competencies rather than by rank, time in service, and need. The OCC model implies a relatively narrow band over which institutional controls have either primacy or legitimacy. This shift from INS to OCC was viewed as a change in organizational structure which, it was implied, would be reflected in the role orientations or organizational commitments of military members.

As part of a quality of life survey, Stahl, Manley, & McNichols (1977) derived and validated measures of the INS-OCC role orientations for Air Force members. Higher-ranking soldiers were found to have higher INS and lower OCC scores than junior soldiers. INS scores were positively correlated with rank, time in service, and career intent. Negative relationships were found for the OCC orientation.

The U.S. Army Research Institute (ARI) has had a long standing research interest in organizational commitment, the dimensions on which commitment varies, and the implications of such variations for career decision-making and other behaviors. The INS-OCC scales, as derived by Stahl et al., provided ARI a convenient tool to investigate some of these phenomena.

Earlier this year, results of three administrations of the INS-OCC items to Army members were presented (Tremble and Brosvic, 1986a, b). Factor analysis confirmed two factors, with

---

[1] The opinions expressed are those of the authors and do not necessarily reflect the position or policy of the U.S. Army Research Institute or the Department of Army.

[2] Presently at the Psychology Department, Glassboro State College, Glassboro, NJ.

each best represented by the items originally intended to measure the INS or OCC orientations. In all three samples, item loadings replicated the INS factor. However, item loadings on the OCC factor were reverse from expectation in two of the samples.

The INS-OCC items have been included in two more recent investigations. This paper summarizes these recent findings with results reported earlier to examine further the reliability and validity of this operationalization of Moskos' INS and OCC orientations for members of the Army.

Methods

The eight Air Force INS-OCC items, modified for the Army, were included in five surveys of Army personnel from 1981 to 1986. Table 1 describes the five samples of responding soldiers. In each survey, soldiers used 5-point scales, anchored from "strongly disagree" (1) to "strongly agree" (5), to respond to the items. Although the surveys differed in purpose, each contained measures of three variables originally used to validate the INS-OCC scales: rank, career intent, and time in service.

Results

For each sample, a factor analysis (principal components with varimax rotation) of the INS-OCC items yielded two factors accounting for 40-47% of the total variance (see Table 2). Relative magnitudes of item loadings on the two factors were generally comparable across samples.

The INS factor reported by Stahl et al. was duplicated in each sample. With one exception, the four items with high loadings on this factor were as expected. Agreement with these items expressed the ascendancy of mission over personal concerns and needs for greater concern for national security and more discipline and supervision of junior personnel. The major exception involved the 1984 sample of combat arms soldiers for which the item on "concern for national security" had little response variance and loaded negatively on the INS factor.

A factor defined by the expected loadings of four items duplicated the OCC factor in the 1983-1986 samples. In these samples, responses to the OCC items indicated beliefs of too many non-job related demands, greater equity in civilian employment than in military service, and the undesirability of an Army post as a place to live. In the two 1981 samples, however, factor loadings were algebraically opposite from those in the other samples and, thus, suggested a globally positive orientation toward Army work and living conditions. Given its loadings, this factor for the 1981 samples was reverse from expectation and was labeled occupational-reverse (or OCC-R).

Following Stahl et al., items with a negative loading on a factor were reverse scored. A scale score for each factor was then computed as the mean response to the four items with highest

loadings on the factor. Table 2 presents descriptive statistics for the scales: means, standard deviations, and intercorrelations between INS and OCC (or OCC-R as appropriate for the sample).

Correlations generally confirmed the validity of the INS scale (Table 2). As expected, INS was significantly and positively correlated with career intent, rank, and time in service in all samples.

Opposite relationships were expected for OCC. In confirmation, significant negative correlations between OCC and the validating variables were obtained in the 1983 sample. In the 1984 and 1986 samples, correlations were relatively weaker. The expected negative relationships were obtained with rank and time in service in only the 1986 sample. In the 1984 sample, correlations between OCC and career intent and time in service were significant but positive.

For the two samples producing the OCC-R factor, significant positive correlations between career intent and OCC-R scale scores were obtained. OCC-R was also positively correlated with time in service in the sample with greater variation in grades and time in service (1981a).

Discussion

Together with past research, these results generally support the reliability and validity of Stahl et al.'s measures of INS and OCC for Army members. Evidence is especially supportive for the INS scale. In all samples, factor analysis produced a factor defined by the expected loadings. While not necessarily strong, correlations between INS and rank, time in service, and career intent were positive as expected.

The single deviation from expectations tended to be associated with a restriction in variance. In the 1984 sample, the item on "concern for national security" failed to load in the expected direction and magnitude. This item received especially strong and uniform endorsement in the 1984 sample, composed exclusively of members of combat arms units. The other samples were more diverse in both type of unit of assignment and opinions representing this item.

Compared to the INS scale, the OCC scale appears to be somewhat less robust. A factor defined by the expected factor loadings emerged in three samples. In those samples, the expected negative correlations with the validating variables were not as consistently obtained. In the remaining two samples, item loadings on the OCC factor were reversed and suggested a globally positive orientation toward the Army. Given this orientation, correlations between OCC-R and the validation variables were generally positive.

These mixed findings suggest a need for further research on scale development, especially for the OCC scale, and on the variations in the commitments of soldiers to the Army in relationship to sample characteristics. In particular, the stability of the INS factor suggests the possibility that a commitment similar to INS applies to all soldiers. This notion is supported by the frequent finding that "service to country" (or a phenomenon similar to it) motivates enlistment and retention in the Army. However, organizational commitments anchored on the values and norms associated with the OCC orientation may vary more with the characteristics of the soldiers sampled. One characteristic indicated by this research is the career maturity of the sample. The two 1981 samples in which OCC did not emerge were the least mature as indexed by rank composition. It is possible that the importance of professional skills and the opportunities for their reinforcement--values underlying the OCC orientation and represented in the questionnaire items used here--are differentially salient depending on the status of career maturity. This possibility is being explored in research examining the INS-OCC role orientations in soldiers at differing stages in their military careers.

References

Moskos, C. (1977). From institution to occupation: Trends in military organization. Armed Forces and Society, 4, 44-50.

Stahl, M.J., Manley, R., & McNichols, C.W. (1977). Operationalizing the Moskos institutional-occupational model: An application of Gouldner's cosmopolitan-local research. Journal of Applied Psychology, 63, 422-427.

Tremble, T.R. and Brosvic, G.M. (April, 1986). Cross Validation of Institutional and Occupational Role Orientations in the Military. Paper presented at the meeting of the Eastern Psychological Association, New York, NY.

Tremble, T.R. and Brosvic, G.M. (August, 1986). Institutional-Occupational Role Orientations in the Military: Organizational Retention. Paper presented at the meeting of the American Psychological Association, Washington, DC.

Table 1

Demographic Characteristics of the Sample

	Sample				
Variable	1981a	1981b	1983	1984	1986
Grade (% in grade E4 or less)	64%	96%	55%	58%	51%
Career Intent (median)	Not Reenlist/ Undecided	Undecided	Reenlist	Undecided	Undecided/ Reenlist
Time in Service (in years)	2.9	1.2	2.0	2.6	2.5
Number of Respondents	301	1672	1531	416	4807

Table 2

Summary of Results

	Sample									
Rotated Factor Loadings	1981a		1981b		1983		1984		1986	
	INS	OCC	INS	OCC	INS	OCC	INS	OCC	INS	OCC
Mission Accomplishment	.62	-.05	.60	.22	.57	-.27	.62	-.27	.59	-.27
National Security	.67	-.04	.60	.04	.66	-.01	-.43	-.26	.59	.11
More Supervision	.66	.24	.65	.03	.68	-.04	.74	.01	.67	-.19
More Discipline	.73	.07	.66	.03	.69	.02	.75	.07	.74	-.06
Relative Equity	-.12	-.63	.03	-.69	-.08	.73	-.02	.72	-.06	.77
Job Opportunities	-.02	.75	.10	.69	.08	-.67	.24	-.67	.22	-.55
Non-Job Activities	.27	-.49	-.09	-.34	-.01	.62	.09	-.61	-.02	-.62
Post Good	.31	.59	.05	.70	--	--	.31	-.63	-.24	-.62
% Variance Accounted For	45%		40%		46%		47%		45%	
Scale Descriptions										
Mean	3.44	2.84	3.04	2.45	3.36	2.88	3.17	2.96	3.27	2.91
Standard Deviation	.79	.75	.75	.75	.70	.82	.73	.59	.77	.60
Validating Correlations										
INS and OCC	.14		.23		-.05		.14		.12	
Career Intent	.48	.51	.21	.40	.24	-.29	.45	.17	.32	.02
Grade	.39	.02	.26	.06	.39	-.22	.38	.02	.25	-.12
Time in Service	.47	.32	.19	.07	.33	-.14	.42	.12	.31	-.08

Note. The "Post Good" item was not included in the 1983 administration. Correlations greater than 0.05 were statistically significant, given sample sizes.

*MAR-VALS: A MICROCOMPUTER-BASED EXTERNAL EVALUATION SYSTEM

by

Robert H. Kerr, Commander, Canadian Forces

&

Edward G. Barnett, Lieutenant(N), Canadian Forces

Phone Contact: (CSN)447-4881/(Centrex)1-902-427-4881

*Mar-Vals: Maritime Command Validation System

Introduction

An external evaluation mechanism is a critical requirement if any systematic approach to training is to be effective. Graduates of training programs and their supervisors must be asked to comment on the quality of training as it relates to the graduate's ability to perform job oriented tasks. Several authors have stressed the need for job related performance measures to achieve this goal (Butler, 1972, pp. 167-168; Denton, 1977, P. 26; Forman, 1980, p. 51). These evaluations must be structured in such a way to identify not only on-job related training problems but on-job related non-training problems as well. Notable performance technologists such as Mager (1970) and Gilbert (1978) have clearly identified that such non-training problems exist in the job environment. If precise solutions are to be developed to reduce the 'performance gap' then both types of problems must be considered. Failure to do this may result in misdirected, non-cost effective solutions which widen this gap.

Maritime Command (MARCOM) has identified the external evaluation requirement in its systematic approach to training Canadian Naval personnel. This process is referred to as "Training Validation" in that system. As new technology is being introduced into the naval job environment, training mechanisms are continually trying to keep pace. MARCOM sees training validation as the only solution to ensuring that new technology transition periods do not create a job-training imbalance. This paper will review a tailored approach that has been developed by MARCOM to address this need.

Background

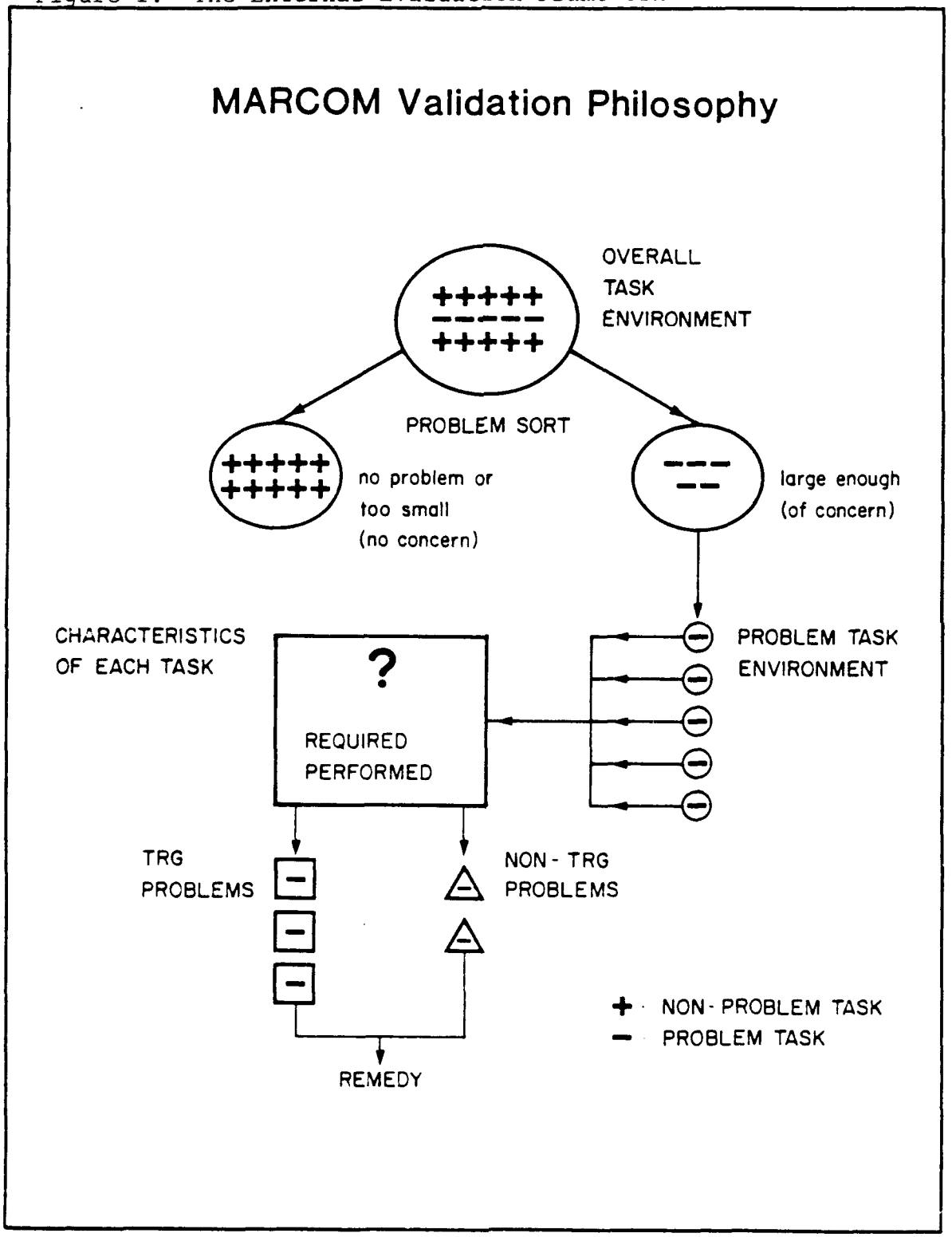
As early as 1972 MARCOM was developing a validation system. This system was not implemented on a large scale basis as naval operational requirements and limited financial/personnel resources precluded its full development. These initial efforts however produced a performance oriented algorithm which was further refined as the validation process evolved. This question-based algorithm was directed in questionnaire format to both graduates of courses and their supervisors. Kerr and Tyerman (1984) provide a detailed discussion of this particular approach.

The early 1980s saw a growing increase in naval technology resulting in a complete reorganization of the Canadian naval occupation structure. The training system that supported this reorganization was required to adjust accordingly. A standardized validation process was seen to be a critical component to ensure that occupational task requirements were reflected as closely as possible in an efficient and effective training structure. MARCOM and its principle training unit, Canadian Forces Fleet School Halifax, were both committed to that end.

The validation philosophy that would encompass such a system had to discriminate training problems from non-training problems for any specific problem task environment as seen in Figure 1. The validation algorithm used in the data collection process would review the characteristics of each task by seeing if they were required and performed for a specific occupation level. Training and non-training problems would be identified within the review and appropriate remedies suggested.

A validation system was proposed in 1985 which would strive to incorporate the best of the previous efforts in that area. To keep

Figure 1. The External Evaluation Framework



pace with the new technologies, a computerized validation strategy was proposed which would provide timely, user-friendly data outputs. The following overall development guidelines and benefits were considered:

1. minimum personnel requirements - fewer personnel required in the validation process to prepare and process data collection questionnaires;
2. machine processed data - resulting in faster and more flexible data analysis with fewer data transfer errors;
3. consistent validation formats - to reduce validation planning and implementation times and to facilitate the training of personnel on the system;
4. more objective outputs - producing clear, standardized outputs to provide a better decision-making vehicle for senior trainers; and
5. future portability - providing a system that would lend itself to a more portable approach with increased use of computer hardware and software by the navy.

System Components

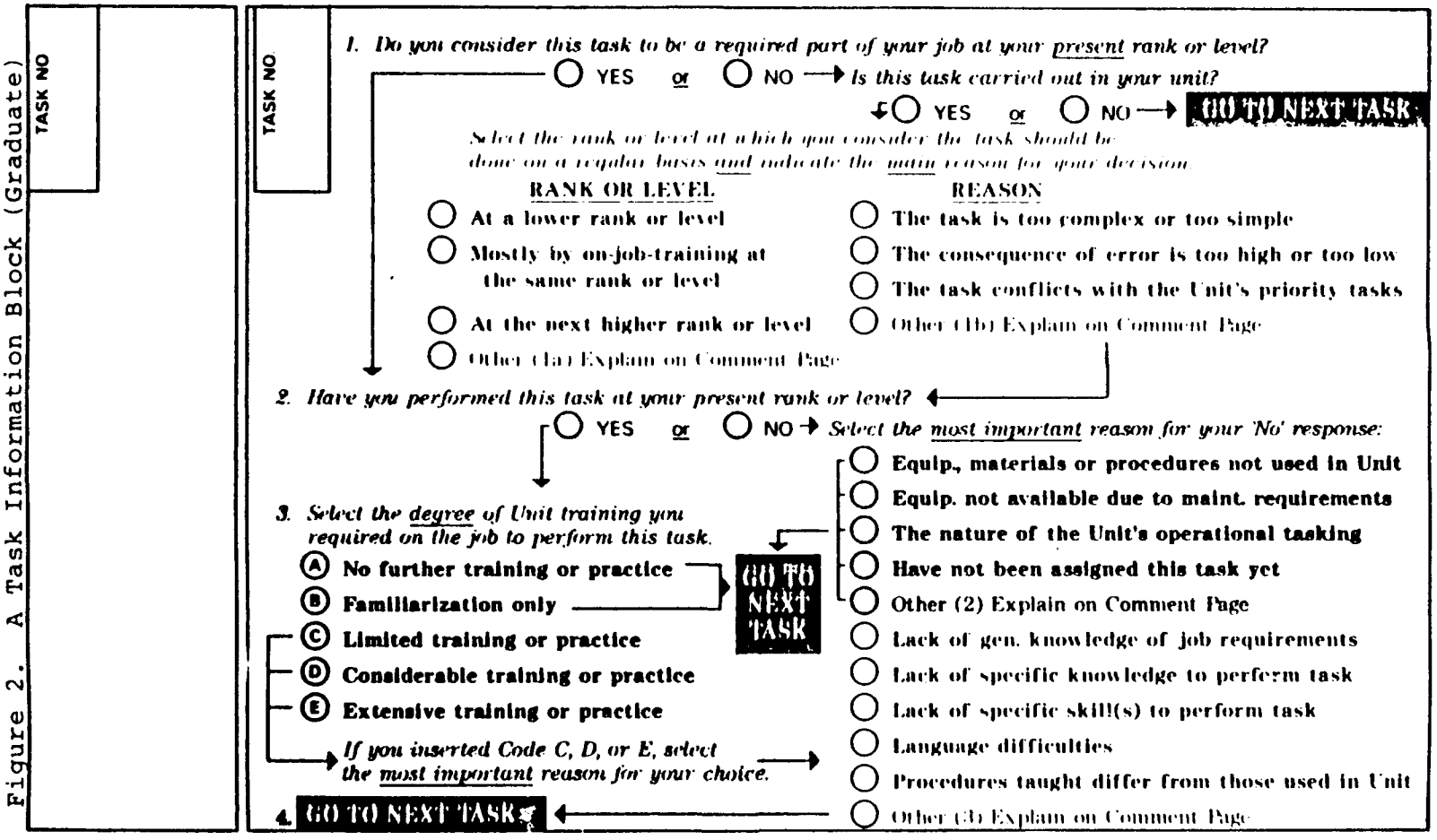
A commercial contract was established in early 1986 to provide such a tailored, computerized validation system. The system was designed such that data could be collected for any particular naval occupation using optically scannable questionnaire booklets. These booklets include a biographical information sheet, an additional comments sheet and task data collection sheets containing the validation algorithm. A maximum number of eight tasks can be overprinted on each data collection sheet, each task being contained in an information block which presents the entire validation algorithm (see Figure 2). Separate data collection sheets were developed for both graduates and supervisors using the same format.

Mailing labels are generated from a central Canadian Forces information system, identifying any specific set of graduates with the names and addresses of their current employers. Questionnaire booklets are forwarded by mail and returned for processing on an optical mark reader. The data are recorded onto a 9-track tape which in turn is placed in a tape reader. The tape reader transfers the data to the microcomputer which then analyzes it using an adapted statistical software routine, "SPSS" (Statistical Package for the Social Sciences). This routine generates specific reports which are readily produced with a computer printer. These reports provide cumulative data summaries for decision-making purposes as well as more detailed individual information if more rigorous analyses are required.

The current system trials have involved the following resource requirements:

1. Hardware -
   a. 1 - Optical Mark Reader System,
   b. 1 - Microcomputer System (640K with a 10-Megabyte Hard Disc),
   c. 1 - 9-Track Tape Reader, and
   d. 1 - Printer;
2. software -
   a. 1 - SPSS Analysis Program, and
   b. 1 - OMR Program; and

Figure 2. A Task Information Block (Graduate)



3. personnel -
   a. 1 - Training Development Adviser (for task development, data processing, analysis and overall project coordination),
   b. 1 - Subject Matter Expert (for task development assistance), and
   c. 1 - Administration Clerk (for label procurement, typing, mailing routines and data processing assistance).

Decision Path Framework

The algorithm in Figure 2 represents the core of the data collection process. A similar routine exists for the supervisor, though the questions are worded for his perspective. This information block requires the respondent to make several decisions as he progresses through the algorithm. Seven decision paths or 'tracks' are possible as the respondent answers the questions for each task. Six of the decision paths are further amplified with qualifying information or 'discriminators' after each question. These discriminations provide valuable cues as to whether or not a problem exists for a particular task and if it is training or non-training related.

At one end of the validation spectrum an individual might feel a task is not required and not carried out in his unit. If that respondent received course training to perform such a task a non-training, management problem may exist. This task may not be carried out when it should or may be included as an unnecessary job requirement.

Another individual may feel a task is not required at his level but that it should be carried out at another. This decision path allows the individual to qualify the level at which he feels it should be done with a specific reason to support that choice. This again represents a non-training problem. If the respondent has performed the task he may indicate that not only is the task at an inappropriate level but that he required limited to extensive training to perform it. This is further qualified by indicating a reason for the degree of extra training chosen. A training problem is indicated here that relates to the non-training problem identified at the start.

A task may be perceived as required but not performed. The respondent must then qualify why that is the case. If equipment or procedures are not used in the workplace a non-training, resource related problem may exist. The ideal case is where a task is seen to be required and performed with no further training or practice required.

This decision path allows a vast amount of data to be accumulated within a standardized framework. Graduate and supervisor responses can be easily compared as the decision path structures are the same for both. The algorithm's most valuable asset is its ability to clearly distinguish training from non-training related problems.

Formative Evaluation

A small group trial of this system was conducted in September 1986 at Fleet School Halifax to evaluate the face validity and useability of the data collection instruments and to evaluate the overall utility of the system's hardware/software integration. Twelve graduates and eleven supervisors participated in the trial which involved validating a junior management course. Respondent feedback indicated that minor clarification was required as to how the questionnaires should be

completed and that a more precise set of instructions should be provided with the questionnaires. Several respondents indicated the value of having the validation algorithm printed beside each task for ease of task reference and form completion. The system hardware/ software components worked well, though it became clear that larger data sets would involve a requirement to increase the microcomputer's storage capability in the future. Data reports were produced as planned with only minor format amendments being required.

Future Perspective

The authors intend to continue trialing the current system until they become more familiar with its components and complete any first level system refinements. Ongoing validations are scheduled for mid-1987 in conjunction with a planned MARCOM validation policy. A parallel project is currently underway which will use video display terminals to generate the validation algorithm. It is the authors' contention that as the naval environment increases its use of computer technology that validation input systems can be wholly developed and implemented via this medium. This area will be explored further as validations continue to be conducted and an understanding of the process increases. Both authors look forward to assisting the Canadian Navy in pursuing an active validation program and exploring more effective and efficient methods to achieve that end.

References

Butler, F.C. (1972). Instructional systems development for vocational and technical training. Englewood Cliffs: Educational Technology.

Denton, J.J. (1977). A field tested evaluation model to assess a CBTE program. Educational Technology, 17(3), 23-27.

Forman, D.C. (1980). Evaluation of training: present and future. Educational Technology, 20(10), 48-51.

Gilbert, T.F. (1978). Human competence. New York: McGraw Hill.

Kerr, R.H. and Tyerman, D. (1984). External evaluation revisited - an experimental update. Proceedings of the 26th Military Testing Association, 1984, 2, 763-768.

Mager, R.F. (1970). Analysing performance problems. Belmont, Calif.: Fearon.

How Army Veterans View Their Military Experiences

Melvin J. Kimmel and Glenda Y. Nogami

U.S. Army Research Institute for the Behavioral and Social Sciences[1]

The Secretary of the Army considers recently separated veterans to be a valuable Army resource. He believes that their Army experiences, the attitudes they have formed as a result of these experiences and their portrayal of these experiences and attitudes to others will affect the civilian world's perception of the military. Because of their potential significance, the Secretary tasked the Army Research Institute to survey these veterans regarding their Army experiences and willingness to continue to identify with the Army.

The Secretary was especially interested in the attitudes and opinions of a particular group of recently separated Army veterans: enlisted soldiers who left active duty after successfully completing one term of service. Consequently, much of the analyses, reports and briefings to date have focused on this group of one-term Separatees. In general, the results have been very encouraging. Most one-term Separatees report that their experience was rewarding, especially for the self-development opportunities it offered, and indicate a willingness to continue to work with the Army as civilians. (Kimmel et al. 1986).

The question remains, however, as to the generalizability of these findings to other separation groups. For example, Gade et al (1984) report that those who attrite before completing a full service term do not have the same motives and influence sources as one-term Separatees, and it may be that their Army experiences and attitudes are much more negative. Those who successfully complete more than one term of service might be expected to be even more positive toward their Army experience than one-term soldiers and more willing to continue to identify with the Army. The present effort was designed to test these hypotheses.

Method

The population of interest was defined as enlisted Army personnel who separated from active duty between October, 1981 and September, 1984 (N=333,481). Stratified random samples were drawn for each of four separation groups: enlisted soldiers who attrited before completing one full term (one-term Attritees); enlisted soldiers who successfully completed one term of service (one-term Separatees); veterans who successfully completed more than one term but left before retirement (Mid-careerists); and those who retired after 20 or more years of service (Retirees). The one-term Separatee sample was made intentionally large relative to the other separation status groups because of the Secretary's special interest in this group of veterans. However, the random samples for the other separation groups were sufficiently large and representative to allow for generalizations to their respective populations as well.

---

[1]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

The survey instrument consisted of 117 forced-choice and 27 open-ended items and took approximately 30 minutes to complete. It was designed to provide information on the past service experiences of Army veterans, their attitudes toward the Army, the perceived impact of the experience on their current lifestyles, and their willingness to continue to identify with the Army. The survey was mailed to potential respondents with an accompanying cover letter signed by the Secretary of the Army. Those who did not return the initial survey after two weeks were sent a reminder post-card and a second mail-out. When the multi-wave survey mailings failed to obtain a response, attempts were made to contact and interview the veterans by telephone. Response rates varied by separation status. Retirees showed the highest response rate (82%, N=412), followed by the one-term Separatees (50%, N=2,566), Mid-careerists (47%, N=515) and one-term Attritees (43%, N=715). Because the percentages of returns from the different ethnic and gender groups sampled were quite similar to their respective population percentages, we believe the data are a representative of the population as a whole.

The data to be presented were weighted (adjusted) to more accurately reflect responses that would have been obtained if the entire population had been surveyed. This was done to compensate for unequal sampling rates, reduce sampling error, and dampen the effects of nonresponse bias (Cochran, 1977).

Results

The Army Experience. As shown in Figure 1, the Army was a positive experience for the majority of respondents within each separation status group.[2] Most of the former soldiers indicated that they were proud to have served their country, found the Army both satisfying and valuable, and would join again if they had it to do all over. While the experience was generally positive, regardless of separation group status, there were differences in relative degree. As might be expected, career Retirees showed the highest percentages of positive attitudes and one-term Attritees the smallest percentages on each of the four measures, although over 50% of all status groups were positive.

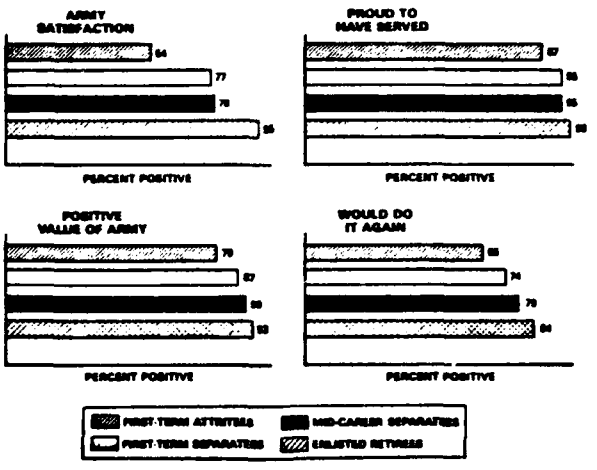


FIGURE 1. Army Experience Attitudes by Separation Status

[2]Each forced-choice attitude scale consisted of 2 degrees of positive and 2 degrees of negative. Percentages responding in the positive direction are presented.

When we asked respondents who considered their experience to be valuable to explain how it was valuable, we found that over 75% in each separation status group stated the Army had enhanced their self-pride, helped build their self-confidence, enabled them to become more self-disciplined, and improved their leadership skills and ability to work with others. A slightly smaller, but still a large percentage of those who found the experience valuable within each separation group (over 70%) said that the Army also had allowed them to establish their independence, enhanced their ability to make friends, and helped them gain a respect for authority and an openness to new ideas. Again, the relative percentages stating that the Army had a positive impact on these self-development characteristics differed by separation status. Career Retirees had the highest percentages stating that the Army had had a positive impact on these self-development characteristics, with percentages ranging from 81% for the Army's impact on ability to make friends to 95% - 97% on self-confidence, self-pride, ability to work with others, and leadership ability; while the percentages for one-term Attritees were lower relative to the other separation status groups, ranging from 71% for ability to make friends to 86% who said the Army had enabled them to become more self-disciplined. Figure 2 summarizes these findings.

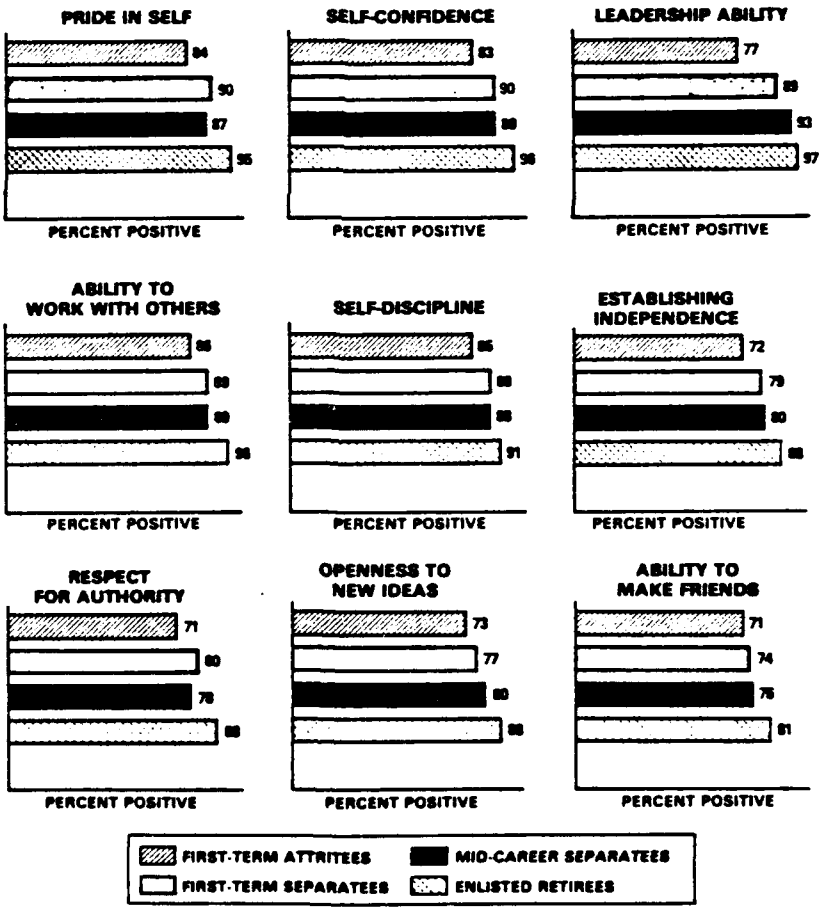


Figure 2. Impact of Army on Self-Development for Veterans Who Considered the Army Experience Valuable

While the veterans saw the Army as a good place for self-development, they did not consider it especially helpful in developing strong family relationships. Only 30% of all married one-term Separatees and Attritees, 38% of the Mid-careerists, and 55% of the Retirees said that the Army had had a positive impact on their relationship with their spouse. The percentages reporting a positive impact of the Army on relationships with their children also were relatively low: 30% for one-term Attritees, 36% for one-term Separatees, 36% for Mid-careerists, and 57% of the Retiree group.

One aspect of the Army experience that received mixed reviews was job skills training and development. For those who felt the Army was a valuable experience, 89% of the Retirees, 79% of the Mid-careerists, 71% of the one-term Separatees and 56% of the one-term Attritees said that the Army had a positive impact on learning specific job skills. Not receiving job skills training was seen as a definite deficiency of the Army by those 13% of the total sample who rated their Army experience as not valuable. Only 18% of the one-term Attritees, 24% of the one-term Separatees, 29% of the Mid-careerists and 47% of the Retirees in this group indicated that the Army was helpful for learning specific job skills.

Veterans as Army Alumni. Another indication of how veterans feel about their Army experience is found in a series of questions concerning their willingness to continue to identify with the Army. When asked if they would like to join an association for former Soldiers, 73% of the Retirees and Mid-careerists, 62% of the one-term Separatees and 55% of the one-term Attritees replied positively; and over 90% of the Retirees and Mid-careerists, 82% of the one-term Separatees and 72% of the Attritees indicated an interest in receiving an Army newsletter or magazine.

Their responses also indicated a willingness to act as Army ambassadors. Ninety-five percent of the Retirees, 86% of the Mid-careerists, 83% of the one-term Separatees and 73% of the Attritees said they speak positively about the Army when talking to friends and acquaintances; while 66% of the Retirees and Mid-careerists, 56% of the first-term Separatees and 52% of the Attritees expressed a willingness to help recruiters identify potential enlistees. When asked about their current level of involvement with Army recruiting, over 80% within each separation group said that they had spoken to at least one person about joining the Army, and 63% of the Mid-careerists, 60% of the Retirees, 53% of the one-term Separatees and 49% of Attritees said they would be willing to talk to groups of high school students about the Army.

What do they tell people about joining the Army? Again, we see very positive results. Less than 5% of the Retirees, Mid-careerists and one-term Separatees and only 8% of the Attritees discourage potential recruits. Sixty-eight percent of the Retirees, 53% of the Mid-careerists, 46% of the first-term Separatees and 37% of the Attritees emphasize the opportunities available in the Army and encourage them to join. The remainder suggest that the potential recruits find out more about the Army and decide for themselves.

Discussion

The results suggest that the Army was a rewarding experience for the vast majority of recently separated enlisted soldiers - regardless of separation status. Separation status differences were found on most items. However, in almost every instance, the difference was only in degree of positive feelings expressed. Even those people who left active duty before completing a full term found the Army a positive experience, although the percentages were somewhat smaller than those for the others separation groups.

Taken as a whole, these results paint an optimistic picture of the future for both the veterans and the Army. The Army has provided these former soldiers with the opportunity to develop personal qualities that will benefit them as civilians. These veterans, in turn, will continue to serve the Army as goodwill ambassadors. Instilled with a pride in serving this country, traditional Army values, and exemplary personal qualities, they are able to show the civilian world first-hand, through their words and deeds, what the Army has to offer. They not only enhance the Army's image, but their own as well.

As might be expected, some elements of the Army experience were not rated as especially rewarding; particularly, issues related to job skill development and the Army's impact on family relationships. With respect to the former, it should be emphasized, that negative ratings were largely attributable to that small percentage of our total sample (only 13%) who said their overall Army experience was not valuable. With respect to the negative impact of the Army on family relationships, the Army has begun to address this issue vigorously in recent years through the Army Family White Paper (1984) - which articulates an Army philosophy for developing families of excellence, the Army Family Action Plan (1983) - which provides a management tool for implementing this philosophy, and the many research efforts and program changes resulting from these documents (as described in The Army Family Action Plan III, 1986). If the survey project described in this paper was replicated in two or three years, it would not be surprising to find that the Army's impact on family relationships was rated as positively as its impact on self-development.

References

Cochran, W. G. (1977). Sampling Techniques, Third Edition. New York: Wiley.

Chief of Staff, U.S. Army (1983). White Paper: The Army Family. Washington, D. C.: Headquarters, Department of the Army

Community and Family Policy Division, Human Resource Development Directorate, Office of the Deputy Chief of Staff for Personnel (1984). The Army Family Action Plan. Washington, D. C.: Headquarters, Department of the Army.

Community and Family Policy Division, Human Resource Development Directorate, Office of the Deputy Chief of Staff for Personnel (1986). The Army Family Action Plan III. Washington, D. C.: Headquarters, Department of the Army.

Gade, P. A., Elig, T. W., Nogami, G. Y., Weltin, M. M., Hertzbach, A. and Johnson, R. M. (1984, May). Motives, Incentives and Key Influences for Enlistment, Reenlistment, and Attrition in the U.S. Army, Proceedings of the Second Symposium on Motivation and Morale in the NATO Forces, Brussels Belgium, 227-250.

Kimmel, M. J., Nogami, G. Y., Elig, T. W. and Gade, P. A. (1986). The one-term soldier: A valuable Army resource, Soldier Support Journal, 13(5), 4-7.

MEASURING ATTITUDES TO SERVICE IN THE BRITISH ARMY

by

Valerie Morris

Army Personnel Research Establishment

INTRODUCTION

1. The Manpower Studies section of the Army Personnel Research Establishment (APRE) is concerned with, amongst other things, retention issues in the British Army. We are at present running several projects which are designed to assess attitudes of officers and other ranks to Army life and to investigate why people choose to leave prematurely.

2. One of these, the Continuous Attitude Survey (CAS) for soldiers, has been running since 1976. The responses enable us to observe changing attitudes over time and detect new areas of concern. 500 soldiers are sent a questionnaire each month and reports on the analysis of the data are sent to the sponsor each half year. The response rate has always been excellent - generally over 80% and the soldiers appear to welcome the opportunity of making their opinions known.

3. The questionnaire covers a wide range of topics, including biographical details, questions on training, work in present location and best and worst liked features of Army life. We have recently redesigned the questionnaire and expanded the range of topics covered. From the beginning of this year respondents have been given the opportunity to write additional "open-ended" comments to qualify or expand their replies, and we find that generally around 40 percent include additional information. These comments are frequently related to specific situations experienced by the respondent, but occasionally serve to highlight areas of concern not directly covered by the questionnaire. As an example, recently several soldiers have expressed dissatisfaction with medical facilities, particularly for families. It may be the case that a directly related question on this topic should be included. Similarly, there is an increasing number of criticisms of clothing and equipment, and yet there is no specific topic in the questionnaire. Obviously the "open-ended" section is going to prove very useful.

4. One fact which has emerged from the comments is the desire for some sort of feedback. Time and again there are requests to know what the survey results are and how they will be used, but at present no suitable arrangements exist for dissemination of the results on a general basis.

5. Another main survey conducted by APRE is aimed at determining the reasons underlying officers' decisions to leave the Army prematurely.

OFFICER PREMATURE VOLUNTARY RELEASE (PVR)

Introduction

6. In the late 1970s there was a rapid increase in the number of officers at the rank of Captain and Major seeking to leave the Army prematurely. As a result, Premature Voluntary Release (or PVR) became a significant factor in officer undermanning. The Director of Manning (Army) requested that APRE assess the attitudes to Army life of those officers who applied for PVR.

7. Besides providing management with information on conditions of service which, if improved, are likely to maximise retention, attitude surveys can also serve a secondary function. By correlating expressed intentions gained from questionnaires with later retention behaviour, the prediction of future retention rates can be made.

8. It was planned originally that an initial study should be carried out to establish a baseline measure of officers' PVR intentions, and to determine which aspects of Army life were important to those who plan to stay and those who wish to leave, as well as to identify areas of dissatisfaction. Accordingly, an attitude questionnaire was designed and sent to over 700 officers at the rank of Captain and Major, between the ages of 25-36 years. 563 were completed and returned, a response rate of 73%.

9. Following a preliminary analysis of the results in 1980, PVR rates began to fall to an acceptable level and further work on this study was given a low priority. However, towards the end of 1983 PVR rates began to rise again and the 1979 questionnaire responses were re-analysed to explore how intentions at one time compared with later actions. The results suggested that PVR was predictable to some extent, but that the predictive value of the original questionnaire could have been enhanced by more direct enquiries into officers' intentions. When the questionnaire was being compiled, the sponsors were concerned that, by asking a specific specific question as to whether the officer DID intend to leave the Army, the idea actually to do so may be introduced. In consequence, instead of asking directly whether an individual was going to PVR, the question was phrased "Do you consider the Army to be you main career?

10. The analysis revealed very few differences between leavers and stayers in their responses to most questionnaire items. Although job satisfaction was the best feature of Army life for most officers, it was also the feature which, if improved, might induce those who intend to leave to remain. The other main areas in which improvements were sought included:

a. Provision of a house purchase scheme;

b. Turbulence, (since the burden of moving house often falls upon the wife in military life);

c. Responsibility, (in that many officers felt there was inadequate delegation from higher ranks);

d. Overstretch, (as this was seen to affect adversely both job satisfaction and promotion prospects).

11. However, once again PVR rates stabilised and further research was shelved until 1984, when PVR rates began to give cause for increasing concern.

THE CURRENT PVR STUDY

12. As a result of the growing numbers of officers seeking to leave the Army prematurely, APRE was asked to design and run a regular survey of PVR leavers' attitudes to Army life, as opposed to a single "one-off" survey.

13. In Spring 1985 a series of interviews was held, both with officers whose intent was to leave the Army, and with those who intend to remain. Several officers' wives were also invited to contribute to the discussions. Interestingly, it seemed that both stayers and leavers held a similar views on the drawbacks of a military career. Using the information gained during these interviews, a pilot questionnaire was designed which contained sections on a range of issues, such as pay and conditions, family life, job satisfaction and so on.

14. The pilot questionnaire was trialled during the Autumn of 1985. This was to ensure that the questionnaire covered topics relevant to decisions to PVR, and that it was expressed in suitable terminology. The results of the pilot study will be briefly discussed later.

15. The PVR survey proper began in January this year and every officer whose PVR application is accepted will be included in the survey. The PVR questionnaire, together with an explanatory covering letter, is being sent to each officer for im to complete and return to APRE anonymously.

16. In addition, we are running a survey of "stayers" this year. This means that we will send a similar questionnaire to a sample of officers who, at present, do not intend to curtail their service.

STAYERS' SURVEY

17. The stayers survey is included for comparative purposes. That is, in looking at officers' attitudes to the Army it is important to be able to identify those attitudes to Army life which are associated with the decision to leave. For example, it may be that PVR applicants will express dissatisfaction with Army pay. This will only be meaningful in understanding PVR if stayers as a group are relatively less dissatisfied, or if sources of dissatisfaction are different for stayers and leavers. Discussions are at present underway to make this a regular survey, and it is expected that such a survey will begin shortly.

PILOT STUDY

18. The results to be discussed here are all drawn from the pilot survey which was undertaken at the end of last year. We are finding that responses to the questionnaire proper are very similar.

RESULTS

19. A rather large proportion of the respondents were aged 50 and over, but inspection of the responses showed that age was an important factor in determining attitudes to the Army and reasons for leaving.

20. For the youngest group of officers (aged 20 - 29 years), over half cited two reasons as being "very important" in their decisions to PVR. These were: "Appropriate point at which to begin a new career" and "Army lifestyle no longer attractive". Other important factors cited by over 40% of this group were "domestic disruption and effects on family life in general" and " lack of job satisfaction".

21. Although 67% considered this to be a good time to start a new career, only 15% said that an offer of an alternative occupation was a very important reason for leaving. Almost 50% of this group said that the effects of Army life on their wife's long term career prospects was "very or quite important" in their decision to PVR.

22. In the 30 - 39 age group, five reasons were each cited by over half the respondents as being "very important" in their decisions to leave prematurely:

a. Appropriate point reached at which to begin alternative career;

b. Domestic disruption and effect on family life generally;

c. Army lifestyle no longer attractive;

d. Lack of job satisfaction;

e. Offer of alternative occupation.

23. To explain further, every one in this group said that domestic disruption was a very or quite important factor in their decision to leave, and over 70% considered that the effects of Army life on children's education and wife's long term career prospects were very or quite important. Less than 30% of the 40 - 49 age group and less than 16% of the 50+ group held similar views.

24. It was mentioned earlier that job satisfaction was an important factor in decisions to PVR, particularly within the two younger groups. In general, the majority of all respondents held positive views about the variety, challenge, responsibility and foreign travel associated with Army life, but other aspects of job satisfaction were looked on rather less favourably. Among these, "resources for realistic training and exercises" were said to be unsatisfactory by over 60% of the youngest age group, and by over 50% of the 30 - 39 group. Two thirds of the two younger age group, and by over 50% of the 30 - 39 group. Two thirds of the two younger groups stated that job satisfaction was "very important" in their decisions to leave, while this factor was "very important" to just a quarter of the 40 - 50 group and one third of the 50+ group.

25. It should be repeated that these results are based on too small a sample for us to draw firm conclusions, but they are illustrative of the kind of information we will be able to supply regularly and routinely.

26. Since retention of trained personnel is a major issue at present, both the Continuous Attitude Survey and the Officer PVR survey has generated some considerable interest. For both groups of respondents, particularly the younger men, turbulence and family issues are important, as is the prospect of decreasing job-satisfaction. What can be done about these issues remains to be seen.

27. On both questionnaires the "free comments" sections are providing valuable additional information. I have not formally "analysed" these yet, but early indications suggest that, apart from specific personal problems, the difficulties engendered by moving from post to post every two years or so leads to domestic stress, and this, coupled with what respondents see as shrinking prospects for future job satisfaction, gives rise to often reluctant decisions to PVR.

28. Not all officers cite job satisfaction or the need to be begin a new career, however. One stated (under the section on Domestic Circumstances and Family Life) that "it is difficult to keep my dogs"!

FINAL COMMENTS

29. The value of regular attitude surveys applies in many contexts and not just to Army personnel. Our experience has shown that there are several advantages in running longer term evaluations rather than single studies.

a. Timescale. There is a minimal time lag between the detection of a problem (for example, a rising exit rate) and the availability of objective information to deal with it. This is in contrast to a single "one-off" survey where the planning, sampling, questionnaire design and analysis have to be carried out from scratch each time.

b. Administrative convenience and cost. Our experience of running the Continuous Attitude Survey among soldiers and the Exit Quetionnaire for officers suggests that having once established the machinery (ie. sampling methods, distribution, analysis etc) the financial and personnel costs of such procedures are very low. For those actually completing the questionnaires the investment of time would be about 20 minutes, (unless, as some do, the respondent chooses to add about six sheets of notes to his replies).

c. Mangement Information. The availability of up-to-date, quantifiable and scientifically collected information can make an extremely valuable contribution to several personnel management processes. A regular survey is not as dependent on detecting differences between stayers and leavers for the prediction of wastage as is a single survey. Across-the-board trends in attitudes can be related to PVR or wastage trends in such a way that possible causal factors can be identified.

Copyright (C) Controller HMSO. London. 1986.

DEVELOPMENT AND EVALUATION OF AN INTERACTIVE COMPUTER-BASED SIMULATED PERFORMANCE TEST

Jeffrey A. Cantor
Director of Training

C. Lee Walker
Manager, Testing and Evaluation

DDL OMNI ENGINEERING

The process of testing and evaluation by computer-based interactive video has become an area rich for research and development. This area has been alive at DDL OMNI for some time through work performed for the U.S. Navy. The desirability for computer-based testing and evaluation relates to a need to promote meaningful feedback between training and its' activity assessment. Recent developments in microcomputer and video technology have combined to provide a communications tool which allows for a blend of auditory and visual stimuli. Thus it is now possible to both receive and transmit the essential elements of effective communication quickly and efficiently.

An electronic environment involving multiple stimuli allows for the functions of training and assessment to form a continuum rather than to exist as separate and often conflicting functions. Williams and Gayeski (1985) state that just as interactive instruction makes possible information presentation based on individual needs, interactive assessment obtains information from people in a personalized and efficient manner.

It is important to note that most interactive training designs provide for embedded evaluation either by inherent branching or overt test item responses. The concepts and ideas expressed in this paper recognize this fact, and suggest extended applications - beond passive embedded evaluation (branching) to overt uses of interactive video testing and evaluation.

A review of the literature in the application of this technology suggests that the medium provides certain distinct advantages over pencil and paper tests and performance tests; advantages which capitalize on, rather than replacing, these more traditional approaches to evaluation. Within the area of test administration Gayeski & Hutchinson (1983) state these advantages:

- Overall evaluation time is shorter as a result of the use of item formats, scenarios, etc. Additionally, the testing process can be terminated once a pre-determined decision point is reached.
- Motivation of the examinee can be controlled, as branching will permit the examinee to quickly reach challenging items and test areas. Motivation is an important aspect of interactive video evaluation which stems from two sources: (a) feedback on the selection of the individual's decisions/choices; and (b) a perceived increase in fidelity to actual situations. The latter is due to results in both improved performance and increased credibility of the test results to the examinee. Timing of test items and responses are controllable, so that skills involving fixed time reactions or rapid decision and response can be effectively assessed. Measures of speed and accuracy can be recorded for each response.
- Fidelity increases over paper and pencil testing because of the combined visual and aural stimuli;
- Expense is decreased compared to supporting performance testing because of reduced test monitor and equipment requirements;

- Scheduling of the testing session can be at the convenience of all concerned. Limited proctoring is necessary as security is incorporated into the machine. There is also flexibility in scheduling because of (1) the ability of the computer to manage the testing process and (2) a reduction in conflicts over use of resources (i.e. rooms, labs, equipments).

We further suggest that there are psychometric advantages in the use of interactive video for testing and measurement.

- Objectivity in measurement is enhanced over that available through performance observation. Actions are offered in response to stimuli; steps are assessed in performing a procedure; answers to discrete questions are recorded non-judgmentally and consistently.
- Branching can be permitted to the extent that it contributes to assessment of a person's abilities, but at the same time constrained such that non constructive divergences can be eliminated.
- Adaptation of the test to an individual's current skill level: In problem-solving situations this can lead to varying paths and sequences relative to one's current competency level. Therefore each person can be successful in solving a particular problem while the test diagnoses the strengths and weaknesses of the examinee's solution. Issues of test difficulty become less of a concern due to the ability of the computer to generate unique sets of items. The incorporation of video will allow for optimum use of this essential stimulus into the item formats and facilitate the assessment of these areas of human performance.
- Combination of cognitive and skill measurement within the same instrument: This is particularly important when dealing with actions based on rules and principles in which correct response for a single case is not sufficient to infer understanding of the rule or principle.
- Interactive video information presentation results in better acceptance of the testing process. This is due to a perceived appropriateness of the test items (face validity).
- Varied response modes: New methods such as "unalerted" responses and selective responses add to the ways in which information can be gathered.

A Typical Test Description and Application

In early 1978, Data-Design Laboratories (DDL) began research and development of interactive video testing technology. The original application of this technology was in the Fleet Ballistic Missile (FBM) Weapons System Training Program (Braun & Tindall; 1974, Braun, Tindall and Robinson; 1975, Robinson & Walker; 1977). A proprietary pencil-and-paper process termed the Decision Measurement System (DMS), which provides interaction between the test items and the examinee, was used as the point of departure for the interactive video test.

The DMS is a unique testing device that was used to assess the ability of technicians to correctly diagnose operational or maintenance problems. It was used as a "skill" measurement device in the Personnel and Training Evaluation

Program (PTEP), where skill is defined as problem-solving ability, rather than manipulative or perceptual-motor ability. The DMS used an answer-until-correct response mode to ensure that each examinee knows the answers to prior test items before proceeding to answer the question at hand. A "single thread" solution was thus used in the DMS so that the examinee was returned to the main solution path at each point (item) in the test sequence.

To illustrate some issues and options in interactive video testing we will draw upon testing material prepared for the Poseidon Missile program. For this test the visual stimuli consisted of panel and gauge indications associated with a missile launch. Auditory stimuli consisted of the stream of reports (both directed and background) occurring in the launch process, as well as related mechanical noises. Questions in each case required the respondent to mentally summarize what he had seen and heard and then to project an action response. A diagram will be used to depict each question type.

Figure 1 depicts the most basic questioning method. A visual segment is played and then a written multiple-choice question is posed. In this case, as in all others within this exercise wrong answers are remediated. Questions of this type simply replace the accustomed written "set up" description with a sound/motion description. Replay can be permitted or denied at the author's initiative.

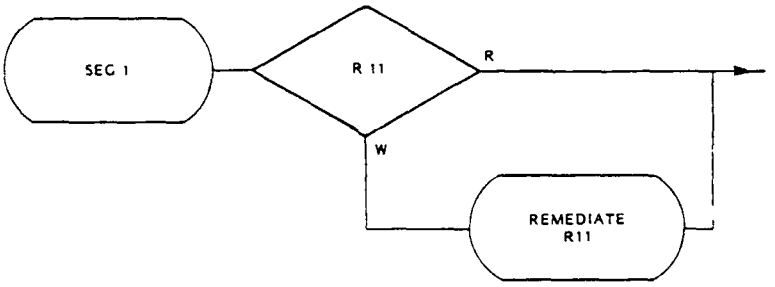


Figure 1

Figure 2 uses a "gating" question to establish some level of problem recognition and follows with multiple choice questions on specific actions. The gating question is an "unalerted response". No formal question is posed but the student must recognize that an action to missile systems is required because of changes to the navigation system. If the respondent indicates that action is required before it actually is, the system records the fact. If the respondent does not recognize a need for action, remediation is given without posing the question on specific actions. Effectively this program segment records data on three objectives:

- Recognize the indicators and reports associated with a normal countdown

- o Understand the relationships between missile gyro initialization process and the status of the Ships Inertial Navigation System
- o Know what action to take to maintain the proper missile gyro initialization process

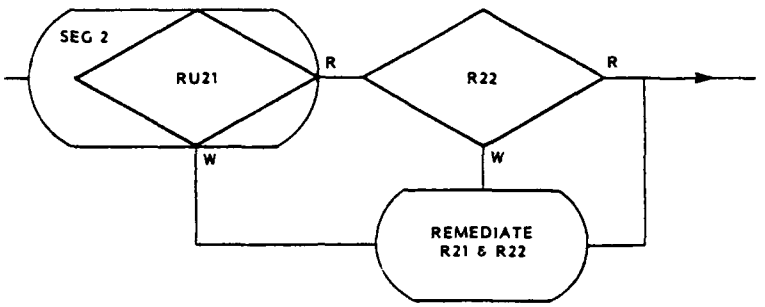


Figure 2

Figure 3 presents a situation where two pieces of information are gathered by a single response. The respondent is required to both recognize that action is required and then to take the action properly. The "set up" which is very rapidly presented visually is one that would require a complex written explanation that would be almost impossible to provide without telegraphing the answer. Objectives satisfied are:

- o Recognize the indicators and reports associated with a normal countdown
- o Know the circuit relationship between three indicator lights and a key operated switch
- o Know the firing procedures for a specific faulted condition

Each of these objectives can be treated as a separate scorable unit.

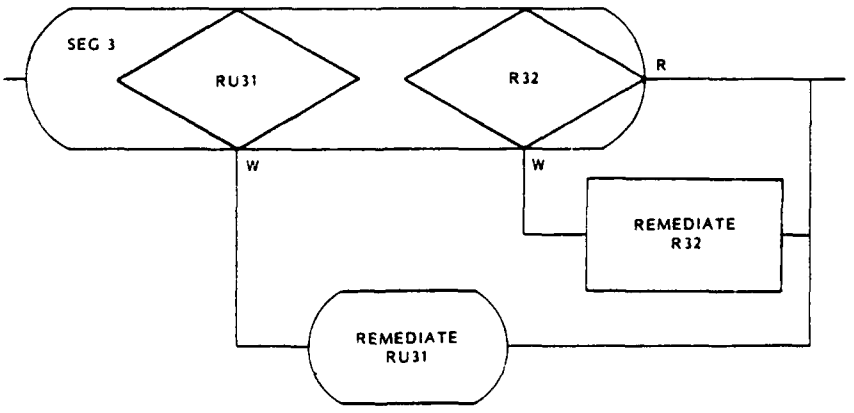


Figure 3

Figure 4 depicts the first step in a fault isolation sequence. A previous question segment required recognition of incorrect timing in the countdown stemming from a problem in the "Missile Compartment". This segment begins by presenting a menu with four sets of indicators permitting the respondent to choose any indicator or in sequence each indicator. These indicators show normal configuration. No score is recorded until the respondent chooses to move past the indicators. A "wrong" mark is given if the one faulted indicator has not been reviewed at this point. This replicates a situation where there are several possible causes to a problem and a narrowing process is part of the solution path. The opportunity to look randomly about the whole Fire Control and Launcher system was not provided for two reasons. First, it was intended to constrain the question to recognition of indicators on specific equipments, and secondly "look anywhere" is very expensive to support with pictures and programs. In this case it would have been an added degree of sophistication and experience which would have told us very little more about the respondent's ability to deal with the problem posed. The questions in this segment related to the following objectives:

- o recognize correct indicators on specific equipment for this stage of the countdown
- o know the circuit configuration for the indicators which initially flagged the problem in sufficient detail to tell which reviewed indicators could have contributed to the problem.

It should be noted that the four indicators used here are not all located in the same shipboard spaces or school laboratories so that replication of this test point in a performance test would be both cumbersome and time consuming.

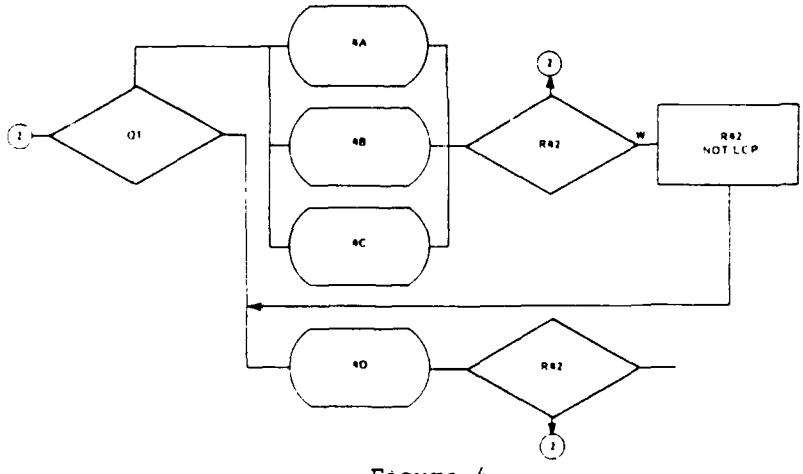


Figure 4

With each of these, response time can become a scoring factor in addition to the actual responses.

Scoring can of course be simplified by limiting response modes or question types. In doing this, however, the test designer fails to effectively use the testing power of the interactive video medium.

Scoring

Interactive Video Testing can combine aspects of both performance testing and written testing in the way measurement points are presented and answers recorded. In addition, because the medium can pose questions in ways unavailable to either written or "hands on" testing and record answers in ways that are difficult to achieve in more common testing methods, new measurement forms are achieved. When different forms of measurement are incorporated in a single test, complex scoring situations are developed. The of these complex situations are still more the property of intuition and empiricism than in any accepted theory of measurement. In the several items discussed above there are:

- o Conventional multiple-choice questions
- o Questions with situational recognition followed by multiple choice items
- o Questions with single responses answering two questions
- o Sort-and-select questions
- o Unstructured situations

Summary

Interactive video should be viewed as a measurement domain with unique properties and capabilities and not just a convenient way of delivering conventional tests. The combination of measurement modes available enable great flexibility in but at the same time present new problems in scoring and testing, equating tests. We must be cautious as we move more fully into interactive video testing that we don't permit the exuberance of "Look what we can measure." to carry us away from the sober reality of "What does it mean?"

References

1. Braun, F.B. and Tindall, J.E., A new sequential multiple choice testing device. Paper presentation at the Military Testing Association Conference, 1974.
2. Braun, F.B., Tindall, J.E. and Robinson, M.A., The Decision Measurement System - Current Status. Paper presentation at the Military Testing Association, 1975.
3. Gayeski, B. and Hutchinson, L., Interactive Assessment Instructional Innovator (28) 2, Feb, 1983.
4. Robinson, M.A. and Walker, C.L., The Decision Measurement System as a means of testing performance by simulation. Paper presentation at Computer Assisted Test Construction Conference 1977.

Gunnery Indices as Measures of Gunner Proficiency

Bob G. Witmer

U.S. Army Research Institute - Fort Knox Field Unit

Armor crewmen perform a variety of tasks, many of which are important in achieving their mission of attriting enemy forces. Tank gunnery is the most visible and perhaps the most critical of these tasks. Few armor leaders would dispute the importance of skilled gunnery performance in land warfare. Increasing gunnery proficiency remains a high priority for armor leaders and trainers, but improvements in training and selection procedures have been slow to occur. In part, the lack of improvement has been due to the difficulty in accurately measuring live-fire gunnery performance. Gunnery measures have typically suffered from low reliabilities due to errors induced by the fire control system, round-to-round dispersion, and inaccurate scoring techniques.

These errors have been largely eliminated in gunnery simulators such as the Unit Conduct of Fire Trainer (UCOFT) that accurately record multiple measures of gunnery speed and accuracy. The UCOFT for example, records identification times, fire times, hit times, hit or miss, and aiming errors for each round fired, and several derived measures of gunnery performance. The capability to obtain accurate multiple measures of gunnery performance eliminates the reliability problems usually associated with measuring live-fire tank gunnery performance, but it introduces some additional problems. Evaluators wishing to determine the effects of training programs or other variables on gunnery proficiency now have a number of possible measures of effectiveness (MOE's) from which to choose. These evaluators can either select a measure of effectiveness from among several possible MOE's or combine measures into a composite criterion. Another alternative is to enter several MOE's into a multivariate analysis of variance (MANOVA). The MANOVA solves for a best combination of the measures of gunnery effectiveness for determining the effects of the independent variables.

Selecting a single measure of effectiveness has certain advantages over the multivariate approach. With a single measure of effectiveness, univariate analysis of variance can be employed to detect the effects of the independent variables. Univariate techniques are simpler than multivariate techniques and the results of univariate analyses are easier to interpret. The manner in which the MOE's are combined by multivariate procedures is a function of the relation between the MOE's and the independent variables under investigation, and hence may vary from one investigation to the next. The multivariate combined criterion variable will always change from one application to the next because the weights of the MOE's are empirically determined. Consequently the researcher cannot know what the MOE will be until the analysis is complete. In contrast, a single measure of effectiveness (e.g., hit time) is always the same variable regardless of the independent variables used.

Composite Measures of Gunnery Performance

A single measure of gunnery performance, however, may not adequately describe the behaviors required for gunnery proficiency. To better describe gunnery proficiency, some researchers have developed composite measures of gunnery effectiveness. Composite measures combine the advantages of including several dimensions of gunnery proficiency with the inherent advantages of simple univariate analysis procedures. Typically, the composite gunnery measures developed have consisted of relatively simple combinations of two measures. For example, Harris, Melching, Morrison and Goldberg (1982) combined accuracy in aiming the laser rangefinder with firing accuracy to obtain a composite measure of accuracy. Eaton, Johnson, and Black (1980) measured tracking proficiency by converting time and error scores to z-scores and adding the z-scores.

Each year individual tank crews must qualify by scoring at least 700 points on a series of 10 day and 10 night gunnery engagements with a 70 point minimum on 7 of the 10 engagements. Collectively the qualifying engagements are known as Table VIII. Field Manual 17-12-1 (1984) provides scoring charts for determining a crew's score for each Table VIII engagement. Engagement scores are a composite function of hit times and own tank exposure times, and vary from 0 to 100. To score 100 on a dual-target engagement, the first and second targets must each be hit within a prescribed time interval. If the first target is hit and the second target missed, the score is a function of the time to hit the first target and the time that the firing tank was exposed before returning to a defilade position. The composite scores of the day and of the night engagements are summed to yield separate day and night Table VIII scores.

Composite measures may be used to evaluate gunnery performance on gunnery simulators. The UCOFT computes three composite measures, Target Acquisition (TA), Reticle Aim (RA), and System Management (SM). Target Acquisition is based on the time to acquire targets and number of target identification and classification errors. Reticle Aim is a function of time to fire the first round, time to kill, and magnitude of the aiming error. System Management is determined by the number of switch setting errors prior to and at the time of firing. TA, RA, and SM are evaluated on a four point grade scale (i.e., A, B, C, F) with a score of A=4 corresponding to the fastest times and fewest errors and F=1 corresponding to the slowest times and most errors. For example, to receive an "A" on TA, acquisition time must be less than or equal to five seconds with no identification or classification errors. An "F" is received if acquisition time is greater than 15 seconds or if the number of identification errors, classification errors, or both exceeds two (UCOFT Instructor's Handbook, 1982).

Bonder has proposed a composite measure of gunnery effectiveness that is inversely related to the enemy attrition rate (Taylor, 1980). Bonder's measure, denoted E(Txy), is the expected value of the time for an individual Y firer to kill an X target. E(Txy) is a function of time to acquire a target, time to fire the first round after the target is acquired, time to fire a round following a hit and following a miss, flight time of the round, probability of a hit on a round following a hit and on a round following a miss,

and probability of destroying a target given that it is a hit. Although E(Txy) has been used primarily for combat modeling purposes, it might also be useful as a composite measure of gunnery proficiency. Some simulators (e.g., VIGS, UCOFT), however, do not provide the requisite measures for calculating E(Txy).

Gunnery Indices

Composite measures such as those described by Taylor (1980), Eaton et al. (1980), Harris et.al. (1982) and Table VIII measures were developed for different purposes and hence use different variables combined in different ways. The common characteristic of these measures is that they combine two or more separate measures of gunnery proficiency into a composite performance measure. Composite measures that are based on a series of observations of gunnery performance may provide a useful metric for evaluating gunnery proficiency. Such measures serve to index individual gunnery proficiency and henceforth will be referred to as gunnery indices. Total scores on Table VIII and TA, RA, and SM averages for a UCOFT exercise are examples of gunnery indices. Gunnery indices may be based on accuracy measures, on measures of gunnery speed, or on a weighted combination of speed and accuracy.

A new comprehensive measure of gunnery proficiency, the Gunnery Index (GI) is proposed as an alternative to traditional gunnery measures and existing gunnery indices. The GI, shown below, is computed from performance measures that are routinely provided by the Videodisk Gunnery Simulator (VIGS) and the UCOFT.

$$\text{Gunnery Index (GI)} = (.5W1\sqrt{ABC} + .5W1D + W2E) \times 100$$

Where A = Number of First Round Hits/Number of Targets Presented

B = Number of Hits/Number of Rounds Fired

$$C = \frac{\text{(Number of Targets Presented - Number of Targets Not Engaged)}}{\text{Number of Targets Presented}}$$

D = J/Average Aiming Error

E = K/Average Hit Time

And J is the smallest average aiming error from the center of mass exhibited by a given population of gunners for a particular set of engagements; $0 < J \leq$ smallest average aiming error for the sample.

K is the fastest average hit time exhibited by a given population of gunners for a particular set of engagements; $0 < K \leq$ fastest average hit time for the sample.

W1 and W2 are weights assigned by the evaluator based on the judged relative importance of accuracy and speed. W1 + W2 = 1.0.

The Gunnery Index includes several measures that are thought to be important to tank gunnery. First round hits, total hits, number of rounds expended, number of targets not engaged, aiming error, and hit time are all accounted for. Aiming error is a measure of the distance in mils from the reticle aim point to the center of mass of the target. Hit time is the time elapsing between the target appearance and the time that the round strikes the target. Times to acquire targets (acquisition times) and times to open fire on targets (opening times) should effect gunner proficiency only to the extent that they influence hit time and therefore were not included in calculating the Gunnery Index.

Gunnery Index Psychometric Properties

The Gunnery Index (GI) has some interesting properties. It allows the researcher or evaluator to determine the relative weighting of time and accuracy in measuring gunnery proficiency. The weights may be based on the judged relative importance of speed and accuracy components by gunnery experts, or weights can be assigned on the basis of the variance to mean ratios of each component with greater weights assigned to the more variable component. The range of scores for GI varies between a number near zero and one hundred. When there are no first round hits, GI is determined by aiming error and hit time. When no hits are obtained, hit time is assumed to be infinite and GI is solely a function of aiming error. When only one round is allowed per target, A = BC, and GI = (.5W1A + .5W1D + W2E) x 100.

Composite measures of gunnery proficiency like GI encompass a larger proportion of the gunnery behavioral domain than do single measures of gunnery proficiency. Therefore, we might reasonably expect such composite measures to better represent true gunnery proficiency. Confidence in GI as a valid measure of proficiency increases to the extent that it varies predictably with certain variables. GI should increase as soldiers receive gunnery training and practice their gunnery skills. Similarly variables that are expected to decrease proficiency such as longer target range, reduced visibility, or firing under degraded conditions should cause GI to decrease. Confidence in GI further increases if moderate correlations between GI and other gunnery measures are obtained and if the reliability of GI is high.

A Test Case for the Gunnery Index

To begin evaluating the reliability and validity of GI as a measure of gunnery proficiency, data collected in a previous experiment (Witmer, 1986) were reanalyzed using the Gunnery Index (GI) as the sole measure of proficiency. Witmer (1986) investigated transfer of training between two tank gunnery simulators for 24 novice gunners. The results of that study as determined by a Multivariate Analysis of Variance (MANOVA) showed that performances on the VIGS and UCOFT were significantly correlated, but that skills learned on one device did not transfer to the other device. Test-retest reliability was reasonably high for the accuracy and speed measures except for aiming error, and significant increases in performance were obtained as a function of practice on each device.

Reanalysis of the Witmer (1986) data using ANOVA and GI as the single criterion measure with speed and accuracy components weighted equally showed significant increases in GI scores as a function of practice on VIGS and UCOFT ($F_{1,22} = 80.1$, $p < .0005$). GI scores on the VIGS and UCOFT were significantly correlated (Pearson $r = 0.42$, $p < .05$) but no transfer was obtained from one device to the other. Test-retest reliabilities using GI as the criterion measure and correcting for test length were .70 and .64 for the VIGS and COFT performances, respectively. These results suggest that GI is sensitive to the effects of simulator training and that GI scores are reliable when based on an adequate sample of performance on these gunnery simulators. Respectable test-retest reliabilities were obtained despite the fact that the GI scores for the novice gunners were significantly higher for the retest than for the test and that one component of GI, aiming error, has been shown not to be very reliable (Witmer, 1986).

Using the Witmer (1986) data, correlations between GI and the individual measures comprising it were computed. The correlations of the VIGS GI composite with its constituent components are as follows: proportion of first round hits ($r = .60$, $p < .001$), total number of hits ($r = .68$, $p < .0005$), aiming error ($r = -.56$, $p < .002$), and hit time ($r = -.79$, $p < .0005$). Similarly the GI composite of UCOFT performance correlates significantly with the following variables: proportion of first round hits ($r = .91$, $p < .0005$), total number of hits ($r = .90$, $p < .0005$), aiming error ($r = -.78$, $p < .0005$) and hit time ($r = -.82$, $p < .0005$). These results suggest that strength of the relationship between Gunnery Index scores and scores on each of the measures comprising it is substantial. This finding increases our confidence that GI scores are useful composite measures of gunnery performance.

Uses and Limitations of the Gunnery Index

The Gunnery Index proposed in this paper provides a meaningful composite measure of gunnery proficiency that relates highly to traditional measures of gunnery accuracy and speed. GI is very sensitive to changes in gunnery behavior and permits fine discriminations among different gunnery performances. For example, two gunners with the same number of first rounds hits, the same number of total hits and the same average hit time may still differ on GI as a function of aiming errors, number of rounds fired, and number of targets not engaged. The sensitivity of GI to small changes in gunnery proficiency parameters makes it ideal for determining the effects of training programs or experimental treatments on proficiency. Gunnery indices such as GI may be useful for measuring learning on training devices and for quantifying overall proficiency on these devices. Because GI measures the performance relative to a best obtained performance, it may be useful in selecting or assigning gunners or tank commanders.

Like other composite measures, GI may not be appropriate in every measurement situation. Generally composite measures are preferred for decision making and evaluation applications, but not for research applications where the primary goal is understanding the relationships between various predictors and separate criterion dimensions (Casio, 1978). If the goal of re-

search is to understand the manner in which different variables affect particular aspects of gunnery proficiency in order to derive better training prescriptions, then GI should not be used. When the researcher needs to know, for example, that visual discrimination training improves hit time, but has no effect on aiming error, or that tracking practice improves aiming error but has no effect on hit time, multivariate procedures should be used, rather than gunnery indices. If the goal is to diagnose specific deficiencies (e.g., poor tracking, poor speed) in an individual's gunnery performance, then detailed observations of the various gunnery proficiency measures are required and GI may not be applicable. On the other hand, if the goal is to evaluate the overall gunnery proficiency of individual gunners or to assess training program effects, environmental effects (e.g., fog, darkness), or the effects of using degraded gunnery techniques on gunnery proficiency, then gunnery indices such as GI may be very useful.

REFERENCES

Cascio, W. (1978). Applied psychology in personnel management. Reston, Virginia: Reston Publishing Company, Incorporated.

Eaton, N. K., Johnson, J. & Black, B. A. (1980). Job samples as tank gunnery performance predictors (ARI Technical Report 473). Fort Knox, KY: Fort Knox Field Unit, US Army Research Institute for the Behavioral and Social Sciences.

General Electric Company, Defense Systems Division, Simulation and Control Systems Department (1982). Instructor's utilization handbook for the M1 Unit-Conduct of Fire Trainer (U-COFT), Vol I. Daytona Beach, Florida: Author.

Harris, J. H., Melching, W. H., Morrison, J. E. & Goldberg, S. L. (1982). Development and evaluation of a stabilized gunnery training program. (HumRRO Report No. FR-MTRD-(KY)-82-1). Alexandria, Virginia: Human Resources Research Organization.

Taylor, J. G. (1980). Lanchester-type models of warfare, Vol II (USARO Technical Report 16403-IM). Research Triangle Park, N.C.: US Army Research Office (DTIC No. AD-A090843).

US Army Armor School (1984). FM 17-12-1, Tank Combat Tables M1. Fort Knox, Kentucky: Author.

Witmer, B. G. (1986). Device-based training and transfer between the Videodisk Gunnery Simulator (VIGS) and the Conduct of Fire Trainer (COFT). ARI Draft Technical Report submitted for publication.

Stinger Team Performance in the Realistic Air Defense Engagement System (RADES)

David M. Johnson and John M. Lockhart

US Army Research Institute for the Behavioral and Social Sciences, Fort Bliss

Current US Army Forward Area Air Defense (FAAD) weapons include the man-portable Redeye and Stinger infrared-seeking missile systems; the vehicle-mounted Chaparral infrared-seeking missile system; and the towed or vehicle-mounted Vulcan gun system. The FAAD environment will be characterized by multiple, hostile and friendly, jet and helicopter aircraft. Some of these aircraft will be attacking local targets, while others will be transiting to attack targets in the rear (Little and Vane, 1986). Some aircraft will be attacking the air defenders themselves, since this is SOP among military pilots. Meanwhile, air defenders will be expected to do their job amidst smoke, noise, ground-fire, and the other accompaniments of modern warfare.

An air defender's job is to destroy hostile aircraft, or force them to abort their mission, while refraining from engaging friendly aircraft. Many subtasks must be performed expertly in order to achieve this goal. Once alerted, air defenders search for aircraft. Aircraft may be detected by any crewmember, regardless of his duty position. The aircraft's location is then communicated to the unit leader and gunner. Once a target is visually detected, gunners will begin tracking the target, ranging it (often visually), and interrogating it with the Identification Friend/Foe (IFF) subsystem, if available. Meanwhile, the leader will be attempting to identify the target visually, with binoculars. If identified as hostile the leader will issue an engagement command. The gunner will then engage the target, assuming he has acquired it with his weapon and it is within range.

Purpose

The RADES testbed was developed to provide a cost-effective means of performing controlled research on issues affecting FAAD system performance. Since so many of the current FAAD engagement steps require the human visual system, factors which affect the visual system would be expected to affect FAAD performance. The purpose of the present research was to investigate the performance of Stinger teams as they engaged fixed-wing (FW) and rotary-wing (RW) aircraft under conditions chosen to vary the visual information available.

Specifically, hostile and friendly FW aircraft (i.e., jets) were presented singly flying either a pop-up or a lay-down attack maneuver against either a terrain or a sky background. Hostile and friendly RW aircraft (i.e., helicopters) were presented singly in either a frontal aspect (face-view) or a quartering aspect (side-view) against either a terrain or a sky background. Finally, two hostile RW aircraft were presented simultaneously with both in either a frontal or a quartering aspect against either a terrain or a sky background.

Data were collected on a range of Stinger team engagement actions (i.e., detection, IFF interrogation, identification, engagement command, IR lock-on, superelevate, fire). For the limited purposes of this paper, however, we need only report data on detection, identification, and fire to adequately characterize the engagement process.

Method

Participants

Twelve Stinger teams stationed at Fort Bliss, Texas, participated. Each team consisted of two soldiers, an E4/5 team leader and E3/4 gunner.

The RADES Simulation

RADES is located at Condron Field, White Sands Missile Range, New Mexico. This desert area contains mountains 10 km to the west and 60 km to the south. Visibility is usually in excess of 60 km. Skies are usually clear.

Space limitations preclude a detailed description of the RADES simulation; which can be found elsewhere (Drewfs, Frederickson, Johnson, and Barber, 1987). RADES employs FAAD crews and teams, manning instrumented FAAD weapon systems, in simulated engagement of subscale aircraft. RADES aircraft are of two types— flying 1:7 scale "jets" and nonflying 1:5 scale "helicopters". The jets represent friendly (USA) and hostile (USSR) attack aircraft. The helicopters, which pop-up from hidden positions and hover, represent friendly (USA) and hostile (USSR) attack/utility aircraft. The friendly aircraft presented in the current experiment were the F-16 (FW) and AH-1 (RW). The hostile aircraft presented were the MiG-27 (FW), Mi-8 (RW), and Mi-24 (RW).

FW aircraft were presented singly, flying either pop-up or lay-down maneuvers, incoming from an azimuth which either had a terrain or a sky background. FW offset was approximately 1.5 km (fullscale) from the Stinger team. An automatic position/location system determined the location and range of the FW aircraft during trials. RW aircraft popped-up, under computer control, for durations of 40 seconds. RW aircraft were presented either singly or doubly, in either face-view or side-view, against either a terrain or a sky background. RW targets were positioned behind sand dunes, when lowered, at a fullscale distance of approximately 3 km from the Stinger team. All targets were equipped with an infrared radiation (IR) source that the Stinger weapon could acquire and lock-on to. All targets were painted in "sand and spinach" desert camouflage colors.

All teams engaged targets with the same Stinger Tracking Head Trainer (Training Set, Guided Missile, M134). The IFF Simulator was modified to produce an "unknown" return upon interrogation. Interface electronics automatically recorded critical Stinger engagement events (i.e., IFF interrogation, IR lock-on, superelevate, fire). Verbal engagement events (e.g., detect, identify, engagement command) were keyed-in during trials by a data collector who was wired into the team's commo net. All weapon and verbal engagement events were automatically tagged as to time and range of aircraft at occurrence. In addition, sensors attached to the Stinger automatically kept track of where the weapon was pointing in azimuth and elevation.

Procedure

One team was brought to the RADES site per day. On the hour-long trip out teams were briefed about RADES and questions were answered, where applicable. Stinger field manuals were provided for review of key engagement actions, command and control, etc. Once at RADES site, a team was given an operations order stating mission, enemy, sector of responsibility (90 degrees), sector boundaries, and Primary Target Line. A trial began with an air defense alert condition "red". During a trial the team engaged targets "as if it were the real thing". Each trial lasted from one to three minutes. A trial ended with an air defense alert condition "white". Between trials, the Stinger was lowered and the team sat in a bunker with their backs to the RADES range.

The Weapons Control Status for all trials was "tight" (i.e., positive, visual, identification of hostile intent required before fire). The IFF return was always "unknown". All trials received an early warning of approximately 30 seconds. Teams were not cued as to aircraft type, identification, or azimuth. All teams used 7x50 binoculars for identification but not for detection.

The FW portion of the experiment contained 8 engagement trials per team [intent (hostile/friendly) x maneuver (pop-up/lay-down) x background (terrain/sky) = 8]. The single RW portion of the experiment also contained 8 trials per team [intent (h/f) x aspect (front/side) x background (t/s) = 8]. The double RW portion contained 4 trials per team [aspect (f/s) x background (t/s) = 4]. All teams received a different, counterbalanced, presentation order of these same 20 engagement trials in this repeated-measures design.

Results

Data were analyzed separately for the three sub-experiments (FW, RW-Single, RW-Double) and the three dependent variables (detection range/time, identification range/time, fire range/time). The dependent measure of performance for the FW sub-experiment was fullscale aircraft range from Stinger team at event occurrence. The dependent measure of performance for the RW sub-experiments was time of event occurrence from target visual availability. Times accumulate across engagement events. Generally, long ranges and short times mean good performance.

Fixed-Wing Results

All of the FW aircraft presented were detected. Detection range was analyzed by a three-factor, repeated-measures Analysis of Variance (intent x maneuver x background). There was a significant main effect of target background upon detection performance ($F=135.35$, $df=1/11$, $p<.001$). The same aircraft when presented against a sky background were detected at a greater mean range (8.1 km) than when they were presented against a terrain background (3.5 km). All other main effects and interactions were not statistically significant.

All of the FW aircraft presented were identified. Identification range was also analyzed using a three-factor, repeated-mesures ANOVA. There was a significant main effect of target background upon identification performance ($F=8.06$, $df=1/11$, $p<.025$). Again, the same aircraft when presented against a sky background were identified at a greater mean range (3.4 km) than when they were presented against a terrain background (2.3 km). No other main effects or interactions were significant.

Eighty-five percent of the hostile FW aircraft presented were fired upon. Fire range for the hostile targets was analyzed using a two-factor, repeated-measures ANOVA (maneuver x background). The main effects and interaction were not statistically significant. However, closer analysis of these data for fire ranges showed this conclusion to be grossly misleading. The same targets when presented against a sky background were engaged at 3.4 km incoming, while they were engaged at 3.1 km outgoing when presented against a terrain background. (Crossover for these flights was approximately 1.5 km.) This is a substantial difference in favor of the sky background conditions.

Single Rotary-Wing Results

Ninety-four percent of all RW aircraft presented were detected. The six

trials on which a target was not detected were all trials in which a helicopter (either friendly or hostile) was presented in frontal aspect against a terrain background. Detection times were analyzed using a three-factor, repeated-measures ANOVA (intent x aspect x background). No main effect of background was found. There was, however, a significant background by intent interaction ($F=22.15$, $df=1/11$, $p<.001$); wherein the detection time was longer for friendly aircraft with terrain background than for friendly aircraft with sky background, and vice versa for hostile aircraft. There was a significant main effect of aspect ($F=23.74$, $df=1/11$, $p<.001$). Aircraft presented in side-view were detected in 7.2 seconds, while those presented in frontal aspect were detected in 10.0 seconds. There was also a significant main effect of intent ($F=76.44$, $df=1/11$, $p<.001$). Hostile aircraft were detected in 6.3 seconds, while friendly ones required 10.9 seconds. Finally, the interaction effect of aspect by intent approached statistical significance ($F=3.56$, $df=1/11$, $p<.10$). This interaction resulted from the fact that the aspect effect (described above) was larger for friendly aircraft than it was for hostiles. None of the remaining interaction effects were significant.

Ninety-three percent of all RW aircraft presented were identified. The seven trials on which a target was not identified were all trials in which a helicopter (friendly or hostile) was presented in frontal aspect against a terrain background. Identification times were analyzed using a three-factor, repeated-measures ANOVA (intent x aspect x background). Again, there was not a main effect of target background; but background did interact significantly with aircraft intent ($F=13.46$, $df=1/11$, $p<.01$). Friendly aircraft were identified more rapidly when presented against a sky background than when presented against a terrain background, and vice versa for hostiles. There was a significant main effect of aspect ($F=27.35$, $df=1/11$, $p<.001$). Aircraft presented in quartering aspect were identified in 13.6 seconds, while those presented frontally required 19.2 seconds. There was a significant main effect of intent ($F=43.56$, $df=1/11$, $p<.001$). Hostile aircraft were identified earlier (13.1 sec.) than friendly aircraft (19.7 sec.). Finally, there was a statistically significant interaction of aspect by intent ($F=12.22$, $df=1/11$, $p<.01$). This interaction resulted from the fact that the size of the aspect effect for friendly aircraft (14.9 sec. < 23.6 sec.) was substantially larger than that for hostile aircraft (12.2 sec. < 14.0 sec.). None of the remaining interaction effects were significant.

Overall percent correct identifications (number correct/total possible) for RW aircraft were 77. Percent correct identifications were greater for targets presented against sky (92) than against terrain (63); and also greater for targets presented in side-view (83) than those presented face-on (71).

Eighty-seven percent of all hostile RW aircraft presented were fired upon. The mean fire time for all hostile engagements was 18.0 seconds. Fire times were analyzed using a two-factor, repeated-measures ANOVA (aspect x background). No significant differences were found.

Double Rotary-Wing Results

Engagement event times presented for the double, hostile, RW trials were based on the first of the two simultaneously presented targets to be engaged.

Detection times were analyzed using a two-factor, repeated-measures ANOVA (aspect x background). There was a significant main effect of target background ($F=12.96$, $df=1/11$, $p<.01$). Targets presented against a sky background were detected in a mean time of 4.0 seconds, while those presented against terrain required 5.7 seconds. No other main effect or interaction was significant. The mean detection time over all conditions of the double RW

sub-experiment was 4.9 seconds. By comparison, the mean detection time over all conditions of the single RW sub-experiment, for hostile aircraft only, was 6.3 seconds. An analysis of these detection times using a t test for correlated samples showed this difference to be significant ($t=2.60$, $df=11$, $p<.05$).

Ninety-seven percent of all presented RW targets were identified. Of the three targets which were not identified, two were presented in frontal aspect against a terrain background and one was presented in side aspect against a terrain background. Identification times were analyzed using a two-factor, repeated-measures ANOVA (aspect x background). None of the identification time differences were statistically significant. The mean overall time for identification was 12.4 seconds. By comparison, the mean overall identification time for the hostile targets in the single RW sub-experiment was 13.1 seconds. This difference was not statistically significant. The mean overall percent correct identifications (number correct/total possible) was 87.

Fire times were analyzed using a two-factor, repeated-measures ANOVA (aspect x background). None of the fire time differences were statistically significant. The mean overall time for fire was 17.3 seconds. By comparison, the mean overall fire time for the hostile targets in the single RW sub-experiment was 18.0 seconds. This difference was not significant.

Discussion

Subscale aircraft were presented to Stinger teams under conditions designed to vary the visual information available. This was done because so many of the current FAAD engagement procedures require specifically visual activities on the part of the air defenders. As has repeatedly been shown in the Results Section, varying the visual information available to air defenders, in ways which are likely to occur on the battlefield, changes their performance substantially. We will discuss the relevant variables in turn. Unfortunately, space restrictions force us to limit this discussion.

Maneuver: No differences in performance were found for FW aircraft flown in two quite different attack maneuvers. It turns out that these maneuvers did not differ during that portion of the flight paths where detect and identify responses were recorded. Kirkland (1972) also found no difference in performance associated with similar maneuvers, when elevation was held approximately constant.

Background: This was a powerful source of variance in the current experiment. Since all RADES aircraft were painted desert camouflage colors, the relative perceptual contrast between aircraft and background was less when they were presented against desert terrain than when presented against sky. As a result, engagement performance was seriously impaired when either FW or RW aircraft were presented against the terrain background. Similar contrast effects have been shown for FW aircraft models by Baldwin (1973a) and for fullscale RW aircraft by CDEC (1978).

Aspect: This variable had a sizeable effect upon the engagement of single RW aircraft, especially friendlies. Any aircraft when presented head-on provides a smaller visual target than when it is presented side-on. As a result, FAAD engagement performance which depends upon visual detection, visual identification, and visual tracking of the target is sure to be impaired. This effect was magnified in the case of the friendly AH-1 which is noted to be particularly narrow in frontal cross-section. This effect of presentation aspect has been shown for FW models by Baldwin (1973a) and by CDEC (1980) for fullscale RW aircraft.

Aircraft Size: As alluded to above, the smaller the aircraft, all other things being equal, the more poorly air defenders will perform when attempting to detect, identify, and track it. The scale models of friendly RW aircraft presented in this experiment were smaller than the hostile RW aircraft because the actual, fullscale aircraft are smaller. Hence, engagement performance is poorer for the friendlies. Size also accounts for the interactions of background and aspect with intent. Baldwin (1973a) has reported similar size effects for the identification of FW models. CDEC (1978) also reported significantly poorer detection performance for the smaller of their two fullscale RW targets.

Multiple, Simultaneous Targets: The time taken to detect the first of two hostile RW targets was less than the time taken to engage a single hostile RW target. In the former case there were two engageable targets within the same 90 degree search sector; in the latter case only one. Hence, this difference was probably caused by an effectively reduced search sector for the "double" condition. In one of the earliest reported experiments on the detection and identification of aerial targets by ground observers, Wokoun (1960) found just such an effect of reducing search sector size.

References

Baldwin, R. D. (1973). Relationship between recognition range and the size, aspect angle, and color of aircraft. HumRRO Technical Report 73-2. Alexandria: George Washington University.

CDEC (1978). Helicopter acquisition test. CDEC Test FC094. Fort Ord: US Army Combat Developments Experimentation Command.

CDEC (1980). Helicopter relative detectability test. Final Letter Report, B046091L, March, 1980. Fort Ord: US Army Combat Developments Experimentation Command.

Drewfs, P. R., Frederickson, W., Johnson, D. M., & Barber, A. V. (1987). Validation of the Realistic Air Defense Engagement System (RADES). ARI Technical Report. Alexandria: US Army Research Institute for the Behavioral and Social Sciences (in press).

Kirkland, G. C. (1972). Evaluation of the deployment of a lightweight air defense weapon system (LADS). Project No. 44-69-05. Quantico: Marine Corps Development and Education Command. (Declassified 1978)

Little, J. H. & Vane, M. A. (1986). Forward Area Air Defense. Air Defense Artillery, Spring, 1986 (pp. 12-16). Fort Bliss: US Army Air Defense Artillery School.

Wokoun, W. (1960). Detection of random low-altitude jet aircraft by ground observers. HEL Technical Memorandum 7-60. Aberdeen Proving Ground: US Army Human Engineering Laboratory.

EVALUATION OF COMPUTER BASED TRAINING

Lieutenant S. Latimer, B.A., B.Ed. (Hons), Royal Australian Navy, Staff of the Royal Naval School of Educational and Training Technology (RNSETT).

INTRODUCTION

1. The Royal Navy is currently undertaking a large scale development of Computer Based Training (CBT). The cost of this innovation requires that systematic evaluation studies be carried out. The author is involved in the planning and conduct of these studies. The purpose of this paper is to highlight some of the difficulties confronting evaluators of CBT and to make some practical suggestions for this work. To provide some basis for discussion an evaluation of one of the Royal Navy's first CBT projects, in which the author participated, will be outlined.

2. Before proceeding it is important to clarify some terms. The Royal Navy defines CBT as the 'use of computers in any part of a training system' (RNSETT, 1984). CBT is seen to consist of Computer Managed Instruction (CMI) and Computer Assisted Instruction (CAI). CMI relates to the use of computers to assist in the setting and marking of examinations, statistical analysis, production of reports, and the identification of training failures. CAI is the direct use of the computer to assist in the learning process. This paper is only concerned with the latter use of computers in training. 'Evaluation' is taken here in its narrowest sense to be the gathering of information about an instructional package in order to judge its worth.

A STUDY

Evaluated Training

3. The study evaluated some CAI materials delivered on a network CBT system manufactured by a well established computer firm. These materials are implemented within career courses for electrical technicians conducted at the Royal Navy's major training establishment. The CAI consists, first, of 12 lessons (each about 75 minutes in duration) on radar theory provided as part of a 4 week module entitled Radio Frequency Techniques. These lessons were produced by a commercial courseware authoring house. The remaining CAI comprises 15 hours on mathematics. This is given to trainees who have failed electrical theory modules because of deficiencies in basic computation skills. All the mathematics computer lessons were authored by service personnel.

Evaluation Strategy

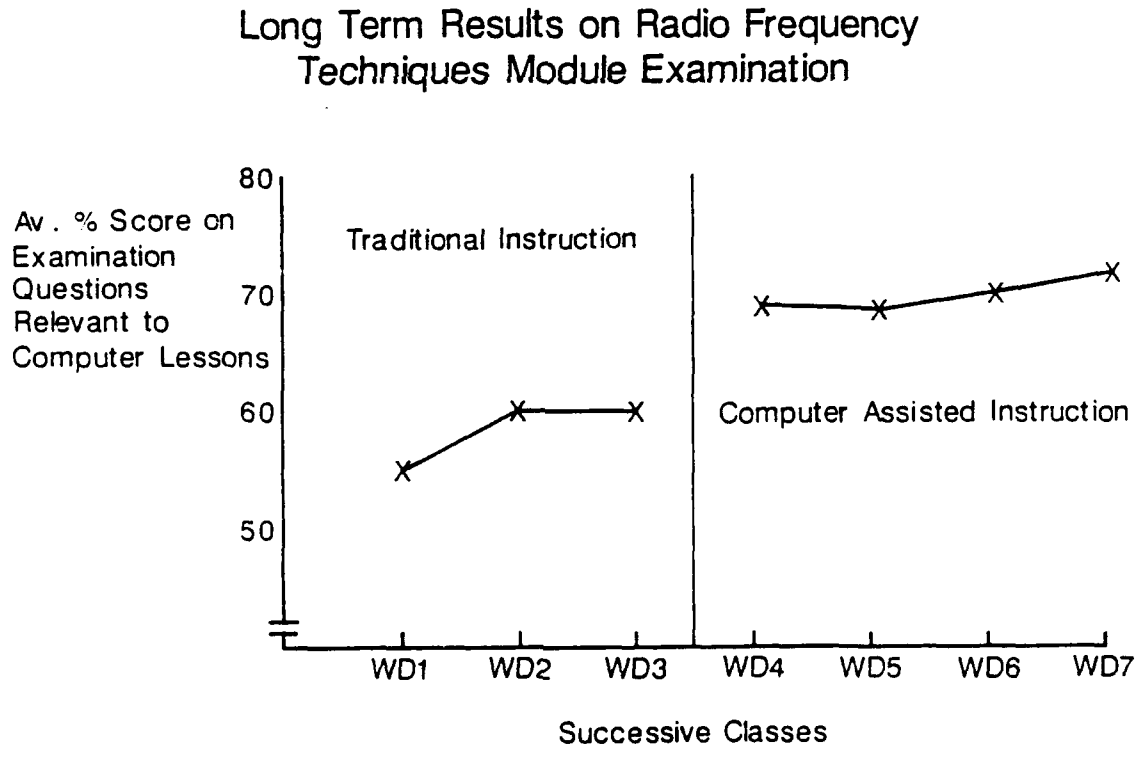
4. The major concern of the study was to identify any changes in the quality of training resulting from the inclusion of computer delivered lessons in the instructional programmes, and where possible to determine those features of CAI most responsible for any such gains or losses. The methodology of the evaluation involved, first, longitudinal and cross-sectional designs with examination/test scores as the dependent variable and instructional treatment as the independent variable. The second part

of the methodology involved the analysis of more qualitative data such as the content, structure and sequence of the computer lessons, the use of the aid by instructors, the attitudes of trainees and instructors towards CAI, and the school management of the computer facility.

Some Findings

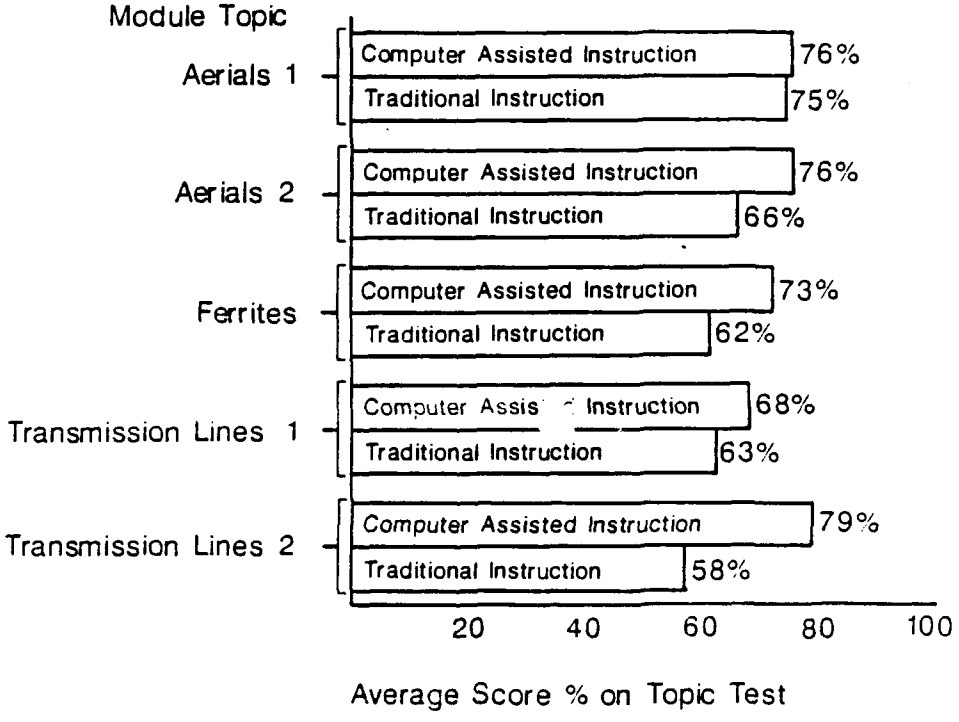
5. The Radio Frequency Techniques module is routinely assessed by a written examination. About half of the examination questions relate to topics which now involve CAI. It was informative to examine the scores of successive classes prior to and after the implementation of the computer lessons. As shown on Figure 1, there was a marked increase in achievement on the module examinations which coincided with the introduction of CAI. Although not indicated on Figure 1, the scores on the examination questions not relevant to the content of the computer lessons remained fairly constant over the period of the study. These results were interpreted as indicating that the introduction of CAI increased achievement because had other factors been responsible (viz. higher ability trainees, increased instructor effectiveness) an improvement would have been observed across all the examination questions.

Figure 1



6. The second analysis carried out for the radar theory computer lessons involved two classes completing the Radio Frequency Techniques module in parallel. For each topic one class was given CAI while the other class was given traditional instruction. The teaching method was alternated between the two classes to control for differences in ability. Short answer tests covering each topic were administered. The results of this comparison study are presented in Figure 2. This indicates that for each topic CAI produced higher average scores than traditional instruction. Although the magnitude of the difference varied between topics, these results were generally interpreted as showing the CAI to be superior to traditional instruction, at least in terms of trainee achievement.

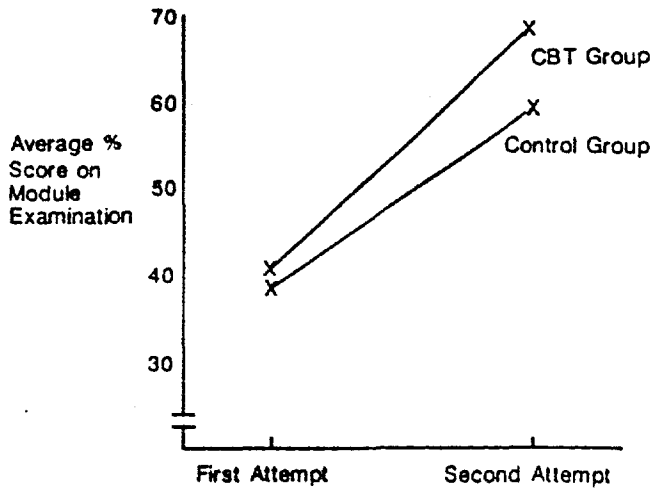
Figure 2 Short Term Comparison of Teaching Method for Radio Frequency Techniques Module



7. The experimental evaluation of the mathematics computer lessons involved trainees who had failed on an electrical theory module. These trainees were divided into two groups, one received the mathematics computer lessons prior to re-sitting the module examination while the other received no remedial instruction of any type. The average scores for the two groups on both administrations of the examination are given in Figure 3. The groups performed at about the same level on the first attempt while the trainees completing the computer lessons obtained a much higher result than the control trainees on the second attempt. This was interpreted as indicating a benefit from the mathematics computer lessons in providing remedial instruction.

Figure 3

Effect of Remedial Mathematics Computer Lessons



8. As mentioned earlier, the study also involved several analyses of more subjective data. One of these examined the structure of the radar theory computer lessons to see how well they conformed to the appropriate instructional model. Because the materials were designed as tutorials they were compared to the 'events of instruction' model proposed by Gagné and Briggs (1979). This analysis was conducted by coding each computer lesson frame as the instructional event to which is related most. Some of the lessons were shown to contain most of the components of a tutorial in the correct proportion. Other lessons, however, provided insufficient recall of pre-requisites, contained too much straight forward presentation of information, asked too few questions, and provided inadequate feedback. All the computer lessons failed to assess mastery of the lesson objectives.

9. Another qualitative analysis involved the use of the radar theory computer lessons by the instructors. From interviews with the instructors and from some observation of their classes it became clear that the computer lessons were used as either revision or primary instruction. The

latter application was the most interesting. Within particular topics the computer lessons replaced lecture presentations and did so in less time (2/3 to 1/2). However, any time savings were lost because the instructors conducted follow-up discussions to the computer lessons. It appeared, therefore, that while this particular use of CAI did not result in reduced training time (at the group level) it did lead to more interactive learning.

10. The attitudes of the trainees and instructors towards CAI were also measured in the study. Generally, the trainees found the computer lessons to be motivating, easy to use and valuable. They were divided on their preference for computer delivered lessons over traditional instruction. The instructors responded positively on such factors as the suitability of the medium, the difficulty of the content, the appropriateness of the learning process, and the likely outcomes in terms of increased trainee motivation and achievement. Interestingly, the instructors did not expect CAI to save time overall or to decrease their workload.

11. As a final point on the results of the study some attempt was made to relate the findings in the experimental evaluations to the more qualitative data. For instance it was found that the radar theory computer lessons shown to have little advantage over traditional instruction in the comparison study contained too little or too much of some instructional events.

SOME COMMENTS AND SUGGESTIONS

12. Having described the study it is now appropriate to consider the adequacy of its methodology and of its outcomes. Beginning with the methodology, the experimental evaluations of the computer lessons reported above are all of questionable validity. The long term study of the examination results did not control for extraneous variables. The short term comparison study had no pre-test. The pre-test - post-test analysis did not include instructor delivered remedial instruction as one of its treatments.

13. Many of these problems are unavoidable in a setting such as that of the study reported here. First, naval trainers are unwilling to set up controlled experiments if this means denying some trainees the use of the computer lessons. Second, naval courses are often so tightly programmed that there is no time for pre-testing. Third, the small through-put on specialized naval courses and the inflexibility of course programming precludes the setting up of comparison groups.

14. However, the military evaluator can compensate, in part, for weaknesses in experimental evaluation design by also collecting more subjective data in less rigid observational frameworks. This can involve much more than attitudes, as in the study reported here. Hopefully, there will be consistencies between both types of analyses and so add strength to the conclusions drawn from the evaluation.

15. Some comment can also be made about the study in terms of its outcomes. The substantive conclusion in the study was that the introduction of CAI had resulted in some improvements in trainee achievement on the relevant instructional programmes. While this evaluation outcome may satisfy the needs of the training developer it is far short of the type of information required by other participants in naval training. The users of course graduates, in this case the Fleet, are most interested in how well CAI can equip the man for his operational performance. Training managers, at the school and organizational level, are concerned with how CAI impacts on the usage of limited training resources. Clearly, the military evaluator must make the scope of the evaluation wide enough to gather data on the full range of possible effects and not just consider information about the relative instructional effectiveness of CAI.

CONCLUSIONS

16. The high cost of CAI packages requires that systematic evaluations be carried out. The practical problems confronting the military evaluator are such that there is a danger of the experimental evaluations being superficial. To counter this the evaluation studies should also include more qualitative analyses which do not suffer from the same types of validity problems. Further, the military evaluator should specify the widest possible set of evaluation objectives which direct the study to consider the concerns of the users and the training managers, as well as those of the training developers. Such measures are necessary if the full value of CAI is to be demonstrated. It would be unfortunate if the military were to abandon this innovation on the basis of inadequate evaluations.

REFERENCES

Gagné, R.M., and Briggs, L.J. Principles of Instructional Design (Second Edition). New York. Holt, Rinehart, and Wintson, 1979.

RNSETT, Computer Based Training A Guide for Trainers. Portsmouth, UK: Royal Navy, 1984.

Disclaimer

The views expressed in this paper are those of the author and do not necessarily reflect the official policy of the Royal Naval School of Educational and Training Technology or the Royal Navy.

NPRDC October 1986

THE COURSE EVALUATION SYSTEM

John A. Ellis
Frederick G. Knirk
Barbara Taylor
Barbara McDonald

Navy Personnel Research and Development Center
San Diego, California 92152

The Course Evaluation System

Introduction

The Course Evaluation System (CES) is a tool for evaluating the effectiveness and appropriateness of a course or a segment of instruction (for example, a module or lesson) within a course. The procedures were designed to be used by Curriculum and Instructional Standards Officers (CISOs), course managers, senior instructors, educational specialists and any other personnel involved in training. The system focuses on the processes of instruction in the traditional Navy classroom. The Navy currently teaches over 7000 courses. The great majority of these are presented in traditional classrooms and laboratories with a group of students being taught by a single instructor. Recent directives by the Chief of Naval Education and Training and the Chief of Naval Operations indicate that this training format will continue to play a predominant role in Navy education and training. Given this situation, the quality and effectiveness of Navy classroom instruction is an important concern. Recent evaluations of lecture-type instruction in civilian schools have shown that instructional quality is highly variable, non-standard and often poor. Similar problems have been observed in Navy schools (Ellis, 1985, VanKekerix, Wulfeck & Montague, 1982). In addition, research on learning has shown that variables affecting student achievement can be effectively controlled. Yet, there have been no systematic attempts to explore the applicability of controlling these or other potentially useful variables in Navy classroom training.

To develop the CES, military and civilian training-related research studies were reviewed for those variables which most significantly affect learning. Current Navy classroom evaluation procedures were also reviewed. The CES represents a synthesis of the above information and is focused on variables which most affect the quality of Navy training.

The CES gives course managers and instructors the capability to pinpoint problems in ineffective courses and make revisions. Specifically, this system is designed to evaluate course quality by assessing the adequacy and consistency of the three primary components of instruction; objectives, test items, and the instructional presentation. There are two evaluation components of the CES: Objective and Test Adequacy and Consistency Evaluation and Presentation Evaluation. Additionally, there is a preliminary step that involves classification of the objectives to be evaluated and separating broad objectives into smaller units. The CES user may choose to apply either one or both CES components. In addition, an evaluation may involve an entire course or be limited to specific lesson(s) or module(s) that have problems.

For objective and test evaluation, course/lesson/module objectives and test items are examined to determine their adequacy and how well they relate to the tasks the student must perform on the job. The objectives are then matched to related test items to check for consistency. This is accomplished by determining if the conditions, action(s), and standard(s) specified in the objectives match those contained in the test items.

The presentation evaluation involves reviewing the instructor and student guides and observing the instructor's presentation in the classroom. Both the consistency of the presentation with the objectives and the adequacy of the presentation are evaluated.

To apply the Course Evaluation System you need the following resource documents: the course objectives (usually found in the curriculum outline), the course tests, Instructor Guide(s) and Student Guide(s).

The remainder of this paper discusses who should use the CES and describes how to classify objectives. For detailed information about to apply the CES using the appropriate forms write to:

John Ellis
Code 51
Navy Personnel Research and Development Center
San Diego, CA 92152
or call at AC 619 225 6434 or AV 933 6434.

Who Should Use the Course Evaluation System

The Course Evaluation System was designed to be used by personnel responsible for course management, course revision, course evaluation, instructor evaluation, and maintenance of instructional quality. In the Navy these people include senior enlisted personnel who are course instructors and managers, Curriculum and Instructional Standards Officers, civilian educational specialists and training specialists, and non-government contractor personnel. This list includes people who are subject matter experts (SMEs) (e.g. instructor personnel) and people who are not SMEs (e.g. educational specialist). Because of this it is important to point out that the CES can only be used if you are an SME yourself or if you have at least one (preferably two or three) SMEs available to assist you in your evaluation. In addition, if you are an SME you should have knowledge of the basic principles of instruction, that is, you should have at least completed basic instructor training.

Classifying Objectives

In order to use the CES, the evaluator must know how to classify objectives and how to separate broad objectives into smaller more meaningful, measurable and relevant objectives. Background information on these topics are available in NPRDC SR 79-24, The Instructional Quality Inventory, II. User's Manual and in NPRDC SR 83-2, Handbook for Testing in Navy Schools. However, enough guidance is provided in this section to enable accurate classification of objectives.

There are five different types of objectives. The types are based on the types of tasks performed on the job. The five types are:

REMEMBER (R)

USE-UNAIDED TRANSFER (UUT)

USE-UNAIDED NO TRANSFER (UUN)

USE-AIDED TRANSFER (UAT)

USE-AIDED NO TRANSFER (UAN)

Classifying objectives into one of the five types is a three step process.

Step 1: Determine if the student is to Remember or Use the information. A student can either REMEMBER information or USE the information to do something. This distinction corresponds to the difference between knowledge and application. The following two objectives illustrate the REMEMBER-USE distinction.

> REMEMBER: The student will draw the symbol for resistor.
>
> USE: The student will set up a Simpson 260-5p multimeter for measuring resistance.

These two objectives differ with respect to what the student is supposed to do. In the first item, the student has to REMEMBER something. In the second, he has to apply or USE his knowledge. The REMEMBER-USE distinction is a simple one. The determination can usually be made by looking at the action in the objective or test item. Typical action verbs are listed below. The ones on the left usually indicate REMEMBER-level tasks, while the ones on the right usually indicate USE-level tasks.

REMEMBER	USE	
name	apply	operate
state (from memory)	remove	repair
list (from memory)	analyze	adjust
recall	derive	calibrate
remember	demonstrate	replace
write (from memory)	evaluate	assemble
recognize	solve	disassemble
explain (from memory)	prove	calculate
select	sort	troubleshoot
describe	maintain	load
identify	compute	predict
	determine	unload

Figure 1. Action Verbs

The distinction between the verbs "Remember" and "Use" may be a bit tricky when you have an objective like the following:

> Using the manual NA 01-S3AAA, list the location and function of the fuel dump valve.

At first glance it might be tempting to classify this a Remember objective. But on closer examination, what the student is actually required to do is to locate the information in the technical manual and record that information on a piece of paper. What the student must demonstrate is the ability to

use the technical manual and therefore, this is a USE Level objective.

Step 2: If the task is USE determine if it is USE-AIDED or USE-UNAIDED. If it is USE-UNAIDED the student has no aids except his own memory. If it is USE-AIDED the student has a job aid to perform the task. This can be determined by looking at the "conditions" part of the objective. Anything that replaces the need for memory counts as an aid.

AIDS include:

1. A list of procedure steps from a technical manual or MRC card.
2. A formula for solving problems and a description of how to use the formula.
3. A statement of a rule or a set of guidelines for troubleshooting and repairing a piece of equipment.

Normal tools, materials, etc., are *NOT* aids.

Step 3: If the task is USE determine whether it is TRANSFER or NO TRANSFER.

A NO TRANSFER objective requires the student to perform a specific procedural task the same way every time. It consists of an ordered sequence of steps designed to accomplish a specific task, which needs to be demonstrated in only one way. There is no requirement that the student transfer or generalize performance to new situations. A good rule of thumb is that if you can answer YES to "if you've seen one you've seen them all" or "if you've done it right once or twice you can do it right under any circumstance" then you have a NO TRANSFER objective.

The following are samples of NO TRANSFER objectives.

1) Tie a bowline knot within 15 seconds. (Use-Unaided)

2) Destroy classified documents under routine conditions using the outline in OPNAVINST 5510.1. (Use-Aided)

3) Weigh a CO2 fire extinguisher and record its weight in accordance with the furnished MRC. (Use-Aided)

Key words or phrases that you might see in NO TRANSFER Level objectives are listed below. The student will:

apply	remove
operate	replace
repair	assemble
adjust	produce
calibrate	destroy

Remember that a NO TRANSFER objective can be Use-Unaided *or* Use-Aided. The aid is a list of steps to be performed.

A TRANSFER Level objective can also be a sequence of steps. However, TRANSFER Level objectives can be applied in a variety of situations or on a variety of different equipments or objects. In other, words they involve tasks and jobs that are more complicated than no transfer objectives. Instruction for TRANSFER objectives requires many different examples and practice items so that as many situations as possible are covered. For TRANSFER objectives it is not possible to teach every possible situation or example in class. For example, if you wanted to teach 5th graders three digit multiplication, it would not be practical to give them every three digit problem there is. Instead, students are taught a rule for dealing with the different types of problems. You would then give them problems that represent the different types of possible problems so that they could practice using the rule. Finally, you would test them on new problems to see if they had learned how to apply the rule. To decide if you have a TRANSFER objective you need to determine if students will be required to deal with problems or situations on the job that may not be covered in class.

The following objectives are at the TRANSFER Level.

1) Given video tape recordings of radar scopes displaying jamming, the student will classify the type of jamming used for each display. (Use-Unaided)

2) Given the formula for capacitive reactance, instruction about how to apply it, and the values of a frequency and capacitance from a schematic, the student will calculate capacitive reactance. (Use-Aided)

3) The student will solve for total power in a DC parallel circuit. (Use-Unaided)

Notice that the TRANSFER Level objective may be Use-Aided *or* Use-Unaided. The aid is at least a statement of the formula or rule to be applied and should include guidelines for when and how to apply it. Key phrases that you might see in a TRANSFER level objective are given below.

The student will:

solve	find
derive	translate
prove	program
calculate	add
troubleshoot	subtract

In summary, there are three steps in classifying an objective.

1) Determine if the student is to REMEMBER or USE information.

2) If the student is to USE information, determine whether the task is USE-AIDED or USE-UNAIDED.

3) If the student is to USE information, determine whether the task is TRANSFER or NO TRANSFER.

References

Ellis, J.A. (April 1985). Review of the Air Intercept Controller Basic Course; NPRDC TR 85-22. San Diego, CA: Navy Personnel Research and Development Center.

Ellis, J.A., & Wulfeck, W.H. (October 1982). Handbook for Testing in Navy Schools; NPRDC SR 83-2. San Diego, CA: Navy Personnel Research and Development Center.

Ellis, J.A., Wulfeck, W.H., & Fredericks, P.S. (1979). The Instructional Quality Inventory, II. User's Manual; NPRDC SR 79-24. San Diego, CA: Navy Personnel Research and Development Center.

Van Kekerix, D. L., Wulfeck, W. H., & Montague, W. E. (1982). Development, Test, and Evaluation of the Computer-assisted Study Management System in a Navy "C" School: Summary Report; NPRDC TR 82-33. San Diego, CA: Navy Personnel Research and Development Center.

VALIDATING VOCATIONAL TRAINING MATERIALS THROUGH THE INSTRUCTIONAL QUALITY INVENTORY (IQI)

George M. Usova, Ph.D.
Shipyard Instructional Design Center, Atlantic
Department of the Navy

The purpose of this paper is to describe the Instructional Quallity Inventory (IQI) -- an instructional validation process. Validation can be conducted in a variety of ways, depending upon the length and complexity of the instruction. Deciding what validation strategy to use must be based upon the nature of the material, cost effectiveness, timeliness, and resources available. This paper discusses the IQI as a validation strategy whose success has been supported by research and substantiated by experts in the field of training.

What is validation? The guidebook, Interservice Procedures for ISD (1975) defines validation as,

> a process through which a course is revised until it is effective in realizing its instructional goal. The validation process therefore is testing instructional materials on a sample of the target population to insure the materials are effective.

This operational approach in the Interservice definition emphasizes the student tryout.

On the other hand, Dick (1983) refers to the validation process as formative evaluation and describes validation in more general terms:

> Formative evaluation is the process used to obtain data for instructors to use to increase the efficiency and effectiveness of their instructional material. The emphasis in formative evaluation is on the collection of data in order to revise the instructional materials, to make the materials as effective as possible.

In the above definition, change and revision to material to increase its effectiveness is the major thrust. Dick does not specify the method through which this change must occur.

In a similar manner, Komoski (1983) defines validation as,

> evaluation conducted during the developmental or prepublication stage of an instructional materials life. Conceptually, it may include any sort of relevant information that can be used to improve a material's performance while that material is still in a dynamic enough state to be changed for the better.

The key term in Komoski's definition is "prepublication". Realizing that evaluation is an ongoing process and that the improvement of learning materials are considered possible for as long as they continue to be used, it is, nevertheless, necessary to conduct a validation at the prepublication stage to assure a "first cut" improvement effort.

Paralleling this concept, Scriven (1967) states that

> The role of formative evaluation is to discover deficiencies and successes in the intermediate versions of the new curriculum ... a debugging operation.

Finally, addressing the issue of validation from a content perspective and not solely from a learning efficiency viewpoint, Nathenson and Henderson (1980) say,

> Expert appraisal and valuing are equally as important as student data. Content quality, both input and potential output is a different issue from learnability. No amount of data from students can validate the quality of the potential outcomes or curricular excellence.

Adding further support is Geis's (1973) definition of "validated instruction" which includes "not only effectiveness, but also relevance and accuracy of content".

Summarily, validation is defined in a variety of ways. The terms validation and formative evaluation are used interchangeably; however, the common thread that runs through all definitions is the essence of change toward improvement to the materials.

How can validation be conducted? The answer to this question is: In a variety of ways. When the term validation is mentioned, the most common image conjured in the minds of most trainers is that of a small group of representative students involved in a tryout; however, as the research reveals, there are a variety of equally acceptable validation strategies that yield similar results and may in fact be more efficient, cost effective, and less cumbersome.

ONE-TO-ONE AND SMALL GROUP

One such approach to validation is the one-to-one student trial or the clinical approach. The clinical approach is called for when utilizing the "learner verification" strategy for formative evaluation. In this approach, the evaluator observes the actual completion of the training lesson by the learner. Any difficulties encountered are then noted by the evaluator. This provides instant identification that is cataloged and can be investigated once the lesson is completed. This is a far more accurate tool than trusting to the fallible memory of the student to recall problems after the examination. Further, as a result of their 1982 study, Low, Thurston, and Brown concluded:

> Research on learner verification has strongly established that a clinical or tutorial approach to data collection results in revisions to instructional materials which are at least as effective as revisions in large scale group-based methods.

Interestingly, in research conducted by Wager (1983) and Kandaswamy (1976) comparing effectiveness of validation between the one-to-one and small group approaches found that

> materials revised according to feedback from the one-to-one sessions with mixed ability levels were as effective as materials revised according to

feedback from a combination of one-to-one and small group sessions alone.

This surprising finding is further reinforced by Carey (1978):

> Employing both the one-to-one and small group stages of formative evaluation may not be necessary for the effective revision of instruction. Formation evaluation may save resources by conducting one-to-one evaluation using a mixture of ability levels.

In sum, the message is saying that one-to-one student trials with students of different ability levels yields the same validation information as small group trials.

EDITORIAL APPROACHES

Another approach to validation is the editorial approach which involves using instructional editing guidelines that incorporate the attributes of effective instruction. Such a guideline might consist of checklists or review forms specifying criteria for sound instructional development features.

Golas (1983) in a comparative study found no differences in instructional effectiveness between modules evaluated (validated) via the collection of student data (tryout) and the use of instructional editing guidelines; furthermore, the cost of developing modules in which student data was collected was substantially greater than costs using editing guidelines.

Komoski (1983) further states that

> Studies have demonstrated that using short, simple instructional sequences -- revisions based on editors' judgments have improved a material just as effectively as revision based on the learner verification and revision of the same material.

Komoski, however does not believe this "editorial approach" to be effective with larger and more complex instruction.

The point here is that there may not be a pressing need to conduct a student tryout (individual or small group) when instruction developed is short and simple. The evidence shows there is little to be gained by it.

IQI

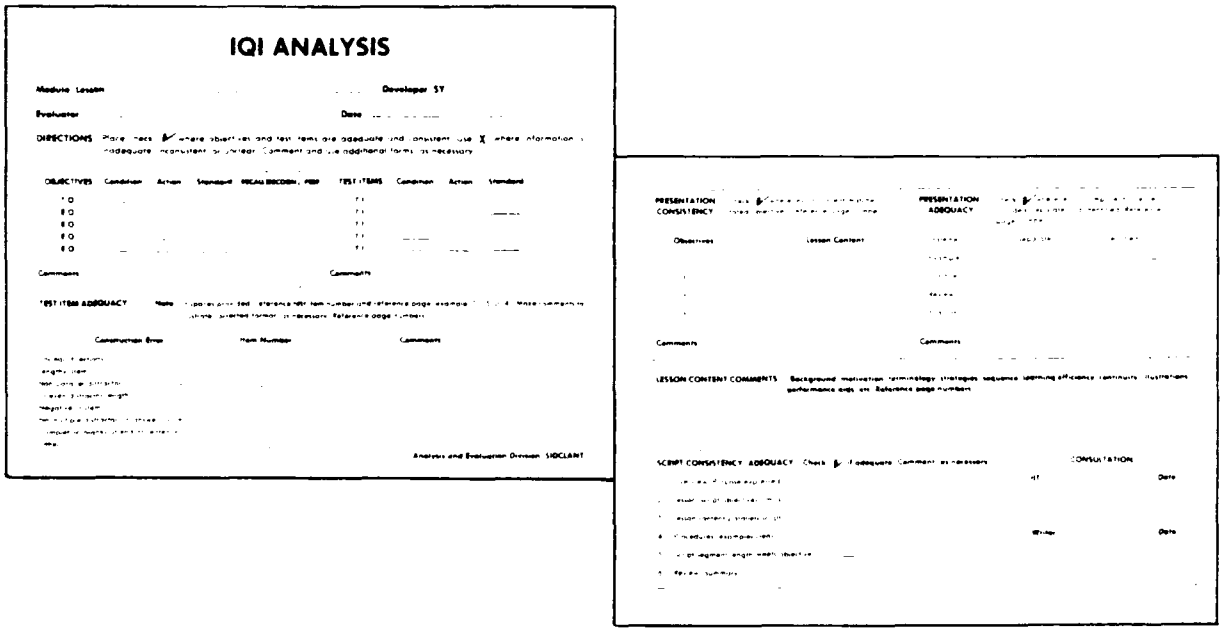
The Instructional Quality Inventory (IQI) is an evaluation device used to assess instruction. It possesses similarities, as a form of validation, to the method previously discussed -- the editorial approach. The IQI, however, is more precise, definitive, and comprehensively known for its specificity in examining objectives, test items, and lesson content for adequacy and consistency. Based upon empirical test studies, the IQI provides a useful framework for determining deficiencies in current instruction (Ellis and Wulfeck, 1983).

In fact, according to Ellis and Wulfeck, "students receiving instruction from IQI material score higher on tests than those receiving instruction from non-IQI material".

The first step requires a trained education specialist or cognitive psychologist to review the lesson's terminal and enabling objectives to assure that the objectives are clear and will stated -- that the objectives contain, as appropriate, a condition, action, and standard. In addition, the objective must address only one learner behavior, be measurable, and of course, be related to the job. (See Figure 1 for the author's adaptation of the IQI for training materials developed in the Shipyard Training Modernization Program.)

The second step is to proceed to the test item(s) which assess that objective. Two questions must be asked here: (1) Is the test item adequate? i.e., the test item must be constructed well enough to be clear, free of hints, and of the proper item format to best address the objective and (2) Is the test item consistent with the objective? i.e., the item must test the requirement established in the objective; otherwise, we will not know whether the objective has been met. Conversely, we must not have extraneous test items that are not applicable to any of the stated objectives (test items that are lesson content driven). After each test item is reviewed with its corresponding objective and appropriate comments and corrections are made, the education specialist then examines lesson content -- again for adequacy and consistency.

Lesson adequacy means that the content must contain information that is clear, separate, and identified to the learner. All lessons should contain the following: (1) statement of learning (2) examples or demonstrations (3) practice (4) review and (5) evaluation. As stated, each of these basic lesson components should be clearly identified within the body of the text and through spacing, heavy bold print, and other media techniques should be separated and identified as such within the body of the text. In sum, these provisions guide and direct the learner so that he has the best possible opportunities to meet his objectives. As can be seen, the IQI has been applied to the lesson's objectives, test items, and lesson content to insure that the lesson has met the requirements of sound principles of training.



IQI ANALYSIS

Figure 1

References

Carey, J. O., "A Fair Return on Your Formative Evaluation Dollar", Unpublished manuscript, Florida State University, 1978.

Dick, Walter, The Systematic Design of Instruction, Tallahassee, Florida, Florida State University Press, 1978.

Golas, K., "Formative Evaluation Effectiveness and Cost", Performance and Instruction Journal, June, 1983, 17.

Kandaswamy, S. et al, "Learner Verification and Revision: An Experimental Comparison of Two Methods", A.V. Communication Review, 24, 1976, 316 - 328.

Komoski, Ken, "The Empirical Improvement of Learning Materials", Performance and Instruction Journal, June, 1983, 3.

Markle, S. "Review", Performance and Instruction Journal, June, 1983, 28.

Nathenson, M. and Henderson, E., Using Student Feedback to Improve Learning Materials, London, 1980.

Phase III Interservice Procedures for ISD, U.S. Government Printing Office, August, 1975, 383.

Scriven, M., The Methodology of Evaluation, AERA Monograph Series in Curriculum Evaluation, No. 1, Chicago: Rand McNally, 1967.

Wager, Jane C., "One-to-One and Small Group Formative Evaluation", Performance and Instruction Journal, June, 1983, 7.

Wulfeck, W., et al., "Instructional Quality Inventory", Performance and Instruction Journal, June 1983, 11.

Measuring Factors Related to Reenlistment Decisions [1]

Alfred L. Smith, Jr.
U.S. Army Research Institute for the Behavioral and Social Sciences

The purpose of this paper is to describe the development and pilot-testing of several scales from the Reenlistment Incentives and Career Decision Making Questionnaire (RICQ). In its revised form, this survey is being administered to a cross-section of enlisted personnel in CONUS and OCONUS installations. The intended outcome of the research is development of a comprehensive, interactive model to explain the reenlistment decision process in order to modify retention efforts. Factors related to reenlistment in the military have been examined at length, but in a number of different combinations with a variety of different focuses in mind, rather than comprehensively. From models that have been developed to explain turnover, at least five major factors have been identified repeatedly as very important : Needs and their fulfilment within the Army, satisfaction with the Army, organizational commitment, occupational stress, and perceived job alternatives. Five scales were developed for the RICQ to measure these factors.

After extensive review of the literature and published instruments which measure factors related to turnover, the first step in developing the RICQ was to conduct semi-structured interviews with soldiers. There were two purposes for this: to identify variables (especially important needs, values) that seem to be most critical in making a decision to reenlist or exit from the Army and to seek confirmation of some preliminary hypotheses about how those decisions are finally reached. Thirty-six first, second or third term soldiers with less than six months of service commitment left were questioned about personal, professional, and family goals, their desires and aspirations, characteristics of their jobs, their satisfactions/dissatisfactions with the Army, perceived opportunities in the civilian job market, job stress, and influence of family on their reeenlistment decisions.

The interview results were used to assist in creating the pilot RICQ. In addition to the five major scales corresponding to the factors noted above, the instrument included demographic and miscellaneous items covering reenlistment bonus programs/options, promotion policies, career goals and intentions, and spouse's attitudes toward the Army. A description of the scales follows.

Needs and Expectancy of Fulfillment

A major focus of the model we hope to develop is on the impact of changing needs over different career stages. Thus, it was important to identify a wide variety of items that may take on differential importance over time. The items comprising the needs scale are 72 statements reflecting needs, wants, and goals that the interviewees identified as being important to their reenlistment decisions. These items covered the following areas: education, leadership, promotions, career, organizational policies, financial

---

[1] The views expressed in this paper are solely those of the author and do not reflect the views of the Dept. of the Army or the U.S. Army Research Institute

stability, job skills, family attitudes/values. Three scores were obtained for each item: an importance rating, a rating of their likelihood of occurrence within the Army (both based on a 5-point scale), and discrepancy score based on the difference between the two. The latter is the score of most interest and is intended to represent the degree to which soldiers perceive the Army as not meeting their needs.

Satisfaction

Extensive research links job satisfaction to turnover and to behavioral intentions to quit or stay which are significant precursors to actual turnover decisions (e.g., Mobley, Griffeth, Hand, & Meglino, 1979, Porter & Steers, 1973). Many approaches conceptualize job satisfaction in the context of match between a person's needs and the rewards provided by the work environment (LaRocco, Pugh, and Eric Gunderson, 1977; Scarpello and Campbell, 1983). It is important to note that, in considering satisfaction as it relates to the Army, it is not sufficient to consider what is specifically "job satisfaction." Serving in the military is a 24-hour job. Moreover, if the individual has dependents, they, too, in essence, are in the Army. Thus the soldier's personal, social, and work needs, including such things as meeting the needs of his/her family, play important roles in determining overall contentment with the Army. Therefore, any measure of satisfaction should encompass job satisfaction as we typicallly think of it as well as satisfaction with various aspects of Army life. The satisfaction subscale of the RICQ attempted to measure this factor on this more global level. The sixteen items covered satisfaction with areas such as : vocational skills acquired, the job, the quality of life, benefits, location, and "overall" satisfaction. A 5-point scale ranging from very dissatisfied to very satisfied was used.

Organizational Commitment

The third subscale was developed to measure commitment to the Army. The concept of organizational commitment has taken on an important role in the prediction of turnover. Most definitions of organizational commitment relate it to the individual's a) sense of involvement in, b) attachment to, and c) identification with an organization (including acceptance of and belief in its goals and loyalty to it) and willingness to put forth effort to remain in the organization. This definition follows from the psychological approach to commitment postulated by Porter and Smith (1970, as cited in Morris & Sherman, 1981). A second approach to organizational commitment derives from exchange theory, which operationally defines commitment in terms of what is required (i.e., what inducements are needed) to get an individual to leave the organization. From this perspective, the better the exchange or payoff as viewed by the individual, the greater the commitment to the organization. Most research has taken the first approach to commitment (Morris & Sherman, 1981; Mobley, et. al., 1979). Most of the 14 items in this scale followed from the psychological approach to commitment. In addition, several were based on the exchange theory approach and considered inducements to leave the organization.

Occupational Stress

Stress can lead to withdrawal behavior (absenteeism, turnover), tension, low job satisfaction, and a host of physiological conditions that are

symptomatic of it. Stress related to one's occupation can be the result of role conflict, responsibility, pressure, job versus non-job conflict, role ambiguity, workload, as well as not having one's needs met. Sixteen items were generated to reflect incidents exemplifying conditions that affect occupational stress. Soldiers were asked to indicate how often these events occurred on a 5-point scale ranging from "Never" to "Always".

Job Alternatives

A number of models of employee turnover (e.g. Mobley,et. al., 1979; Blueborn, 1982) have emphasized the importance of perceptions of the availability and attractiveness of alternative jobs. Within the military settings, job alternatives relate to both MOS changes within the Army and civilian job opportunities outside of it. As suggested above, civilian alternatives also can be attractive because they offer differences in lifestyle, not just employment. The Job Alternatives scale did not include this aspect: The 6 items were related to the job aspect only.

Method

Sample

Eighty-eight soldiers, primarily first-termers, pariticipated in the pilot testing of the RICQ. Criteria for selection were that they were within six months of ETS and eligible to reenlist.

Procedure

Soldiers were administered the questionnaire in groups of no more than six. Explanations of the instructions were given at the beginning of each new section of the survey. None of the scales were identified with the labels used here. For example, the occupational stress scale was merely labelled "Frequency Scale". In addition to debriefing subjects on the purposes of the research, lengthy discussion of the items themselves, reasons for their inclusions, and subjects reactions and suggestions was an integral part of the survey administration.

Analysis

Item analysis consisted of computation of item validities within scales (item-total correlations) and internal consistency reliability (Cronbach's alpha) for each of the scales. To identify whether more efficient scales were possible, after initial analyses, items with validities $<.35$ were removed from the scales and the computations were repeated. Criterion-related validity of the original and reduced-item scales was determined by correlating total scale scores with responses to a single 5-point item measuring Army career intentions.

Results and Discussion

Table 1 summarizes the results of the item analyses and computation of scale validities.

Table 1.
Reliabilities, ranges of item-total correlations, and criterion-related validities for original and reduced-item scales.

Scale	N of items	Alpha	Range of Item-Total Correlations	Validity	p
Needs: Discrepancy	72	.94	-.23 - .68	-.23	.0297
Satisfaction	16	.79	.15 - .68	.37	.0004
	10	.79	.27 - .66	.37	.0004
	9	.79	.37 - .65	.38	.0003
Commitment	14	.86	.18 - .74	.75	.0001
	12	.87	.39 - .68	.74	.0001
Stress	16	.76	.10 - .51	.23	.0277
	9	.77	.30 - .53	.29	.0068
	8	.76	.38 - .56	.28	.0081
Job Alternatives	6	.67	.22 - .51	-.42	.0014
	5	.67	.27 - .50	-.42	.0001
	4	.68	.37 - .53	-.49	.0001

Only results for the Discrepancy scores of the Needs scale are presented since these are the scores of most interest within this scale. Note that the alpha is very high. This appears to be due to a tendency for subjects to rate all the needs as highly important. This sample contained mostly first termers. Observations from the pilot and later data collections indicate that second and third termers do not share this same tendency. Note, also, that this scale contained almost as many items as their were subjects. Moreover, at this stage, since there is interest in the items individually rather than collectively, no attempts were made to reduce the number of items although 13 items had item-total correlations <.35.

Moderate to high reliabilities were obtained for the other four scales. In all cases, the number of items could be reduced without impacting on reliability. In the revised RICQ, six Satisfaction items were eliminated on the basis of the item analysis and four new ones were added based on the formal post-survey discussions with subjects. Similarly, two items were removed from the Commitment Scale and four were eliminated from the Stress Scale. All job alternatives items were retained since there were so few.

The individual scales showed significant, moderate to high validities for predicting career intentions. There was an inverse relationship between the

total discrepancy score and career intentions: The larger the discrepancy between important needs and perceived likelihood of their being met within the Army, the less likely was the soldier to be considering it as a career. This was also true for job alternatives scale: The more soldiers perceive other alternatives as more attractive, the less likely they are to want to be career Army.

Given that this research describes an experimental instrument which was piloted on a small sample, that each scale is at least moderately reliable and validly predicts career intentions is most encouraging. These appear to be good measurement tools for factors which are relevant to the research model we hope to develop. The research now in progress with the revised RICQ will focus on determining the interrelationships of factors which are most relevant to decisions to remain in or leave the Army, i.e. incentives and disincentives to reenlistment. In particular, it will look at the relationship of need fulfillment and expectancies that needs will be met, as they interact with other relevant variables, to the reenlistment decision. This will be considered, not just at one point in time, but at different career/life stages. Given personal growth over time, changes in marital and family status, etc., and accompanying changes in needs and values, it is likely that factors which influence the individual's satisfaction with and commitment to the Army and reenlistment decisions at later tenures will be different from those affecting initial reenlistment.

References

Bluedorn, A. C. (1982). Theories of turnover: Causes, effects, and meaning. Research in the Sociology of Organizations, 1, 75-128.

LaRocco, J. M., Pugh, W. M., & Eric Gunderson, E.K. (1977). Identifying determinants of retention decisions.Personnel Psychology, 30, 199-215.

Mobley, W.H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. Psychological Bulletin, 86, 493-522.

Morris, J. H., & Sherman, J.D. (1981). Generalizability of an organizational commitment model. Academy of Management Journal, 24, 512-526.

Porter, L. W., & Smith, F. J. (1970). The etiology of organizational commitment: A longitudinal study of initial stages of employee-organization relationship. Unpublished manuscript. As cited in Morris, J. H., & Sherman, J. D. (1981). Generalizability of an organizational commitment model. Academy of Management Journal, 24, 512-526.

Porter, L. W., & Steers, R. W. (1973). Psychological Bulletin, 80, 151-176.

Scarpello, V., & Campbell, J. P. (1983). Job satisfaction and the fit between individual needs and organizational rewards, Journal of Occupational Psychology, 56, 315-328.

PRIMARY MOTIVATIONAL VARIABLES IN MILITARY CAREER DECISION-MAKING[1]

Barbara L. McCombs
Denver Research Institute

Despite advancements that have been made in the identification of variables predictive of reenlistment decisions, improvements in models of the career decision-making process, and new developments in tests and measurement, the best predictive models of military enlistment decisions account for no more than 40 to 60 percent of the variance. The error in this prediction has significant cost implications from the standpoint of expenditures in training as well as in the retention of needed skilled personnel to support the efficient operation of a vast array of specialized jobs (Hicks & Nogami, 1984). For these reasons, the search continues for factors that can account for additional variance in the prediction of reenlistment decisions. The purpose of this paper is to describe a new conceptualization of the career decision-making process and the role of a class of individual difference variables in this process that has not been adequately attended to in previous retention or turnover prediction models. The paper also describes the methodology used to construct a battery of measures to assess these variables and the plans for validating the battery with Army enlisted personnel.

The class of variables hypothesized to be primary in the career decision-making process are self-system variables--those processes involved in the self-evaluation of competency and control that underlie motivation to pursue a particular course of action. The emphasis on these types of motivational processes is stressed because of the tacit assumption that human behavior is basically motivated by needs for self-determination and self-development, as well as the need to achieve a sense of personal competency in the achievement of personal development goals. In the employment context of the military, individuals express these needs in a variety of ways which are reflected in their perceptions, expectations, job satisfaction, performance, career intentions, and ultimate career decisions. If the basic motivational processes associated with these needs can be assessed and understood--in combination with the assessment and understanding of related situastional/environmental factors, and interrelationships with individuals' basic intellectual strengths and capabilities--military decision makers will be in a better position to select and match career options with available enlistees and enlisted personnel and thus maximize the probability of retaining needed personnel.

Background. It has been recognized that vocational decision-making involves a complex set of cognitive processes that individuals use to organize information about themselves and their vocational choices, to evaluate alternatives, and to commit to a particular action (Jepsen, 1983). In particular, recent work in human motivation theory by social, cognitive, and developmental psychologists has led to fairly general agreement regarding the particular importance of individuals' perceptions, expectations, and judgments of personal competency (self-efficacy) and personal causation (self-control) in influencing the motivational bases of decisions (e.g., Bandura, 1982, 1984, 1986; Cervone, 1986; Fishbein & Ajzen, 1975; Lefcourt, 1984; McCombs 1984, in press; Manderlink & Harackiewicz, 1984; Paris, Lipsom, & Wixson, 1983; Paris, Newman, & Jacobs, 1986; Weiner, 1976, 1980; White, 1959; Wittrock, 1986). In comparisons of three alternative models of military reenlistment decisions, Motowidlo and Lawton (1984) have found perceptions and expectations to be major determinants of intentions to stay and the final decision to stay. Landy and Becker (1985), however, have argued that there is a substantial amount of basic research still needed to understand how cognitive processes and abilities fit into various motivational models.

[1]This research was accomplished under Army Contract #MDA903-86-C-0114. The views, opinions, and/or findings contained in this document are the views of the author and should not be construed as the official position of ARI or as an official Department of the Army position, policy, or decision.

I believe that motivational theorists have now made the kind of progress necessary to elucidate the role of self-evaluative processes in motivation and decision-making. It is now widely accepted that individuals are active creators and constructors of their own knowledge and experience bases (e.g., Bandurda, 1982, 1984; Harman, 1973; Landy & Becker, 1985; McCombs, 1984, in press; Mischel, 1977; Wittrock, 1986). Those working in the areas of self theories have also generally agreed that the self is a compound set of multiple, hierarchically organized cognitive structures that exert a powerful influence on attention, organization and categorization of information, recall, and judgment (Eccles, 1984; Fleming & Courtney, 1984; Marsh, Parker, & Barnes, 1985; Paris & Cross, 1983; Pervin, 1985; Rogers, Kuiper, & Kirker, 1977; Shavelson & Bolus, 1982; Shavelson, Hubner, & Stanton, 1976). Several theorists have argued that the self acts as the background or setting against which new information, prior experiences, and knowledge are organized into personal schemas (Rogers et al., 1977); and that the self-structure is the largest and most available structure or set of structures in memory, and the central and first structure through which all information flows (Markus & Sentis, 1982; McCombs, in press). As such, the self has come to be conceived as an extremely active and powerful agent in the organization and evaluation of each individual's concept of reality and processing of personal data. When viewed in this light, it is clear that the self is the base set of filters (schemas) through which all information is acted upon and that every decision has a self-referent focus to a greater or lesser degree.

Two self-evaluative processes that have received increased attention are evaluations of personal competency and control. Individuals' needs to maintain positive evaluations of their worth and efficacy, and their needs to be masters of their own fate, influence their perceptions, expectations, and choices. It is argued that how people judge their capabilities for competence and control affects their motivation and behavior and the types of career matches they seek (Bandura, 1982, 1984; Landy & Becker, 1985; Lefcourt, 1984). As Bandura (1982, p.33) has stated, "Self phenomena lie at the very heart of causal processes because, not only do they function as the most immediate determinants of behavior, but they also give shape to the more distal external influences arising in transactions with the environment. Nevertheless, self processes have yet to receive the systematic attention in psychological theorizing and research they deserve."

Research that has been conducted on the role of personal control and competency evaluations has stressed their importance in influencing choice of activities and environmental settings (e.g., Bandura, 1982, 1984, 1986; Lefcourt, 1982, 1984). Generally this research suggests that people will avoid career situations they believe exceed their capabilities, but remain in situations they judge themselves capable of managing. In addition, persons with high needs for personal control will seek out those employment options that allow them to exercise their influences. Butler, Lardent, and Miner (1983) have argued that not only may turnover be due to certain motivational propensities in the individual that interact with aspects of organizational structure and process, but that this view of motivational fit has received little empirical or theoretical attention in the turnover literature. For a number of years, however, research evidence has been accumulating that indicates the importance of variables such as perceived control and competence in positive work attitudes, perceptions of task requirements, job satisfaction, motivation to persist, and success in training (e.g.. Chan, Karbowski, Monty, & Perlmuter, 1986; Dailey, 1970; Booth, Hoiber, & Webster, 1976; Booth, Webster, & McNally, 1976; Gunderson & Johnson, 1965; Kasperson, 1982; Lefcourt, 1984). In addition, early work on the turnover of Navy aviation trainees has indicated the importance of self-system variables such as needs for competence and control as discrimators of those trainees who voluntarily withdrew (Bucky, 1971; Bucky & Burd, 1970).

When motivational and personality variables have been used in turnover research, findings are somewhat disappointing in terms of additional variance accounted for. For

example, Booth et. al (1976) report that motivational variables, such as liking the career field, added only 3 to 8 percent to the variance accounted for in the prediction of success in Navy paramedical training. Furthermore, Arnold and Feldman (1982) have reported that a multivariate model of job turnover which included motivational variables only had an r=.44, with motivational variables contributing only an additional 1 percent to the variance accounted for. In spite of these findings, Motowidlo and Lawton (1984) have recently argued for the inclusion of affective and cognitive factors in models of reenlistment decisions. Included in these factors are perceptions, values, and beliefs. Although Motowidlo and Lawton do not specifically address the self-evaluative processes of competency and control, it is clear from self theories and research that these are primary types of perceptions and beliefs which are antecedents of job satisfactions, expectancies, intentions, and actual reenlistment decisions. Work by those interested in the enhanced prediction possible with self-system variables has suggested that as much as 20 to 30 percent of the variance could be accounted for by the inclusion of these variables in prediction models (e.g., Borman, Rosse, & Abrahams, 1980; Hoyle, 1986).

A significant problem in research with primary motivational variables, however, has been the lack of adequate definitions as well as carefully developed and well validated measures of these constructs (Lefcourt, 1984; Palenzuela, 1984). Until these steps have been taken, therefore, it is not possible to adequately evaluate the contribution of self-evaluative processes to the prediction of career decisions.

Method

Derivation of construct definitions. The basic approach to the definition of the constructs of personal control and competency was one in which available theory and research in this area was critically evaluated. From this evaluation, the following definitions were derived for the constructs of control and competence:

Control is generally defined as individuals' judgments and perceptions of their capabilities to be self-determining and the masters of their own fate, as well as their understandings of the contingencies responsible for success and failure. More specifically, these cognitive self-evaluation processes include (a) locus of control--underlying understandings regarding the locus of responsibility for events as internal (self) vs. external (others, fate); (b) personal control--perceptions of being able to exercise personal responsibility over events; and (c) attributions--tendencies to attribute reasons for successes and failures to internal (ability, effort) vs. external (luck, others) factors.

Competence is generally defined as individuals' perceptions or judgments about their capabilities to interact effectively with their environments and to execute the courses of action that are required to effectively handle particular situations. More specifically, these cognitive self-evaluation processes include (a) self-confidence--judgments of personal confidence with respect to specific capabilities or competencies; (b) adaptability--perceptions of capabilities to easily adjust to new requirements; (c) self-worth--judgments and perceptions of one's inherent value; and (d) competence--perceptions of abilities to exercise adequate control over one's actions.

Scale construction. In keeping with the construct definitions, measures of control and competence were developed to assess each of the dimensions of these constructs. In addition, given the support in the literature for both global and domain-specific assessments of self-system variables (e.g., Fleming & Courtney, 1984; Harter, 1985; Hoyle, 1986), and for trait and state assessments of both global and domain-specific variables (e.g., Bandura, 1982; Mischel, 1977; Nyquist, 1986; Spielberger et al., 1983), separate global

and domain specific measures of control and competence that met the criteria of assessing the theoretically based underlying constructs were examined. Items that were conceptually related to the constructs of interest were selected and modified to fit the global and domain-specific, trait and state assessment needs. New items were generated as necessary to obtain at least 10 items per construct subscale. The same items were used for the preliminary versions of the trait and state counterparts of global and the domain-specific measures, such that subsequent empirical evaluations could determine the best items for these respective scales.

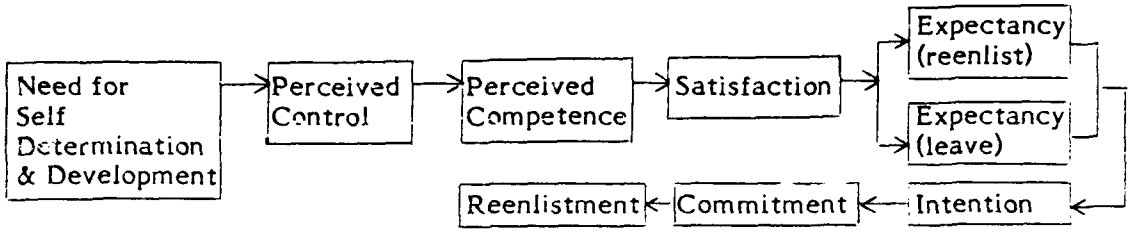
The resulting measures and number of items per subscale are as follows:

Subscales	Control				Competency			
	Global		Domain-Specific		Global		Domain-Specific	
	Trait	State	Trait	State	Trait	State	Trait	State
Control								
1 - Locus of Control	10	10	10	10				
2 - Personal Control	19	19	10	10				
3 - Attributions	11	11	10	10				
Competence								
1 - Self-Confidence					13	13	10	10
2 - Adaptability					10	10	10	10
3 - Self-Worth					15	15	10	10
4 - Competence					12	12	10	10
Total # of Items: (320)	40	40	30	30	50	50	40	40

In addition to the preceding eight measures and their respective subscales, a measure of importance was also developed to assess individuals' evaluations of the importance of being in control and of being competent in the job setting. Many theorists in the area of self-evaluative process assessment have argued that importance or valence measures are necessary to adequately assess the relationships between competence and control and criterion variables of interest (e.g., Bowman, et. al, 1980; Harter, 1985; Lefcourt, 1984; Motowildo & Lawton, 1984; Palenzuela, 1984), and thus items for assessing importance in both these areas were developed. The resulting 40-item importance measure consists of 20 competence items and 20 control items.

Preliminary Causal Model of Reenlistment Decisions

The following preliminary causal model builds on previous work in the area of turnover and reelistment, but adds detail on antecedent self-system variables that have generally been implied as perception or belief factors. The model can be graphically displayed as follows:



Mediating variables in the causal chain are assumed to be individual differences in cognitive abilities, importance ratings or values, as well as specific organizational or environmental variables.

Subsequent Research and Evaluation Plans

We are in the process of completing Phase I of an SBIR project for the Army Research Institute (ARI) in which preliminary data on the psychometric properties of the primary motivational process scales and on the potential predictive relationships between various elements in the preceding causal model and first term reenlistment decisions can be evaluated. The design of the Phase I study consists of the following features;

- Administering the motivational battery to a sample of enlistees (judged by their supervisors as "desirable" in terms of reenlistment) in two specialty areas (MOSs) that differ in the degree of competence and control requirements inherent in their structures;

- Conducting analyses to determine the battery's psychometric properties, including internal consistency and test/retest reliability, preliminary factor analyses, convergent and discriminant validity, and preliminary predictive validity; and

- Constructing a revised motivational battery consisting of the "best" items and scales that can be used in subsequent validation studies with larger samples of enlistees and can contribute to causal modeling and theory building that will support improved reenlistment prediction and military career advising.

In addition to measures of the variables identified in the preliminary causal model, our Phase I research includes measures of ability (ASVAB) and a computer-based ability/performance battery developed by the Essex Corporation. Preliminary data are also being collected via interviews and observations on situational variables as a means to better classify differences along the control and competency dimensions that exist in various MOSs, and which can subsequently be used to better determine matches of individual difference and job characteristics. A long term goal of continued research in this area is to derive possible areas of intervention that can enhance person-job matches and thus retention of needed personnel in the military.

Conclusion. In spite of mixed findings, a central theme that has run through the vast literature related to career decision-making is the complexity of the process and its highly variable nature due to unique individual differences in perceptions, interpretations, and evaluations of decision factors. I believe that if we can get at the basic self-evaluation processes that lie at the core of information processing, we will be better able to define decision factor paths that will greatly improve the reliability and validity of our prediction models. By knowing these basic self-system processes and their role in shaping decision paths, it is possible that we can find discernable patterns that will bring order to the complexity of career decision-making prediction.

References

Arnold, H. J., & Feldman, D. C. (1982). A multivariate analysis of the determinants of job turnover. Journal of Applied Psychology, 67(3). 350-360.

Bandura, A. (1982). The self and mechanisms of agency. In J. Suls (Ed.), Psychological perspectives on the self (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum.

Bandura, A. (1984). Representing personal determinants in causal structures. Psychological Review, 91, 508-11.

Bandura, A. (1986). Social foundations of thought and action: A social-cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

Booth, R. F., Hoiberg, A. L., & Webster, E. G. (1976). Work role motivation as a predictor of success in Navy paramedical training. Military Medicine, 141(3), 183-187.

Booth, R. F., Webster, E. G., & McNally, M. S. (1976). Schooling, occupational motivation, and personality as related to success in paramedical training. Public Health Reports, 91(6), 533-537.

Borman, W. C., Rosse, R. L., & Abrahams N. M. (1980). An empirical construct validity approach to studying predictor-job performance links. Journal of Applied Psychology, 65(6), 662-671.

Bucky, S. F. (1971, January). The California Psychological Inventory given to incomding AOC's and DOR's with normal and "ideal" instructions. Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Bucky, S. F., & Burd, J. (1970, September). Need satisfaction in the idenfication of the DOR. Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Butler, R. P., Lardent, C. L., & Miner, J. B. (1983). A motivational basis for turnover in military officer education and training. Journal of Applied Psychology, 68(3), 496-506.

Chan, F., Karbowski, J., Monty, R. A., & Perlmuter, L. C. (1986, April). Performance as a scource of perceived control. (Technical Memorandum 4-86). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Cervone, D. (1986, August). Availability biases, self-efficacy, judgments, and performance motivation. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Dailey, R. C. (1979). Locus of control, task attributes, and job performance. Perceptual and Motor Skillls, 49, 489-490.

Eccles, J. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), Achievement and achievement motives: Psychological and sociological approaches. San Francisco: W. H. Freeman and Company.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention and behavior: An introduction to theory and research. Reading, MA: Addison-Wesley.

Fleming, J. S., & Courtney, B. E. (1984). The dimensionality of self-esteem: II. Hierarchical facet model for revised measurement sclaes. Journal of Personality and Social Psychology, 46(2), 404-421.

Gunderson, E. K., & Johnson, J. C. (1965). Past experience, self-evaluation, and present adjustment. Journal of Social Psychology, 66, 311-321.

Harman, H. H. (Ed.) (1973, December). Proceedings: Toward the development of more comprehensive sets of personality measures. Symposium of the American Psychological Association, Montreal. Princeton, NJ: Educational Testing Service.

Harter, S. (1985). Processes underlying self-concept formation in children. In J. Suls & A. Greenwald (Eds.), Psychological perspectives on the self. Hillsdale, NJ: Lawrence Erlbaum.

Hoyle, R. H. (1986, August). Factor analytic study of the dimensionality of self-esteem. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Jepsen, D. A. (1983). Career decision-making. In L. Harmon (Ed.), The individual's use of information in career development: From cognitive to computers. Columbus, OH: National Center for Research in Vocational Education, Ohio State University.

Kasperson, C. J. (1982). Locus of control and job dissatisfication. Psychological Reports, 40, 823-826.
Landy, F. J., & Becker, W. S. (1985, July). Adaptive motivation theory. University Park, PA: Pennsylvania State University.
Lefcourt, H. M. (1984). Research with the locus of control construct (Vol. 3). New York: Academic Press, Inc.
Manderlink, G., & Harackiewicz, J. M. (1984). Proximal versus distal goal setting and intrinsic motivation. Journal of Personality and Social Psychology, 47(4), 918-928.
Markus, H., & Sentis, K. (1982). The self in social information processing. In J. Suls (Ed.), Psychological perspectives on the self (Vol. 1, 41-70). Hillsdale, NJ: Lawrence Erlbaum.
Marsh, H. W., Parker J., & Barnes J. (1985). Multidimensional adolescent self-concepts: Their relationship to age, sex, and academic measures. American Educational Research Journal, 22(3), 422-444.
McCombs, B. L. (1984). Processes and skills underlying continuing intrinsic motivation to learn: Toward a definition of motivational skills training interventions. Educational Psychologist, 19(4), 199-218.
McCombs, B. L. (in press). The role of the self-system in self-regulated learning. Contemporary Educational Psychology.
Mischel, W. (1977). On the future of personality measurement. American Psychologist, 32(4), 246-254.
Motowidlo, S. J., & Lawton, G. W. (1984). Affective and cognitive factors in soldiers' reenlistment decisions. Journal of Applied Psychology, 69(1), 157-166.
Nyquist, L. V. (1986, August). The dynamic self-concept: Cognitive and behavioral responses to challenge. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.
Palenzuela, D. L. (1984). Critical evaluation of locus of control: Towards a reconceptualization of the construct and its measurement. Psychological Reports, 54, 683-709.
Paris, S. G., & Cross, D. R. (1983). Ordinary learning: Pragmatic connections among children's beliefs, motives, and actions. In J. Bisanz, G. L. Bisanz, & R. Kail (Eds.), Learning in children: Progress in cognitive development research New York: Springer-Verlag.
Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. Contemporary Educational Psychology, 8, 293-316.
Paris, S. G., Newman, R. S., & Jacobs, J. E. (1985). Social contexts and functions of children's remembering. In M. Pressley & C. J. Brainerd (Eds.), Cognitive learning and memory in children. New York: Springer-Verlag.
Pervin, L. A. (1985). Personality: Current controversies, issues, and directions. Annual Review of Psychology, 36, 83-114.
Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. Journal of Personality and Social Psychology, 35(9), 677-688.
Shavelson, R. J. Hubner J. J., & Stanton, G. C. (1976). Validation of construct interpretations. Review of Educational Research, 46, 407-441.
Shavelson, R. J., & Bolus, R. (1982). Self-concept: The interplay of theory and methods. Journal of Educational Psychology, 74, 3-17.
Spielberger, C. D. et al. (1983). Manual for the state-trait anxiety inventory. Palo Alto, CA: Consulting Psychologists Press, Inc.
Weiner, B. (1976). Attribution theory, achievement motivation, and the educational process. Review of Educational Research, 42, 201-215.
Weiner, B. (1980). The role of affect in rational (attributional) approaches to human motivation. Educational Researcher, 9(7), 4-11.
White, R. W. (1959). Motivation reconsidered: The concept of competence. Psychological Review, 66, 297-333.
Wittrock, M. C. (1986). Student thought processes. In M. C. Wittrock (Ed.), Handbook of research on teaching (3rd ed.). New York: McMillan.

EFFECTS OF VARYING ITEM FORMAT ON TEST PERFORMANCE

Lee E. Day
William F. Kieckhaefer
RGI, Incorporated

BACKGROUND

The Department of Defense is considering implementing computerized adaptive testing to replace paper-and-pencil versions of the Armed Services Vocational Aptitude Battery (P&P-ASVAB). The computerized adaptive version of that test (CAT-ASVAB) is expected to increase the efficiency of selecting and classifying new accessions. (For more information, see Sands, 1984, 1985.)

The item bank for the operational CAT-ASVAB consists of five-option items from the omnibus item pool provided by the Air Force Human Resources Laboratory. Additional items which may be needed to supplement this item bank are available from the experimental CAT-ASVAB item bank. However, these items have four options. Therefore, it is necessary to examine whether mixing items with different formats within CAT-ASVAB subtests will affect examinee performance on the test.

Only one publication relevant to this issue was found. Brittain and Vaughan (1984) studied the effects of mixing items with different numbers of options within the Army Skills Qualification Test. They predicted errors would increase when an item with **n** answer options followed an item with more than **n** answer options, where errors were defined as choosing a non-existent answer option. Consistent with their hypothesis, mixing items with different numbers of answer options caused an increase in errors. However, the authors reported that the magnitude of this increase was insufficient to conclude that the increase in errors was of practical significance.

PURPOSE

The purpose of this study was to examine whether mixing four- and five-option items within subtests of the experimental CAT-ASVAB would increase examinee errors. Since the CAT-ASVAB is being considered as a replacement for the P&P-ASVAB, it is important to examine aspects of the CAT-ASVAB which could affect examinees' performance on the battery. To this end, the research questions investigated were to determine whether mixing item formats affects either: (1) item difficulty, (2) test difficulty or (3) response latencies.

METHOD

Examinees

Examinees in this study were 1200 male Navy recruits at the Recruit Training Center, San Diego, California. Test administrators randomly selected the examinees from available recruits and seated them at individual testing stations. (See Segall, Kieckhaefer, and Day, 1986, for greater detail on the experimental CAT-ASVAB testing equipment and the test instructions.)

Test Administration Procedures

Word Knowledge (WK) and Paragraph Comprehension (PC) subtests were administered by computer using a conventional (i.e., nonadaptive) strategy. The item bank for WK consisted of 64 items; items one through 32 had five options, and items 33 through 64 had four options. Similarly, the PC item bank consisted of 48 items; items one through 24 had five options, and items 25 through 48 had four options. For each four-option item in a subtest, there was a corresponding five-option item which was matched as closely as possible on the item characteristic curve. The software presented some combination of half of the items in a subtest item bank according to the experimental condition. These items were presented in order of increasing difficulty. To control for item position effect, the software presented a given item in the same position within the test in all experimental conditions.

Examinees were randomly assigned to one of six experimental conditions for each subtest, with the restriction that there were 200 examinees per condition. Two conditions were control: the first condition contained all five-option items and the second contained all four-option items. The other four conditions were mixed format: the third condition consisted of alternating five- and four-option items (Mixed - 1:1), the fourth condition consisted of four-option items embedded in five-option items such that every fourth item had four options (Mixed - 3:1), and the fifth condition consisted primarily of five-option items with every eighth item having four options (Mixed 7:1). Finally, in the sixth condition four- or five-option items were randomly selected with the restriction that the examinee received an equal number of four and five option items (Mixed-Random).[1]

RESULTS

Item Difficulty

To compare item difficulties, two-by-two tables were constructed for each item in each mixed format condition. These tables contained the proportions of examinees passing and failing each item in the mixed format condition compared with those proportions on the same items in the appropriate control condition. A Chi-square test assessed the significance of the comparison.

Table 1 summarizes the results of the Chi-square tests. For the WK subtest, only seven of the 160 comparisons (about 4.4%) produced significant Chi-square tests. For the PC subtest, only one of the 120 comparisons (about 0.8%) produced a significant Chi-square test. For all significant differences, the proportion of examinees in the mixed condition was larger than in the control condition.

Test Difficulty

For examinees in each mixed format condition, except the Mixed-Random condition, two number-correct scores were computed for each examinee: one for the four-option items and one for the five-option items. An independent t-test

---

[1] D. O. Segall and K. E. Moreno of the Navy Personnel Research and Development Center provided the tests and experimental design.

Table 1

Number of Significant Chi-Square Tests
Comparing Mixed Format and Control Conditions on Item Difficulty

		WORD KNOWLEDGE SUBTEST		PARAGRAPH COMPREHENSION SUBTEST	
CONTROL CONDITION	EXPERIMENTAL CONDITION	No. Items	No. Significant Chi-Square Tests	No. Items	No. Significant Chi-Square Tests
Five-Option Items	Mixed - 1:1	16	Ø	12	Ø
	Mixed - 3:1	24	Ø	18	Ø
	Mixed - 7:1	28	1	21	Ø
	Mixed - Random	32	3	24	Ø
Four-Option Items	Mixed - 1:1	16	2	12	Ø
	Mixed - 3:1	8	1	6	1
	Mixed - 7:1	4	Ø	3	Ø
	Mixed - Random	32	Ø	24	Ø
Column Totals		160	7	120	1

Note. Two hundred (200) examinees were in each control and experimental condition for a total of 1200 examinees. For all significant differences, the proportion of examinees passing the item in the mixed condition was larger than in the control condition.

assessed the significance of the difference between these scores and the number-correct scores computed on the same sets of items for examinees in the control conditions.

Table 2 presents the results of **t**-tests on test difficulties. The first column shows which mixed-format condition was compared with the control condition indicated. The second column shows the number of items considered in the number-correct score. The remaining columns show results for the WK and PC subtests. For test difficulty, the table shows both the **t**-value and the power of the **t**-test to detect a difference equal to one half of a standard deviation (see Cohen, 1977, pages 20-27) on the paper-and-pencil subtest for WK (35 items) or PC (15 items), adjusted for the number of items used to compute the number-correct score. This standard deviation was computed from the pre-enlistment paper-and-pencil ASVAB scores for a sample of 7515 recruits from all four Armed Services (Day, Kieckhaefer, and Segall, 1986).

Response Latencies

For each mixed format condition except the Mixed-Random condition, average response latencies were computed for each examinee on two sets of items: one

for the four-option items and one for the five-option items. Response latencies were also computed on the same sets of items for examinees in the control conditions. Then, independent t-tests compared the average response latencies between each mixed-format and the appropriate control condition. Because different numbers of examinees received each item in the Mixed-Random condition, a t-test was not computed to compare the response latencies of the Mixed-Random condition with a control.

Table 2 also presents the results of t-tests on average response latencies. The third column of results shown for each subtest presents the t-values for response latencies. None were statistically significant.

Table 2

Results of t-Tests Comparing Mixed Format and Control Conditions on Test Difficulties and Response Latencies

	WORD KNOWLEDGE SUBTEST				PARAGRAPH COMPREHENSION SUBTEST			
		Test Difficulty		Response Latency		Test Difficulty		Response Latency
EXPERIMENTAL CONDITION	No. Items	t-value	Power[a]	t-value	No. Items	t-value	Power[a]	t-value
Comparison With Five-option Control								
Mixed - 1:1	16	.06	.90	-.85	12	-.08	.88	-1 77
Mixed - 3:1	24	-1.09	.94	.47	18	-.21	.93	- 64
Mixed - 7:1	28	-.24	.93	-.98	21	-1.82	.93	.67
Comparison With Four-option Control								
Mixed - 1:1	16	-1.83	.99	1.49	12	1.30	.97	-.72
Mixed - 3:1	8	-1.35	.94	1.84	6	-.98	.85	-1.92
Mixed - 7:1	4	1.35	.86	-.07	3	-1.40	.65	-.28

Note. There were 200 examinees were in each control and experimental condition (1200 examinees total). The t-values shown are from independent t-tests between Control and Experimental Conditions. Positive t-values indicate a larger mean value for the control condition. None are statistically significant.

[a]The values shown are for the power of the t-test, where power is defined as the ability of the t-test to detect a difference equivalent to one half of a standard deviation on the paper-and-pencil subtest.

CONCLUSIONS

Comparison of the proportion of examinees passing an item across test conditions indicated that varying the number of response options did not increase item difficulty. Only 4.4% of the WK comparisons and only Ø.8% of the PC comparisons were significant. Furthermore, comparing number correct scores produced no significant differences in test difficulty across test conditions. These results support the following conclusion:

(1) Mixing items with different numbers of response options does not increase item or test difficulty.

Finally, comparisons of response latencies across test conditions produced no significant differences. Therefore, the results of these analyses indicate that:

(2) Response latencies are not affected by mixing items with different numbers of response options.

REFERENCES

Brittain, C. V., & Vaughan, P. R. (1984). Effects of mixing two-option and multi-option items in a test. **Proceedings 26th Annual Conference of the Military Testing Association Volume I** (369-374). Munich, Federal Republic of Germany: Psychological Service of the German Federal Armed Forces.

Cohen, J. (1977). **Statistical power analysis for the behavioral sciences.** New York: Academic Press.

Day, L. E., Kieckhaefer, W. F., & Segall, D. O. (1986). **Predictive Utility evaluation of computerized adaptive testing in the Armed Services.** San Diego, CA: RGI, Incorporated.

Sands, W. A. (1985). An overview of the accelerated CAT-ASVAB Program. **Proceedings 27th Annual Conference of the Military Testing Association Volume I** (19-22). San Diego, California: Navy Personnel Research and Development Center.

Sands, W. A. (1984). Computerized adaptive testing (CAT) for the U. S. armed forces: Progress and plans. **Proceedings 26th Annual Conference of the Military Testing Association Volume I (327-331).** Munich, Federal Republic of Germany: Psychological Service of the German Federal Armed Forces.

Segall, D. O., Kieckhaefer, W. F., & Day, L. E. (1986). **Predictive utility evaluation of computerized adaptive testing in the Air Force.** San Diego, CA: RGI, Incorporated.

LIVE-FIRE TRAINING

Dr S G Lister, Army Personnel Research Establishment, UK

1. INTRODUCTION

The experience of firing live ammunition has always been considered to be a central part of an effective training system yet very little work has been conducted that has specifically addressed how much 'live' training is necessary or how its use might be optimized. This paper reports two studies recently conducted at the United Kingdom Army Personnel Research Establishment (APRE). Both studies concerned the Royal Armoured Corp (RAC), one looking at Main Battle Tank (MBT) training, the other training for the Anti-Tank Guided Weapon (ATGW) system, Swingfire.

These studies were both prompted by the very high cost of live ammunition and the consequent need to ensure that the minimum amount of ammunition is used that will produce the desired level of proficiency. It was appreciated that this amount is largely dependent upon the quantity and quality of simulator training beforehand as well as the nature of classroom instruction. As such these studies were part of a much larger area of research within APRE's Personnel Psychology Division and the studies reported here were primarily concerned with the varying amounts of ammunition following the existing classroom and simulator training.

2. MAIN BATTLE TANK LIVE FIRE TRAINING

Tank gunners spend a high proportion of their time training on the principal training simulator, the Gunnery Training Simulator (GTS). This comprises a tank turret mounted on a gantry, in front of which is a large screen onto which is projected a slide of a landscape. Mounted on this are a number of static and moving targets, the latter travelling along fixed tracks on the screen. A record is kept of various timings concerned with the engagement sequence.

On completion of simulator training, new recruits go to the nearby Warcop ranges. Ammunition scales were set some 4 years ago although they have recently been reduced to about 81% of this level (or 81% Training Datum Level, or TDL). It is notable that currently no objective assessment of this live fire practice is made. Subsequently, after the now trained tank gunners have joined their regiments, they have the opportunity once a year to attend a crew annual firing practice. Again ammunition scales were set some 4 years ago but were subsequently reduced to 81% of their original level. Between these practices, gunners continue to have access to a GTS.

The aim of the experimental trials reported here was to establish the amount of live ammunition that should be fired

following standard use of the GTS both for initial recruit training and crew annual firing practices. A susiduary aim was to assess whether an increase in training time spent on the GTS, that could be accommadated by range time being saved with certain ammunition reductions, could produce satisfactory standards. The work described here comprises two trials concerned with novice gunners and trained crews respectively. These are decribed separately below.

2.1 Novice Gunners

2.1.1 Method

The Subjects were 42 trainee gunners based at Catterick. They were randomly allocated to one of five Experimental Groups firing 50, 60, 70, 80, and 100 percent TDL, as shown in Table I.

Group	GTS training	%TDL fired	Size of Groups
1	Extra	50	8
2	Extra	60	9
3	Normal	70	8
4	Normal	80	8
5	Normal	100	9

Table I - Novice Gunners, Experimental Groups

The ammunition scales fired by Groups 1 and 2 meant that a complete day could be saved on the range and spent on GTS training instead. This permitted an increase in the number of simulated engagements fired from 134 to 164 (about 22%). After the Groups had undergone the GTS training shown in Table I, an assessment test was then conducted on the GTS. This was designed by the RAC to represent the demands of tank gunnery, comprising a range of target types and engagement methods. Then, following live fire training, a further assessment test was administered, using live ammunition. This was designed to be as nearly as practically possible the same as the GTS assessment test. Both tests comprised 6 targets which were engaged until hit, up to 3 rounds maximum being permitted during a 60 second period. A record of the time to each round fired, and its outcome, was made using high powered lenses.

2.1.2 Results

The results of the GTS and live fire tests are described separately below.

GTS Test

A comparison between Normal (Groups 3, 4 and 5) and Extra (Groups 1 and 2) GTS training was made in terms of time to hit target. The result was nonsignificant ($F=0.83$).

Live fire test

Performance on the live fire test was scored in 2 ways, hit probabilities and time to hit.

Hit Probabilities. The number of hits was calculated for each gunner at 2 time limits, 13 seconds (the estimated operational standard expected of gunner not having to interact with a Commander) and 60 seconds (the total exposure time for each target). At both time criteria the comparison between the 5 groups proved nonsignificant (H=4.52 and 4.64 respectively).

Time to first hit. Performance was analyzed separately for each of the 6 targets. All comparisons revealed a nonsignificant difference between the 5 Group (F=0.67 to 2.33).

2.2 Trained Crews

2.2.1 Method

The Subjects were 58 crews at their Annual Firing Practice at Hohne, Germany. They were organized into 5 Experimental Groups. These Groups fired 50, 60, 70, 80 and 100 percent TDL, as shown in Table II.

Group	%TDL fired	Size of Group
1	50	11
2	60	12
3	70	12
4	80	11
5	100	12

Table II - Trained Crews, Experimental Groups

In these trials, GTS usage was constant across the 5 Groups, quantity of live ammunition being the sole variable. After firing the different TDLs, the subjects all fired a live fire assessment test. Again this was designed by the RAC to reflect the demands of tank gunnery. Each target was exposed for 60 secs during which time the crew could fire up to 3 rounds. A record of the timing to each round fired, and its outcome, was made using high powered lenses.

2.2.2. Results

Performance was again scored in 2 ways, hit probabilities and time to first hit.

Hit Probabilities. This was calculated for 2 time intervals, 18 seconds (the required operational standard of a crew) and 60 seconds (the total exposure time for each target). It was found that although all 5 Groups perform at about the same level at

the 18 second criteria (H=5.62, NS), after 60 seconds there was a significant difference between them (H=15.39, Sig 1%). Further analysis of this difference reveals that performance at either 50 or 60 percent TDL is significantly poorer than performance at 80 or 100 percent TDL.

Time to First Hit. Performance on each target was analyzed separately. All resultant comparisons were nonsignificant (F=0.72 to 1.24).

3. SWINGFIRE LIVE-FIRE TRAINING

Swingfire is a long range anti-armour systen, able to engage tanks at all ranges from 150 to 4000 metres. After an initial automatic gathering phase, the operator assumes manual control using a thumb-stick and guides the missile onto its target. In total he may be required to track for up to about 25 seconds.

Simulator training at the time the trials were conducted was achieved using a tracker trainer that injects a simulated target into the operators sight. This gives him the oppurtunity to fire many hundreds of simulated engagements each year.

The trials desribed here concerned only the training requirements of already experienced operators and had the aim of establishing wnether 3 or 4 live missiles should be fired at Annual Firing Practices both with regard to the immediate and long term relative benefits.

3.1 Method

The Subjects were 30 Swingfire Operators at their Annual Firing Practice at the Otterburn ranges. The trials were in 2 phase.

Main phase. The trials involved a group of 30 subjects firing 4 missiles each. In order that the training value of the firing camp was maintained, it was required that the group fire missiles 1 and 2 at targets and in a method chosen by the Commanding Officer. The resultant trials design is shown in Table III. It can be seen that Missiles 3 and 4 were fired by the same method and at targets of similar range

Missile	Target
1	at regiments discretion
2	at regiments discretion
3	direct optical static
4	direct optical static

Table III - Swingfire trials design

A Control Group was also identified, who fired 3 missiles at the same firing camp. This permitted a follow-up study, at the next Annual Firing Practice, of operators who fired 3 or 4 missiles one year previously.

The Directorate of Land Service Ammunition (DLSA) made closed circuit television (CCTV) recordings of all 3rd and 4th missiles. These were subsequently analyzed and hit probabilities calculated.

Follow-up. This was conducted a year after the main phase of the trials and costituted the normal Annual Firing Practice of those who participated in the initial phase together with those previously identified as Controls. However, it was only possible to include 10 of the original 30 participants in the follow-up study and it was decided only to use a 6 point subjective scales completed by instructors as a measure of operator performance.

3.2 Results

Main Phase. The results of the comparison between the Experimental Group's 3rd and 4th missile DLSA rating showed that performance was superior with the 3rd missile (F=8.46, sig 1%).

Follow-up Phase. The reliability of the subjective rating scales is in some doubt and it should be noted that an initial validation of the rating scales during the main phase of the study revealed a nonsignificant correlation with the DLSA scores (Kendalls Correlation Coefficient =0.22). With this reservation, the results of the subjective ratings are summarized in Table IV.

	1	2	3
Experimental	3.80	3.61	4.22
Control	3.89	3.00	4.00

Table IV - Mean Subjective Ratings for the 3 missiles fired in the follow up study.

The results shown in Table IV were analyzed to see if there were differences between the Experimental and Control Groups. This revealed that there were no significant differences with any missile. The Experimental and Control Group data were combined in order to assess any differences in performance between missiles. Again, this gave a nonsignificant result. That is to say, performance with missiles 1, 2 and 3 were all very similar, there being no evidence of learning from missile 1 to 2, or from 2 to 3.

4. DISCUSSION

The results from both studies support the view that satisfactory training can be achieved using relatively small amounts of training ammunition. This has been particularly clearly demonstrated in the case of the Main Battle Tank where a reduction to 50% has been possible without adversely affecting performance in the case of the novice gunner and to 70% in the case of the trained crew. With Swingfire, firing 3 rather than 4 missiles annually appears to result in no loss in performance effectiveness and there is some evidence from the subjective ratings during the follow up study that only 1 missile may be required. Clearly, further work is required before this scale of reduction in training ammunition could be recommended.

The question that is raised by both studies concerns the way in which live ammunition should be used during training. In the case of both Swingfire and MBT, trainees will have fired many simulated rounds before going to the practice range. Accordingly, it would be surprising if operator ability were much improved by firing a few more live rounds. The skill of firing the weapon system should be thoroughly learned by this point, provided the simulator is giving adequate transfer of training. Rather live fire training should be expected to provide exposure to the atmosphere of firing the actual weapon system and, in doing so, enhance morale and boost confidence. It is important, however, to specify precisely what training can be accomplished during live fire training for a particular weapon system as this affects the nature of the training aids that will be required. Particularly, the requirement for complex full mission simulators may be lessened when it is possible to give operators some experience of using the weapon system itself.

It follows from an analysis of the training value of live-firing in terms of the generation of confidence that there is something to be learned from the experiences of other trainees. Currently, little attention is paid by those not firing to what is happening on the range. The use of Closed Circuit Television (CCTV) Cameras and intercom should enable all trainees to get some training value from all missiles that are fired. Currently the question of range instrumentation is being investigated and, undoubtabley, great improvements can be made at very little cost.

Finally, considerabley more thought is required into the engagement types that should be fired with the reduced ammunition scales that these studies have suggested might be adopted. For example, firing certain engagements have greater training value than others, requiring that certain skills common to all engagements are practiced plus certain others in addition. Clearly all engagement types should be looked at from this point of view.

Copyright c Controller HMSO. London. 1986

FRONT-END ANALYSIS FOR U. S. MARINE CORPS TRAINING: INDIVIDUAL TRAINING STANDARDS

DOUGLASS DAVIS, Ph.D.,
Task Analysis and Design Specialist
Commandant of the Marine Corps (TAP 31)
Washington, DC 20380-0001
(202) 694-2404/1551

ABSTRACT

By the end of fiscal year 1989, the Deputy Chief of Staff for Training (Code T), Headquarters U. S. Marine Corps will have analyzed all Marine Corps occupational specialties (MOS). The results of this concentrated front-end analysis will be individual training standards (ITS) that specify discreet tasks and standards for individual training within each MOS, for all aspects of training, formal and other.

Individual training standards developed for Marine Corps ground, air, and professional development training programs are based upon extensive occupational analyses accomplished through surveys, onsite interviews and validations, and verification of task lists and standards by subject matter experts.

This paper reviews the overall plan and methods used to develop ITS and explains the identification and application of tasks and standards common between or among MOS's and occupational fields. It also explains the placement of ITS within various training programs and the relationships among ITS, collective (team) training standards, and mission readiness of units (presently evaluated by the Marine Corps Combat Readiness Evaluation System).

Being developed in tandem with ITS is the Computer Assisted Systems Approach to Training (CASAT), an automated ITS development and maintenance system. That system is discussed in terms of a front-end-analysis data base and the role that data base should play in an external feedback and appraisal system that could serve to ensure that the data base is current.

U. S. Marine Corps individual training standards (ITS) specify the tasks that marines are to be trained to perform, the conditions under which the tasks are to be trained, and the degree of proficiency to which each task is to be performed. They are more than just standards, as standards are usually defined; roughly, they equate to broad, terminal learning objectives. The Marine Corps has been systematically developing training standards for a

long time. Mission Performance standards of the Combat Readiness and evaluation System and the aircrew standards described in the Training and Readiness Manuals are examples of collective (team) standards. Physical fitness and marksmanship requirements are examples of standards that apply to individuals.

The Training Department (Code T) at the Headquarters, U. S. Marine Corps, has initiated an extensive project to develop ITS. By 1990, all categories of training, including formal school, unit, on-the-job training, etc., will have been evaluated for possible ITS development. More than 750 military occupational specialties (MOSs) will be reviewed; occupational fields (OccFlds) formed by these MOSs will also be reviewed and, if applicable, recommendations for changes will be made. Hundreds of Marines will participate in ITS development; all Marines will have been affected when ITS development has been completed.

In 1981, the Commandant decided to accelerate the development of standards by establishing a Training Department at HQMC, staffing the Department with nearly twice the number of personnel that were in the previous training division, putting a Major General in charge and giving the Department the mission of ". . . formulation, development and publication of individual and collective training standards for all categories of training conducted in Marine Corps units and institutions . . . ."

The development of standards is only one third of the Training Department mission. The Deputy Chief of Staff for Training is also tasked with establishing training policy and allocating resources. In carrying out these responsibilities, a systems approach to training (SAT), frequently referred to as instructional systems development (ISD) is employed.

Before the SAT was implemented by Marine Corps Order 1553.1, most Marine Corps training was based upon intuitive judgment that resulted from personal experience. Courses contained what people thought should be taught. As a result, some critical skills were omitted and others, not as critical, were included. SAT has changed all that by replacing intuitive judgment with information obtained from Marine Corps doctrine and from individuals working in operational units. What is to be trained is based upon an analysis of data obtained from various sources. This ensures that training is provided for those tasks that must be performed by Marines.

The SAT process analyzes training requirements, translates these requirements into training-objective format, selects the proper training strategy, develops effective training delivery systems and provides quality control. It is a systematic but flexible process ensuring that Marines acquire the knowledges and skills needed to accomplish missions. The goal of the process is to get the most out of the resources invested in training by improving performance on-the-job and/or decreasing the amount of the investment.

Since it was adopted in 1969, the Marine Corps SAT process has not changed significantly. What has changed is the responsibility for carrying out the phases of the process. There are five phases in the process: analysis, design, development, implementation, and evaluation. The most time consuming phase is analysis. The major difficulty with implementing SAT in the past was the limited numbers of personnel available in units and schools to conduct analysis. For that reason, the responsibility for conducting analysis was assumed by Headquarters.

Analysis is a systematic set of procedures used to determine what an individual or unit is supposed to do to perform a job or mission sucessfully and what must be accomplished during training world to prepare for that job or mission. The product of the analysis phase is a training standard and, in the case of OccFlds, an evaluation of the OccFld structure.

After the standards are developed, the design and development of training programs, (Phases II-IV of the ISD Model) take place at formal schools and at the Marine Corps Institute in Quantico.

During the analysis phase of ITS development, each task that Marines in a particular billet or MOS are required to perform is assigned to either a formal school or to unit commander in training. Formal school directors prepare programs of instruction based on the tasks provided. The Marine Corps Institute is responsible for developing standardized training packages for those tasks assigned to unit commanders. The packages will be provided to unit commanders to use in their training programs as they see fit.

The priority for ITS development is reevaluated semiannually. Changes to the priority are based on input solicited from FMF (Fleet Marine Force) commanders and principal staff officers at HQMC.

Although the need to develop training standards was established by the BGen Sardo study and the responsibility to do it was assigned to the Training Department, it is in the best interest of the Marine Corps to ensure that we do not duplicate the efforts of others. With that in mind, we conduct a needs assessment and needs analysis prior to beginning standards development. Some categories of training may not merit standards development because appropriate standards have been developed by another service or by another Marine Corps agency; development of standards for a particular category of training is not cost effective; or Marines are trained in other service schools.

The three branches (air, ground, and professional development education) in the Analysis and Design Division each have an analysis and design section responsible for developing standards for air, ground and professional development education, respectively.

We are unable to develop standards in a reasonable period of time without additional resources. We were successful in POM85 and have been authorized funds for training standards in FY85 and FY 86. However, we need to compete to obtain funds for FY88, and FY89.

The additional funds will be used to obtain contractor support. The Naval Training Systems Center in Orlando has assisted us in contracting for aviation training standards. The Naval Personnel Research and Development Center provides the same support for ground training standards. During FY87, the Department of Energy regional office in Idaho Falls and Oak Ridge will also assist.

Individual Training Standards are being developed to provide task standards for individual performance. The Marine Corps also has plans to develop collective training standards (CTS) for team training. Already in existance are mission performance standards (MPS) that indicate a unit's readiness for combat, or to perform missions. The Marine Corps Combat Readiness System (MCCRES) evaluates units in terms of their ability to perform missions.

The direction in which the Marine Corps is heading supports three levels of standards: ITS, CTS, and MPS. Although consideration has been given to the development of a synergistic system of Marine Corps standards that would support the mutual interplay of all three levels, there are no firm plans to develop such a system. The Navy Personnel Research and Development Center (NPRDC) is studying the feasibility of a single Marine Corps standards system.

In confluence with the development of training standards is the development of CASAT (Computer Assisted Systems Approach to Training), an automated data system that will support the development and evaluation of ITS. Development of CASAT is being accomplished through consideration of the requirements that would support a system of individual, collective, and mission standards. It is now too early in CASAT development to know exactly if and how this can be accomplished; however, through careful planning and employment of the NPRDC study findings and recommendations, the next logical step will be the development of an all-inclusive Marine Corps standards system. Such a system would evaluate performance at the unit, team and individual levels and would indicate those weaknesses that must be attended to at each appropriate level in order to ensure that the Marine Corps has a means to measure combat readiness in terms that can be interpreted effectively and efficiently.

MATMEP:
INDIVIDUAL TRAINING STANDARDS
FOR
AVIATION MAINTENANCE MILITARY OCCUPATIONAL SPECIALTIES

LEWIS F. ROGERS, Ph.D.
Colonel USMCR
Officer-in-Charge, MATMEP Development Team
Commandant of the Marine Corps (TDA)
NAS Memphis, Millington, TN 38054
(901) 872-5315/5316

The Marine Corps' Individual Training Standards System (ITSS) is, by design, a simple program devised to train Marines as realistically as possible. The program's goal is to find out what Marines need to know, teach them what they don't know, and then test them to see if they can perform their jobs. The ITSS is a performance based process, the means to an end, resulting in usable, hands-on training developed by Marines for Marines.

To avoid "reinventing the wheel", the Training Department, Headquarters Marine Corps, adopted the Fourth Marine Aircraft Wing's Maintenance Training Management and Evaluation Program (MATMEP) as the ITSS for the aviation maintenance occupational fields. MATMEP was conceived and implemented by the 4TH MAW to address problems of reserve training/billet mismatch caused by demographics. MATMEP drew from and embraced existing Navy and Marine Corps training programs, replacing several duplicate programs with a single standard approach to training management and evaluation.

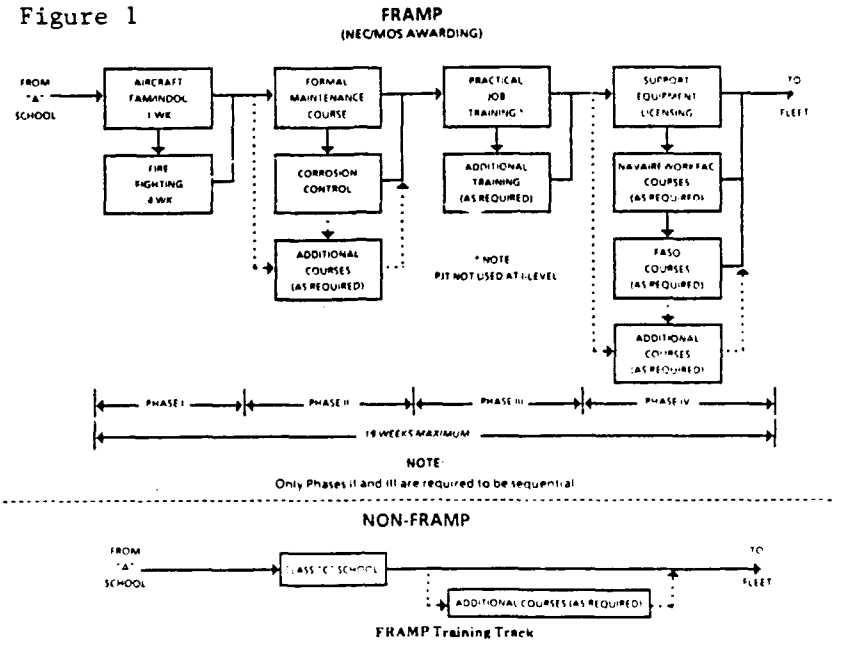
The purpose of this paper is to discuss MATMEP and show how the aspired ITSS goals are met through utilization of existent Navy and Marine Corps programs.

Naval Aviation Technical Training

Formal training in the aviation technical fields falls under the Chief of Naval Education and Training. The training pipeline leading to military occupational specialty (MOS) qualification is shown in Figure 1. The "A" School provides the trainee with basic skills and the fundamental knowledge required to assimilate follow- on aircraft specific training. The "C" School training occurs in Naval Aviation Maintenance Training Group Detachments (NAMTGDs) and addresses training on specific equipment or aircraft systems. Class "C" training provides sufficient background for the trainee to perform required maintenance tasks under close supervision. In an ideal situation, the practical-job-training (PJT) occurs in a Fleet Readiness Squadron (FRS). The Fleet Readiness Aviation Maintenance Personnel (FRAMP) program provides closely supervised instruction from the time of graduation from "A" School to the attainment of an MOS. In the absence of a FRAMP, the PJT occurs in the squadron to which the trainee is assigned. The Marine Corps presently has FRAMP training available on less than half of its aircraft. When the trainee reaches the fleet, Navy

training strategy still dictates, to a degree, internal Marine Corps training since maintenance practices, procedures, and training fall under the requirements of the Naval Aviation Maintenance Program (NAMP).

Figure 1



"developed by Marines for Marines"

The key players for developing ITSS/MATMEP were all Marines, as can be seen from the following infrastructure. A MATMEP Development Team (MDT), made up of reservists who developed the 4TH MAW MATMEP, was activated for a period of one year to provide the nucleus of support for program development. The AV-8B and the CH-53E FRAMPs were designated as MATMEP system managers (MSMs) for the fixed and rotary wing communities. The MSMs were responsible for development of MATMEP models for their respective communities. These models included: prototype task inventories, training standards, and management records for each occupational field. Maintenance training model manager (MTMM) squadrons were assigned for each aircraft model in the Marine Corps. These MTMMs were responsible for the standardization of aviation maintenance training in an aircraft community. Job incumbents from each of the MTMMs provided the subject matter experts (SMEs) required for the development of MATMEP for their community.

"find out what Marines need to know"

A Front End Analysis (FEA) was performed to identify the task, skill, and knowledge requirements of each job or(MOS). Complex analyses, however, were not required for MATMEP for an extensive training system is in place for Naval aviation technical occupational fields, and, if we accept the educational integrity of the aviation maintenance system, a plethora of source material is

available to facilitate the analysis. Source materials used in the analyses included: maintenance instructions, course syllabi, logistics support analyses, Maintenance Material Management (3M) Reports and data from the Naval Aviation Logistics Data Analysis (NALDA) System.

Phase I of the MATMEP analysis process included the conduct and authentication of a task inventory (TI). The MATMEP development team compiled work packages containing the task, skill, and knowledge source material listed above. Duties, large generic segments of work, were extracted from the Marine Corps Military Occupational Specialties Manual for each job or MOS to be devlop-ed. The duties were organized into four nominal groups: general, operational and safety duties, scheduled and unscheduled maintenance duties, maintenance administration duties, and productive indirect work center duties, such as quality assurance and collateral duty inspection, workcenter supervision, instruction, etc.

The development process was initiated by a procedural brief to the MTMM by the MDT. A top down approach was used to develop the TI. The SMEs in the MTMMs took the nominally grouped duties and developed task statements within the framework of the duty outline. The source material packages provided a broad and sound basis for TI development; however, the primary focus was on technical publications and maintenance instruction manuals (MIMs). The prototype MOS developed by the MSM was provided as a guide to the degree of specificity. Once the tentative TI was complete, the results were submitted to the MDT for review. The review covered degree of specificity, technical expression, completeness and elimination of duplication.

After review by the MDT, the TI was formatted and produced as a four level Individual Qualification Record (IQR). A portion of a sample IQR is shown in Figure 2. The four skill progression levels are each defined in terms of task performance, task knowledge, and subject knowledge. Levels I and II equate to an apprentice, Level III to a journeyman, and Level IV to a master technician. Level I training takes place in the NAMTGDs, Level II is covered in the FRAMP and/or in squadron on-the-job training (OJT), Level III is squadron OJT and Level IV takes place in squadron OJT and advanced formal instruction.

Figure 2

DA 8.6

8.6 Performs organizational level maintenance on applicable systems including the removal, installation, adjustment and alignment of the system and the repair or replacement of associated components.

TASK #	TASK DESCRIPTION	REFERENCE	WUC	LEVEL I	LEVEL II	LEVEL III	LEVEL IV
A	VHF/UHF COMM system	AI-AV8B-650-300	62IK0	*11	/ /	/ /	
A-1	R&R ACHIP-AB-7109/A	"	67K2600		• / /	• / /	
A-2	R&R fin cap antenna	"	62X2Z		/ /	/ /	
A-3	R&R KY-58 secure voice #1 or #2	"	67410	*11	/ /	• / /	
A-4	R&R receiver-transmitter, R/T-1250 ARC-182(V) #1 or #2	"	62X2100		• / /	• / /	
A-5	R&R ventral fin antenna	"	62X2F		/ /	• / /	
A-6	R&R U/VHF radio set control, C-10776-1-ARC	AI-AV8B-650-300	62X2600	*11	/ /	• / /	
A-7	R&R switching unit	"			/ /	/ /	

cedures of maintenance instructions, provide both the training objective and the standards for performance evaluation. The task, skill, and knowledge requirements contained in the MATMEP IQR provide, in effect, a syllabus for inservice training. The IQR indicates required training, documents OJT, and reports attainment of satisfactory levels of skill for both formal and informal OJT.

Due to the intensity of typical Marine and Navy flight operations, informal OJT is difficult to schedule, coordinate, or document and often is handled on a "catch as catch can" basis. The key to success of any OJT program is a motivated trainee. MATMEP is essentially a management-by-objectives approach to training, and as such, can provide the required motivation. The IQR gives the trainee a complete list of training objectives for every level of training and in effect integrates the work unit supervisor's training goals with the individual Marine's needs to contribute and develop personally.

While MATMEP is aimed primarily at unit training, it also provides data for more effective appraisal of institutional training. Maintenance Training Requirement Review (MTRR) Conferences are held on a regular basis to review the training pipeline for a particular weapons system. MTRR participants will find the IQRs invaluable, for the task inventory contained in the MATMEP IQR provides a single occupational data base that spans the aviation training pipeline. When a Marine arrives at his assigned squadron, the work center supervisor can review his training record and see precisely, task by task, the training received. With this information, the work center supervisor can more objectively appraise previous training, as well as more effectively schedule and manage individual and work center training.

"test Marines to see if they can perform their jobs"

MATMEP will include diagnostic testing similar to the Navy's Maintenance Training Improvement Program (MTIP). This combination will provide the Marine Corps with a tool to evaluate both performance and knowledge. One of the goals of MTIP is to determine knowledge deficiencies by diagnostic testing and then offer refresher training to eliminate those deficiencies. MATMEP will incorporate MTIP but will develop the test items around the prioritized level progressive tasks in the IQR. MATMEP testing will include both open and closed book items. Development of test items and standard refresher lesson guides will be a task for the MTMM conferences. The deficiencies identified by this testing will drive to a large degree the formal side of unit training. A second goal of MTIP, institutional training appraisal, can be accomplished by a standard Level II test given within ninety days of assignment to the fleet.

While MATMEP provides units with an internal means for training appraisal, the program is not complete without external evaluation. There are many forms of external evaluations - the IG to name but one. These inspections include training appraisal. But,

Phase II is the training analysis phase. In this phase the appropriate NAMTGDs and FRSs reviewed the task list on a task by task basis to determine which tasks were included in their respective curricula. The IQRs were then updated by the MDT to incorporate this input. The tasks included in NAMTGD or FRAMP courses are indicated at the appropriate levels in the IQR with an asterisk. If either of the institutions recommended task additions or deletions the recommended changes were submitted to the MTMMs for their concurrence prior to incorporation into the IQR.

Phase III involved the prioritization and selection of tasks for inclusion in squadron OJT. The IQR is extensive and contains all of the actions authorized by the maintenance instructions. Many of the tasks occur infrequently. The SMEs in the MTMMs reviewed the IQRs and identified tasks that occurred infrequently and could be omitted from the training continuum without detriment to unit safety or mission performance. These tasks, though still included in the IQR, are considered training optional. The remaining tasks are considered training essential and are noted with an asterisk in Level III of the IQR. In cases where no FRAMP exists, high priority tasks are selected for a training syllabus for the squadron OJT required for MOS qualification. These MOS essential tasks are noted in Level II of the IQR with an asterisk. Criteria considered in the selection of task for training include: time spent performng, frequency of performance, and probable consequences of inadequate performance. The Maintenance Material Management (3M) and the Naval Aviation Logistics Data Analysis (NALDA) systems were invaluable data sources for this task selection effort.

Phase IV of the process is the implementation of MATMEP and the validation of the TI/IQR. The MDT has incorporated the data from Phase III and produced updated IQRs. MATMEP is now ready for implementaion as the training management and evaluation program for organizational level maintenance activities. Approximately one year after implementation, the MTMM for an aircraft model will host a review conference. Marines representing each MOS in each squadron in that community will attend the conference. The task list and training priorities will be reviewed and the IQR will be updated. Subsequently, th` IQR will receive update through a perpetual review process. Once MTMM review conferences have been held for all aircraft, a system review will be held by the MSMs to evaluate and update the overall MATMEP process.

"teach Marines what they don't know "

The main purpose of MATMEP is the management and evaluation of OJT or inservice training. OJT is a major element of the technical training pipeline with direction and guidance for this training found in the NAMP (OPNAVINST 4790.2D). As the name implies, OJT occurs in the operational unit and includes formal (lecture) and informal (practical) training, as well as required reading. The classroom instructor for OJT is also the work unit supervisor of the trainee and evaluation of the trainee is primarily based on the capability to demonstrate specific training objectives. The tasks contained in the IQR, when coupled to the step by step pro-

evaluations of this type are primarily aimed at training documentation with the assumption that good training documentation equates to good training.

As a part of a unit's Marine Corps IG inspection, a randomly selected group of Marines is required to take the Essential Subjects Test. Success on the test measures to a degree the effectiveness of a unit's general military training. Unit technical training should be evaluated in a similar fashion using MTIP.

The Marine Corps has an excellent performance based evaluation instrument in its Marine Corps Combat Readiness Evaluation System (MCCRES). The MCCRES was designed to evaluate aircrew performance under conditions as close to combat as possible. Aircraft availability and systems status are a reflection of maintenance training. MCCRESs, however, are surge operations that can be handled with a few well trained Marines while sustained high intensity operations require a broader base of trained personnel. The MCCRES should be expanded to more directly address maintenance training.

Automation

Automated MATMEP qualification records will reduce the training documentation workload of supervisors. Individual, workcenter, and squadron qualification summaries will provide invaluable management data. MATMEP is similar to the Training and Readiness (T&R) Syllabus, a system used to manage flight crew training. MATMEP, like the T&R syllabus, lends itself to the quantification of performance capability on an individual and unit level through the use of qualification percentages. It is, however, the opinion of this writer that upline reporting of unit qualification in the early stages of MATMEP would be detrimental to the program since such reports place the emphasis on unit report cards and not unit training. Unit evaluation is best addressed through IGs and MCCRESs. On the other hand, personnel data on an individual basis, giving level of qualification, system by system, would provide a powerful tool to more effectively handle manpower assignments.

In training development we too often get tied up in process and forget product; program development becomes more important than training. In military training, documentation often overshadows training itself. Hopefully, with MATMEP, we will avoid these pitfalls. MATMEP is grassroots developed, incorporates a variety of evaluatory methods and is based on a dynamic occupational data base that spans the aviation maintenance training spectrum. There is little new in MATMEP; it is a blend of ITSS with existing maintenance training programs and doctrine. With the tightening of budgetary belts, it is time to train better and work smarter.

THE EMERGENCE O. COLLECTIVE TRAINING STANDARDS (CTS) IN THE MARINE CORPS

1. What are CTS? The purpose of the present paper is to discuss a proposed approach for developing and implementing Collective Training Standards (CTS) for the Marine Corps. In this paper, we will address the issues of what CTS are, why they are needed, why their development needs to be integrated with that of other types of performance standards, and how this can be done. The term CTS is used by the Marine Corps to designate performance standards established for collective or team tasks. In the research literature the definition of a team task has varied greatly from one researcher to another. In the present context a team or collective task will be defined broadly as a task where two or more personnel with assigned specific roles are working toward a common goal. Note that this definition does not require that the teams be formally structured (such as platoons or squads). Therefore, ad hoc teams which only form for a specific task would be included. This definition also does not specify the nature of the interactions that occur among personnel in performing the task.

2. Why are CTS needed? CTS, along with Individual Training Standards (ITS) establish performance criteria for determining the training readiness of a unit. Such standards are needed to provide guidance for unit commanders in training and evaluating their personnel. Taken together, ITS and CTS should provide performance standards for all of the tasks that must be performed to accomplish missions. In the Marine Corps, evaluations of mission performance are conducted with the Marine Corps Combat Readiness Evaluation System (MCCRES). It has been questioned whether CTS are really required since MCCRES Mission Performance Standards (MPS) already cover collective performance. However, MPS only provide a very general level of performance evaluation. A single MPS may cover many levels of individual and collective performance. They are not diagnostic. That is, if MPS are not met, it may not be evident which personnel performed which tasks incorrectly. Also, they are not prescriptive. That is, they do not indicate what type of training is required to satisfy their standards. CTS are needed to provide more specific evaluations of collective performance and to provide a link between collective performance and training.

3. How should CTS be developed? At present, the Marine Corps is developing ITS independently of CTS and MPS. A recent report by Lewellyn (1984) supported the concept of separate development and proposed an approach for developing CTS independently of ITS. This involved reviewing MCCRES MPS to identify instances of collective performance and establishing collective performance standards at company, platoon and squad levels.

There are a number of limitations associated with Lewellyn's approach as he describes it. First of all, it is not comprehensive. He indicates that it is designed only to identify

team tasks performed by established units (such as platoons or squads) not to cover ad hoc tasks. Second, his approach is not sufficiently analytic to separate all team tasks from all individual tasks. At the level he conducted his analysis, some identified tasks contain elements of both individual and team performance. He also did not specify which type of personnel perform which type of tasks. Without such information collective training responsibilities will be difficult to determine. Third, his approach does not allow him to relate individual and collective training requirements. This can be an important consideration since research findings have established that individual training prerequisites should precede team training (Dyer, 1985). Fourth, his approach does not distinguish between collective tasks that require team training and those that do not. This also can be an important consideration since team training may be more expensive than individual training and does not appear to benefit all types of collective tasks (Wagner et. al., 1977).

4. Identifying collective performance. Because of the limitations associated with Lewellyn's approach, an alternative approach for identifying CTS requirements has been proposed. It consists of an analysis procedure for identifying collective performance responsibilities and a selection procedure for determining which of the identified instances of collective performance actually require CTS. The proposed analysis procedure is similar to the approach employed by Lewellyn in that collective tasks are identified from MCCRES MPS. It differs in the level of analysis provided, and in the manner that task responsibilities are assigned.

In MCCRES, MPS are separated into stages (designated as tasks). The proposed analysis procedure deals with each stage as a separate block of performance and analyzes mission involvement in terms of the units involved (companies, platoons, etc.) and the functions to be accomplished. In the first stage of this analysis procedure data is provided to indicate what groups are involved in the mission and what functions they perform. It has been found convenient to display such mission involvement data in a matrix format (Wheaton and Johnson, 1980). An example of such a matrix is provided in Figure 1 for the Prepare For Attack Stage of the Attack Mission. The headings at the top of Figure 1 indicate what functions are to be accomplished in this Stage (ordering troops to prepare; distributing ammunition; checking weapons, communications equipment, vehicles and supplies; and increasing security). The headings at the side indicate, in a hierarchical manner, the different units that are involved in performing each of the listed mission functions. Each x indicates where a group is involved with a specific mission function. For example, it can be seen that the Battalion Commander and Staff group of Headquarters and Service (H & S) Company is involved in giving the warning to prepare for attack. This matrix would be used as an aid in identifying instances of collective performance. Each of the functions would be reviewed by subject matter experts to determine whether collective

Mission: Attack Stage: Prepare for Attack

Groups	Functions						
	Give Warn	Dist Ammo	Chk Weap	Chk Comun	Chk Vehic	Chk Supl	Incr Secur
H & S Co	x	x			x	x	
Btl Cdr/Staff	x						
Serv Plat		x			x	x	
Rifle Co	x	x	x	x	x	x	x
Rifle Plat	x	x	x	x			x
Weap Plat	x	x	x	x	x	x	x
Asslt Sect	x	x	x	x	x	x	x
60mm Mort Sect	x	x	x	x			x
Mach Gun Sect	x	x	x	x			x
Weap Co	x	x	x	x	x	x	x
81mm Mort Plat	x	x	x	x			x
DRAG Plat	x	x	x	x			x
Asslt Sqd	x	x	x	x	x	x	x

Figure 1. Mission involvement chart.

performance is involved and what personnel, within or between groups, interact in performing the functions.

The next level of the proposed analysis might be referred to as a function analysis since it separates mission functions into their component tasks for each group. An example of such a function analyses for a rifle platoon is displayed in Figure 2.

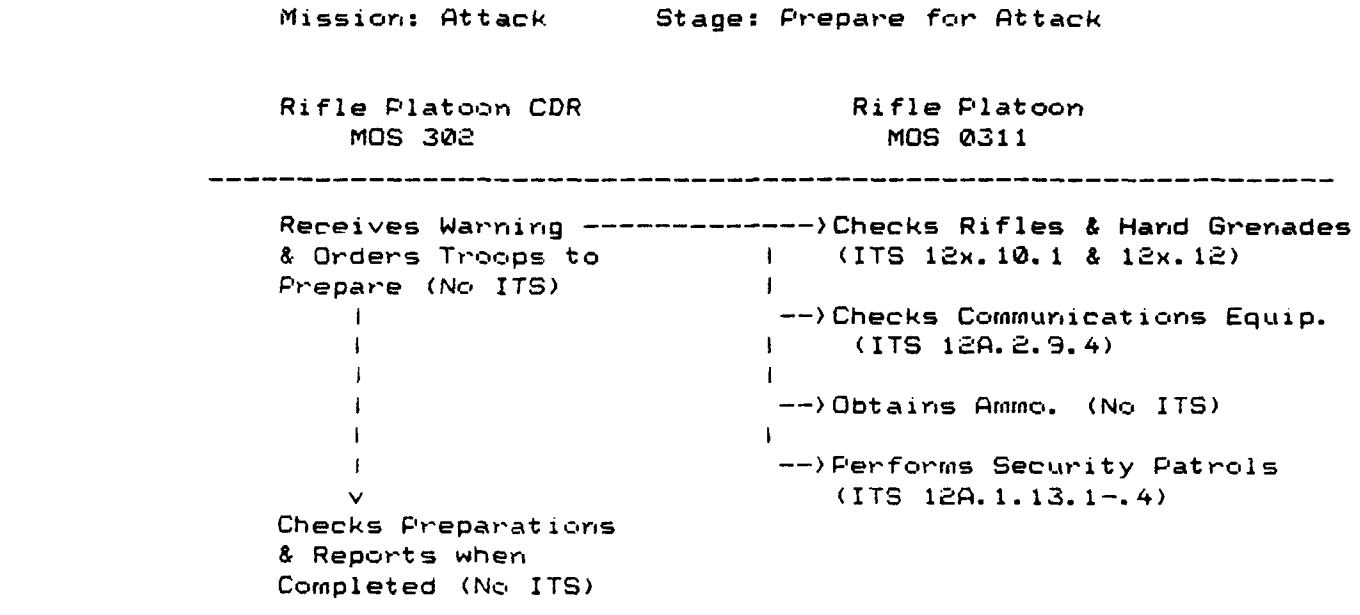


Figure 2. Infantry rifle company rifle platoon function analysis.

It now becomes possible to specify the Military Occupational Specialties (MOSs) involved in the mission and their respective tasks. Rifle platoon members, for example, belong to MOS 0311. It is also possible to determine, for each group, what specific tasks they perform for each of the functions they are involved with. Most tasks are somewhat unique to the equipment on which they are performed. For example, in checking weapons, different groups perform different tasks because they use different weapons. In contrast, checking communications would be the same basic task for all groups if they use the same communications equipment.

In the function analyses, the task descriptions, along with the specification of the involved MOSs, allows for the identification of existing related ITS. In Figure 2., ITS 12x.10.1 and 12x.12 correspond to the tasks of checking rifles and hand grenades. One advantage of this type of analysis is that, in addition to identifying collective performance requirements, it identifies individual performance requirements not currently covered by ITS. For example, for the rifle platoon none of the existing ITS applied to obtaining ammunition. Another advantage of this type of analysis is that it can be used

to indicate which ITS are related to instances of collective performance. This can be helpful in specifying ITS as prerequisites for CTS.

5. Determining CTS requirements. The mission involvement data along with the function analysis provide an overview of collective tasks associated with each group operating within the battalion. The next question, then, is which of these collective tasks require CTS. As was indicated earlier, not all collective tasks require CTS. CTS are only required for collective tasks where group members are to be trained collectively. Therefore, in order to determine which tasks require CTS, it is necessary to determine whether or not the tasks should be trained in a team context.

There are a number of factors that should be considered in determining whether or not a collective task should be team trained. Does the task have to be performed by a team to be practiced? If learning the task requires practice and practice can only be conducted in a team context, then, obviously, the task must be trained collectively. Is coordination required? Wagner et. al. (1977) cited studies that indicated team training might be expected to be more effective than individual training for emergent (non-procedural) tasks where team members must coordinate their actions. For example, in a tank team the driver and the gunner might each need to be aware of what the other is doing in order to coordinate their actions. Such coordination requires an awareness of the position, actions or goals of other team members. Does the task involve communication? Kahan et. al. (1984) in summarizing a Rand Corporation workshop on improving team performance indicated that all participants saw team communication as a critical element in team performance. Turney et. al. (1981) cited several studies which demonstrated that effective teams communicate differently than less effective ones. Team training may be required to promote the communication patterns used by the more effective teams.

Note that it may be desirable to train personnel as a group even if they are being trained to individual standards. For example in some situations team training may be used to promote competition or to build confidence in the capabilities of other team members. In other situations collective training may be provided simply because it allows a large number of personnel to practice the task at the same time. In such situations, where collective training is being used only to enhance individual performance, CTS are not needed.

In addition to determining whether or not CTS are required, it is also necessary to determine the level at which the training is to be provided. The level or levels selected for collective training should depend on the reasons that collective training was selected. For example, if it is being used to promote communication, it should involve all of the personnel involved in the communication process. Therefore, team training should not be confined to established groups such as platoons or squads but

should allow for interactions within or between such groups.

Based on the above considerations, algorithms are being developed to aid subject matter experts in determining what type of CTS are required for specified instances of collective performance. These, along with the previously described analysis procedures, should provide the Marine Corps with an approach for identifying CTS requirements that will allow team training resources to be focused on those areas of performance where they are most needed.

REFERENCES

Dyer, J.L. (1984). Team research and team training: A state-of-the-art review. In F. A. Muckler(Ed.) Human factors review(pp.285-323). Santa Monica:Human Factors Society.

Kahan, J.P., Webb, N., Shavelson, R. J. and Stolzenberg, R. M. (February 1985). Individual characteristics and unit performance (R-3194-MIL). Santa Monica:The Rand Corporation.

Lewellen, M. T. (December 1984). Collective training standards for the Marine Corps (CNT Resch. Memo. CRM 84-40). Alexandria:Center for Naval Analysis.

Turney, J. R., Cohen, S. L. and Greenburg, L. (April 1981). Targets for team skills training (GP-R-43017). Columbia:General Physics Co.

Wagner, H., Hibbits, N., Rosenblatt, R., and Schulz, R. (February 1977). Team training and evaluation strategies: State-of-the-art (HumRRO Tech. Rep. TR 77-1). Alexandria:Human Resources Research Organization.

Wheaton, G. R. and Johnson, E., III. (July 1981). Research on systems analytic approaches to unit evaluation: Specification of performance variables and measures for ES-based ARTEP. Alexandria:U.S. Army Institute for the Behavioral and Social Sciences.

Note: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.

LOGISTICS SUPPORT ANALYSIS (LSA) VALIDATION THROUGH TRAINING EVALUATION

Marty Dilg, Education Specialist
United States Marine Corps, Washington, D. C.

RATIONALE

The Marine Corps, as well as the other services, is acquiring increasingly sophisticated weapon systems. As a result of this sophistication, various problems, trade-offs and alternatives associated with personnel, management and training must be realized, surfaced and evaluated so that a supportable maintenance concept can be developed for each weapon system. Initial Logistics Support Analysis Report (LSAR) data provides one of the primary means through which Marine Aviation evaluates the trade-offs and alternatives for new and modified weapon systems. The use of these LSAR data enables the Marine Corps to analyze task performance: personnel, skill level, skill specialty, work area, and man-minutes requirements.

Based on these elements, LSAR also identifies problems that surfaced in the force structure. The problem areas generated by a policy stance in the areas of acquisition, manpower and training and some ongoing resolutions are addressed in this paper.

INTRODUCTION

The Marine Corps often utilizes other services as acquisition agents to procure and field its weapon systems. Since the Marine Corps is organized under the Department of Navy, the Navy has been chartered with the responsibility for fulfilling Marine Corps aviation requirements, so that the Department of Defense can realize greater procurement efficiencies. Marine Corps aviation weapon system procurement is conducted to a large extent by the Naval Air Systems Command.

In the past, Navy/Marine Aviation acquisitions paid only passing attention to manpower and training requirements as part of the Intergrated Logistics Support Plan. The Navy Manpower/Training models were developed in isolation from the Marine Corps manpower ceiling, thereby straining the established manpower philosophy.

Manpower ceilings are congressionally imposed, thus placing the Marine Corps in a constrained environment. Any real growth in terms of personnel is limited. In spite of such constraints, all requirements must be met. Military Occupational Specialties (MOS's) to meet the maintenance concept are defined in terms of aircraft and systems. Assignments to these requirements are based on a Marine's primary MOS. Manpower must also meet Marine Corps unique ("B" billet) assignments, such as recruiter or embassy duty. To fulfill these types of requirements, a Marine serves a tour of duty outside the parameters of a primary MOS. Another influencing factor in the Marine Corps is the promotion system which is promulgated by Manpower. This system has recently changed from promotion by broad Occupational Fields (OccFld) to promotion by rather limited MOS's.

Formal school training for Marine Aviation is accomplished jointly,

primarily by other services, with most of the formal school training being conducted at the entry level. A structured recertification, requalification or advanced formal school training program is negligable for most MOS's. With the current training and manpower philosophies and in conjunction with the sophistication of the weapon systems, it is frequently difficult to recruit, train, and retain Marines with skills that meet the requirements of organizational, intermediate, and depot level repair and the maintenance concept of the Marine Corps. From these considerations the following problem areas surfaced.

MANPOWER AND TRAINING PROBLEMS

Due to the manpower ceilings imposed on the Marine Corps by Congress, there are many MOS's that are excessively small, that is, populations that contain fifty Marines or less. Another constraining factor from a manpower perspective is geographic location. Some MOS's, because they are predicated on aircraft and systems, can be assigned to a limited number of places, in several instances, only one or two geographic locations. These two factors alone tend to drive manpower personnel toward micromanagement of personnel assets. Manpower is also required to fill "B" billets, most of which are manned at one hundred percent. Each MOS must contribute its "fair share" to fill these Marine Corps unique billets. In doing so, a Marine is removed from the arena that requires him to deal with an extremely complex and sophisticated weapon system. After three years, the normal duration of a "B" billet assignment, technical proficiency and currency are degraded. Additionally, in an effort to recruit and retain personnel in hard to fill MOS's, manpower promulgated a system of guaranteed promotions to sergeant and/or a guarantee of a geographic preference. These last two factors create morale problems in the Fleet Marine Force (FMF). These problems, coupled with other problems that surfaced within training, have been flagged as potentially hazardous.

The training philosophy for the Marine Corps exacerbates the manpower problems. Some MOS's have extremely long entry level training tracks, taking up two-thirds or better of a Marine's initial entry committment. This is due in part to the sophistication of the weapon systems and a hardware oriented, systems approach in teaching Marines their trade. Since the MOS's are identified by aircraft or system, training pipelines are generated to teach to that system or piece of hardware. This frequently results in redundancy between/among the training tracks.

Most Marines with Aviation MOS's are taught by other DOD agencies. The Marine Corps in the past has had limited input into how Marines are trained in these formal schools; consequently, changes in weapon platforms or the identification of skill and knowledge deficiencies must be handled by alternate methods, often creating an East Coast/West Coast Marine Corps and non-standardization of training. The change in a weapon system sometimes skews the maintenance philosophy. Presently, there are intermediate level technicians functioning at an organizational level and vice versa.

STUDY METHOD

Rather than having the Aviation, Manpower, and Training Departments deal

with these issues and problems separately, all three departments decided to approach them collectively. Each department contributed personnel and other assets to the project. The first step agreed upon by the various departments was to quantify existing data. A math model for grade shaping purposes was selected. This math model consisted of percentages of a total population with the percentages broken out by paygrade from E-1 through E-9. Each group of feeder MOS's each OccFld and the total aviation force structure was computed and compared against that math model. The MOS populations studied included all existing, authorized billets within the Marine Corps. The math model enabled us to grade-shape populations in terms of an ideal structure. These were then compared to actual populations.

Tasks for the Aviation MOS's were extracted from various data bases: Logistic Support Analysis Reports (LSAR), Naval Aviation Logistics Data Analysis (NALDA), and Comprehensive Occupational Data Analysis Program-80 (CODAP). These tasks were compared for similarity.

The senior staff noncommissioned officers in the FMF, within the United States were surveyed and interviewed. The objective of the surveys and interviews was to determine if MOS's should be combined based on the tasks performed, skills and knowledges required, and any training problems.

Training tracks for each MOS were compared for length, types of tasks taught, and placement within the training pipeline.

ANALYSIS AND PRESENTATION OF DATA

Based on the math model there were several areas that were identified:

a. MOS's with no promotion potential past a certain rank

b. MOS's with populations of less than ten. See figure 1, "MOS Breakdown (63XX)".

	REQUIRED			
	6363/64/65/67			
E1-E3	9	5	14	
E4	7	5	13	18
E5	5	5	11	32
E6	3	26	10	11
E7	3	18	3	8
E8				
E9				
	27	59	51	69

Figure 1, MOS Breakdown (63XX)

c. MOS's with critical departures from the math model, particularly at the senior noncommissioned officer level. See figure 2, "Grade Shape".

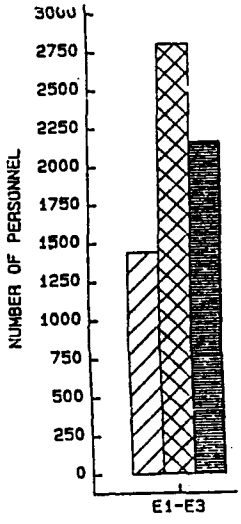


Figure 2, Grade shape

d. Inappropriate grade shaping producing little or no career progression. See figure 3, "MOS Breakdown (64XX)".

	REQUIRED 64 32/33/34/35	ACTUAL 64 32/33/34/35
E1-E3	58 21 37 32	69 90 43 44
E4	67 49 26 40	70 40 27 19
E5	66 65 23 8	66 48 36 22
E6	27 20 14 14	42 30 15 20
E7	18 11 4	9 15 10 3
E8		

Figure 3, MOS Breakdown (64XX)

Tentative consolidations of MOS were formulated from the analysis of tasks, skills, and knowledges available through the LSAR data, training pipelines, interviews and survey data.

Graphic displays of information with narratives recommending changes for each group MOS's were drafted and routed among the Aviation, Manpower, and Training Departments. See figure 4, "Force Structure".

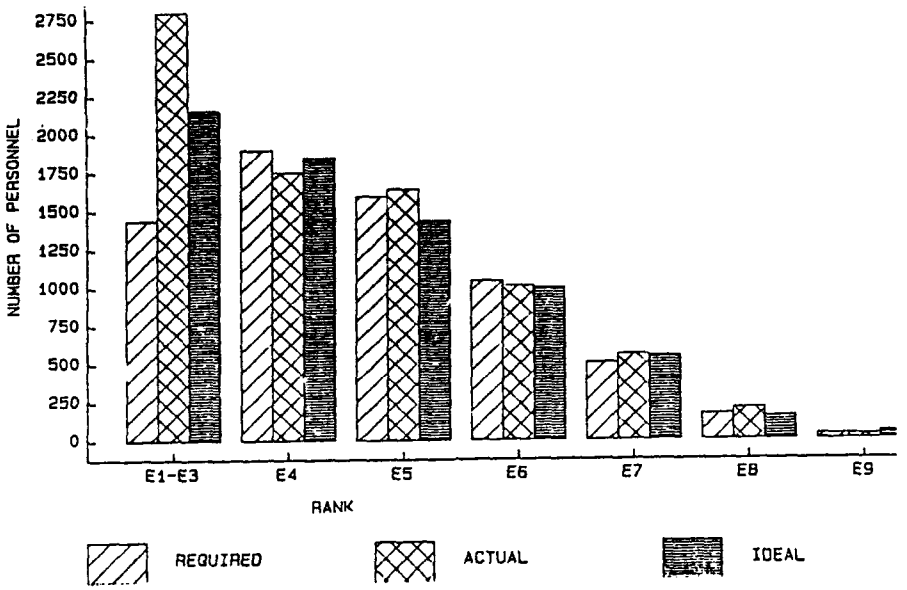


Figure 4, Force Structure

PROGRESS TO DATE

As a result of the information that was routed some manpower and training problems are being solved. In light of the fact that some MOS have no promotion potential or career progression past a certain point, the Commandant issued a statement that in effect guaranteed a qualified Marine the opportunity to progress in his primary MOS from E-1 to E-7. That statement in turn has placed requirements on the three departments involved, particularly within Aviation, to provide that opportunity through appropriate grade-shaping and billet-structure. One of the ways that this is being done is by consolidating MOS's that are alike, such as structures mechanic and hydraulicsman. See figure 5 labeled "MOS Consolidation." Another approach is to break up the training tracks into a tri-level approach so that there is an apprentice, a journeyman, and a master technician within each technical MOS. Entry level training is held to a minimum, thereby reducing cost to train "first term enlistees". Manpower can use highly specialized Marines in "B" billet assignments without a degradation in performance or unit readiness. Training becomes more cost effective because training pipelines are adjusted so that a Marine obtains training throughout a career.

	OLD MOS/S/	E1-E5 PROPOSED MOS	E6-E7	E8-E9
W/C 610	6412/6413 6414/6415/6416	6412 6413	6414	
W/C 620	6432/6433 6434/6435 6423 (ADDITIONAL MOS)	6432 6433 6423	6434	
W/C 630	6472 6474 6475 6476	6472 (AV8 EETS REQR) 6473 6474 6475	6476	
W/C 640	6452/6453/6454/6455 6482	6452 6482 (F/W) 6483 (R/W)	6484	6391
W/C 650	6442/6443/6445 6444/6446	6442 6443	6444	
	6462 6463 6464 6465 6477	6462 6463 6464 6465 6466	6468	
W/C 670	6492/6493 5938 5982	6492 6493 6494	6495 6496 6497	

Figure 5, MOS consolidation

Effects of Soldier Performance and Characteristics on Relationships with Superiors[1]

Ilene F. Gast and Leonard A. White
U.S. Army Research Institute for the Behavioral and Social Sciences

With the increasing emphasis on interactive leadership approaches (Jacobs, 1971; Graen 1976) has come a recognition of the contributions subordinates make to the leadership process. Although leaders may tend to have a characteristic style, they vary their behavior substantially in response to subordinate actions and needs. Graen has shown that leaders form different kinds of working relationships with their subordinates. Relationships range from "in-group" ones characterized by mutual support and trust to "out-group" ones where both parties do only what is required by the formal employment contract.

Graen (1976) notes that relationships formed early in one's career have lasting effects. Based on a longitudinal investigation of management trainees, Wakaybayashi and Graen (1984) conclude that a newcomer's relationship with his or her superior serves motivating and mentoring functions that help the newcomer assimilate into the organization and gain access to information and resources central to the functioning of the work unit. This experience gives newcomers the confidence they need to set higher performance goals. Thus, the relationships that first tour soldiers form with their superiors are not only important from the standpoint of socialization into the Army, but may also affect career progression and leadership potential.

Past research has shown that subordinates' performance is a powerful determinant of subsequent treatment by superiors (e.g., Greene, 1975). Generally, poor performers are more likely to have low quality relationships with their superiors. However, because this phenomenon has been investigated primarily in the laboratory, field research is needed.

There also is evidence that relatively stable personal dispositions enable some subordinates to form more positive relationships with superiors. Graen and his associates (Graen, Novak, & Sommerkamp, 1982) demonstrated the importance of subordinates' growth need strength to the formation of effective relationships with superiors. However, with the exception of Graen's work and research by Hough, Gast, White and McCloy (1986), researchers have not adequately addressed the potential effects of individual differences on subordinates' interactions with their superiors. Such research is needed.

Using data from Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel (Eaton, Goer, Harris & Zook, 1984), this paper examines how working relationships between superiors and subordinates are directly affected by subordinates' job performance, temperament and ability. In addition, this paper explores non-linear effects of soldier ability and temperament on working relationships with superiors.

Method

Subjects

Subjects were 5,123 first term soldiers in 9 military occupational specialties (MOS): 683 infantrymen (11B), 636 cannon crew members (13B), 489

---

[1]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.

tank crew members (19E), 349 radio teletype operators (31C), 618 light wheel vehicle mechanics (63B), 670 motor transport operators (64C), 502 administrative specialists (71L), 487 medical specialists (91A) and 689 military police (95B). Within the sample, 88% of the soldiers were male and 12% female. Of those who reported their racial origin, 23% were black, 3% were Hispanic, 70% were white, and 4% replied "other". On the average, soldiers had been in the Army for 18 months and with their present companies for about a year. To facilitate data analysis, jobs were grouped into four occupational clusters identified by McLaughlin, Rossmeissl, Wise, Brandt and Wang (1982). The Combat cluster included MOS 11B, 13B and 19E; MOS 31C, 63B and 64C comprised the Operations cluster; MOS 71L made up the Clerical cluster and the remaining MOS, 91A and 95B comprised the Skilled Technical cluster.

Instruments

Supervisor Behavior Questionnaire. The authors wrote items to tap categories of supervisory activities identified through analysis of 400 behavioral examples of effective and ineffective leadership. These items required subjects rate statements about their supervisor using a 5-point scale from Very Seldom or Never (1) to Very Often or Always (5). The resulting questionnaire was field tested in a sample of 696 first term enlisted (White, Gast, & Rumsey, 1985) and revised prior to administration in the present sample. Principal factor analysis with promax rotation revealed five factors with eigenvalues greater than one: Inspiration/Support, Participation, Structuring Work, Fairness/Discipline, Work Allocation. The present research employed only the scales corresponding to the first two factors; these scales were most similar to scales measuring qualities of "in group" relationships in previous research (Vecchio & Gobdel, 1984; Novak, 1985). Typical items on the 9-item Inspiration/Support scale included "Your supervisor understands your problems and needs" and "Your supervisor wants to make you give your best effort". The 4-item Participation scale contained items like "You are permitted to use your own judgment in solving problems". Reliabilities (Chronbach's alpha) for these two scales were .82 and .70 respectively.

General cognitive ability. The Armed Services Vocational Aptitude Battery (ASVAB) was administered to all subjects prior to entering military service. A composite of four ASVAB subtests, known as the Armed Forces Qualification Test (AFQT), served as the measure of general cognitive ability.

Temperament. Hough, Kamp and Barge (1984) developed ten scales to assess temperament constructs shown to be related to criteria of work performance in previous studies. The resulting inventory, Assessment of Background and Life Experiences (ABLE), was tested on 470 soldiers at three forts. These data guided revisions to the items and scales. When subjected to principal factor analysis with varimax rotation, the revised scales yielded three factors with eigenvalues greater than one: Dependability, Achievement Orientation, and Emotional Stability. Scales measuring self-esteem, dominance, energy level, and work orientation comprise the Achievement Orientation factor. The Emotional Stability factor assesses the degree of stability vs. reactivity of emotions. The Dependability factor includes measures of conscientiousness, non-delinquency, support for rules and regulations, and respect for traditional alues. Factor scores for the three scales were used in the analyses.

Job knowledge tests. Through job analysis important knowledge areas were identified for each MOS. Project A personnel, assisted by subject matter specialists, developed items to tap these knowledges. The overall job knowledge test score was the percentage of items answered correctly by each soldier.

Hands-on task proficiency tests. Critical tasks were identified to represent the task domain for each MOS. A multiple step proficiency test was developed for each task, and each step was scored pass or fail. For each task, the score was the proportion of steps passed; then these task scores were averaged to yield an overall hands-on test score (Campbell, Campbell, Rumsey & Edwards, 1985).

Army-wide performance rating scales. Eleven 7-point behaviorally anchored rating scales were developed to assess soldier effectiveness across army jobs. These scales went beyond task performance to include aspects of socialization and commitment to the organization. Ten scales covered specific aspects of soldier effectiveness; the eleventh scale required an assessment of overall effectiveness. Supervisors' ratings on this eleventh scale were employed in the present analyses (Eaton et al., 1984).

Procedure

After receiving training in the use of the behavior anchored rating scales, supervisors, in groups of 3-15, evaluated their subordinates. The mean number of supervisors providing the ratings for each ratee ranged from 1.66 to 1.83. Ratings were averaged across raters to form an overall Army-wide effectiveness rating for each ratee. Tests of job knowledge and hands-on task proficiency were also administered to the soldiers.

Results and Discussion

The performance measures (i.e., hands-on, job knowledge and supervisory ratings) were standardized within each MOS cluster. Then, moderated regression techniques were used to examine determinants of leadership within each MOS cluster. The "moderating" effect of one independent variable on another is indicated by a significant increase in explained variance due to entry of the cross-product term after all main effects have been entered into the model. Separate models were constructed for Supportive and for Participatory leadership. Explanatory variables were entered into the equations in sets; models were tested in the following sequence: (a) main effects of individual difference variables, (b) main effects of performance variables, (c) all main effects, (d) all main effects and interactions between ability and temperament variables, (e) all main effects and interactions among temperament variables, and (f) all main effects and all interactions.

Tab' summarizes the results from all models tested. Among the performance mea s, supervisors' assessments of subordinates' performance predicts reported leadership most consistently. Looking across all of the models, work sample performance and job knowledges do not contribute significantly to Supportive leadership. Task proficiency is related to Participation within the Operations and Skilled Technical MOS clusters; job knowledge predicts Participation within the Operations cluster.

Independently, the set of temperament variables accounts for at least as much variance in reported leadership as individual differences in job performance do. When considered apart from the performance measures, the three temperament measures are significant predictors of rated leadership across MOS. When the performance measures are added, the independent contribution of the temperament measures weakens somewhat, suggesting that these measures share variance with supervisory ratings. Cognitive ability is a significant predictor in only one MOS cluster, the combat related jobs. Although ability does not generally make a direct contribution to the prediction of rated

Table 1

Results of Regression Analyses for Each MOS Cluster

MODEL	Main Effects							Interactions						R^2
	AFQT 1	ACH 2	DEP 3	EMOT 4	HO 5	JK 6	SR 7	AFQT X TEMP. 1*2	1*3	1*4	TEMP. X TEMP. 2*3	2*4	3*4	
Inspiration/Support														
Clerical MOS														
INSP = IND. DIF.	NS	*	**	*										.048
INSP = PERF.					NS	NS	**							.042
INSP = PERF + IND. DIF.	NS	NS	**	NS	NS	NS	**							.072
INSP = MAIN EFF. + AFQT*TEMP	NS	NS	NS	NS	NS	NS	**	NS	NS	*				.093
INSP = MAIN EFF. + TEMP*TEMP	NS	NS	**	NS	NS	NS	**				NS	NS	NS	.076
INSP = MAIN EFF. + ALL INTERACTIONS	NS	NS	NS	NS	NS	NS	*	NS	NS	*	NS	NS	NS	.097
Combat MOS														
INSP = IND. DIF.	*	*	**	**										.091
INSP = PERF.					NS	NS	**							.044
INSP = PERF + IND. DIF.	NS	NS	**	**	NS	NS	**							.112
INSP = MAIN EFF. + AFQT*TEMP	*	NS	**	*	NS	NS	**	NS	NS	NS				.116
INSP = MAIN EFF. + TEMP*TEMP	NS	NS	**	**	NS	NS	**				NS	NS	NS	.115
INSP = MAIN EFF. + ALL INTERACTIONS	*	NS	**	*	NS	NS	**	NS	NS	NS	NS	NS	NS	.116
Operations MOS														
INSP = IND. DIF.	NS	**	**	**										.071
INSP = PERF.					NS	NS	**							.031
INSP = PERF + IND. DIF.	NS	**	**	**	NS	NS	**							.087
INSP = MAIN EFF. + AFQT*TEMP	NS	NS	**	*	NS	NS	**	NS	NS	NS				.090
INSP = MAIN EFF. + TEMP*TEMP	NS	**	**	**	NS	NS	**				NS	NS	NS	.088
INSP = MAIN EFF. + ALL INTERACTIONS	NS	NS	**	*	NS	NS	**	NS	NS	NS	NS	NS	NS	.091
Skilled Technical MOS														
INSP = IND. DIF.	NS	*	**	**										.049
INSP = PERF.	NS	NS	**											.042
INSP = PERF + IND. DIF.	NS	NS	**	**	NS	NS	**							.074
INSP = MAIN EFF. + AFQT*TEMP	NS	NS	NS	NS	NS	NS	**	NS	NS	NS				.074
INSP = MAIN EFF. + TEMP*TEMP	NS	NS	**	**	NS	NS	**				*	**	NS	.089
INSP = MAIN EFF. + ALL INTERACTIONS	NS	NS	NS	NS	NS	NS	**	NS	NS	NS	*	**	NS	.089
Participation														
Clerical MOS														
PART = IND. DIF.	NS	**	**	**										.147
PART = PERF.					NS	*	**							.071
PART = PERF + IND. DIF.	NS	**	**	**	NS	NS	*							.165
PART = MAIN EFF. + AFQT*TEMP	NS	NS	NS	NS	NS	NS	NS	NS	NS	*				.165
PART = MAIN EFF. + TEMP*TEMP	NS	**	**	**	NS	NS	NS				NS	NS	NS	.156
PART = MAIN EFF. + ALL INTERACTIONS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	.166
Combat MOS														
PART = IND. DIF.	NS	**	**	**										.082
PART = PERF.					NS	NS	**							.053
PART = PERF + IND. DIF.	NS	**	**	**	NS	NS	**							.108
PART = MAIN EFF. + AFQT*TEMP	NS	*	*	NS	NS	NS	**	NS	NS	NS				.107
PART = MAIN EFF. + TEMP*TEMP	NS	**	**	*	NS	NS	**	NS	NS	NS	NS	NS	NS	.110
PART = MAIN EFF. + ALL INTERACTIONS	NS	*	*	NS	NS	NS	**	NS	NS	NS	NS	NS	NS	.110
Operations MOS														
PART = IND. DIF.	NS	**	**	**										.106
PART = PERF.					**	*	**							.078
PART = PERF + IND. DIF.	NS	**	*	**	**	*	**							.165
PART = MAIN EFF. + AFQT*TEMP	NS	**	NS	**	**	*	**	NS	NS	NS				.163
PART = MAIN EFF. + TEMP*TEMP	NS	**	*	**	**	**	**				*	NS	NS	.167
PART = MAIN EFF. + ALL INTERACTIONS	NS	**	NS	**	**	**	**	NS	NS	NS	*	NS	NS	.168
Skilled Technical MOS														
PART = IND. DIF.	NS	**	**	**										.089
PART = PERF.					*	NS	**							.060
PART = PERF + IND. DIF.	NS	**	**	**	*	NS	**							.120
PART = MAIN EFF. + AFQT*TEMP	NS	NS	NS	NS	*	NS	**	NS	*	NS				.126
PART = MAIN EFF. + TEMP*TEMP	NS	**	**	**	**	NS	**				NS	NS	NS	.125
PART = MAIN EFF. + ALL INTERACTIONS	NS	NS	NS	NS	*	NS	**	NS	*	NS	NS	NS	NS	.131

Note. ACH = Achievement Orientation; DEP = Dependability; EMOT = Emotional Stability; HO = Hands-on Test score; JK = Job Knowledge test score; SR = Supervisory Rating of Job Performance.

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

leadership, past research suggests that it is an antecedent of job knowledge which, in turn affects hands-on performance and supervisors' assessments of subordinate performance (White, Borman, Hough & Hoffman, 1986). Given the mediational role that job knowledge plays, its failure to to have a direct effect on reported leadership in the present investigation is not surprising.

Without exception, the combination of individual differences in temperament and job performance variables accounts for more variance in leadership measures than either set of variables considered alone. However, the addition of interaction terms offers little advantage. In no case do they increase the amount of variance accounted for by more than two percentage points. Further, the interactions have no consistent pattern of significance.

Finally, the two leadership variables appear to differ in how strongly they are predicted by the independent variables. With the exception of the Combat job cluster, the independent variables account for more variance in Participation than in Inspiration/Support. Further, Achievement Orientation is a significant predictor of Participation, but not of Supportive leadership. Thus, in determining the amount of support a leader will provide, subordinate attributes may contribute less than other determinants of leadership behavior (e.g., leader attributes, organizational norms and values, resource allocation), whereas in most MOS, participation may depend more heavily on subordinate characteristics.

In summary, soldiers who report receiving higher levels of support from their superiors tend receive higher scores in dependability and emotional stability and are seen by their superiors as effective performers. Further, soldiers who report more involvement in work related decisions have the preceding characteristics and are also scored as more achievement oriented.

The present research successfully extended past research in two important ways. First, it demonstrated in a field setting that performance predicts reported leadership. Second, although performance affects soldiers' treatment by their superiors, individual differences in job-related temperament factors are at least equally important. Further, both sets of variables make independent contributions to the prediction of reported leadership. Because treatment by superiors can be predicted from relatively stable individual differences, supervisory treatement should be expected to generalize across supervisors and through time. Thus, subordinates who negotiate more effective relationships with their superiors during the first tour should be expected do so throughout their careers. Additionally, it is likely that future bosses will see these individuals as more effective.

Future research should trace the careers of individuals in the Project A database to determine if, in fact these predictions hold. Furtner, the present research assumed one-way causality; future research might address the bi-directional causality of superior-subordinate interactions.

References

Campbell, C. H. Campbell, R. C., Rumsey, M. G., & Edwards, D.C. (October, 1984). Development and Field Test of Task-Based MOS-Specitic Criterion Measures. (Technical Report No. 717). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (October, 1984). Improving The Selection Classification and Utilization of Army Enlisted Personnel: Annual Report, 1984 Fiscal Year. (Technical Report No. 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Graen, G. B. (1976). Role-making processes within complex organizations. In M. Dunnette (Ed.), Handbook of industrial organizational psycnology. Chicago: Rand McNally.

Graen, G. B., Novak, M. A., & Summerkamp, P. (1982). The effects of leader-member exchange and job design on productivity and satisfaction: Testing a dual attachment model. Organizational Behavior and Human Performance, 30, 109-131.

Greene, C. N. (1975). The reciprocal nature of influence between leader and subordinate. Journal of Applied Psychology, 60, 187-193.

Hough, L. M., Kamp, J. D., & Barge, B. A. (1984). Utility of Temperament, Biodata, and Interest Assessment for Predicting Job Performance: A Review and Integration of the Literature. Minneapolis: Personnel Decisions Research Institute.

Hough, L. M., Gast, I. F., White, L. A., & McCloy, R. (1986, August). The relation of leadership and individual differences to job performance. Paper Presented at the Meeting of the American Psychological Association, Washington, D. C.

Jacobs, T. O. (1971). Leadership and exchange in formal organizations. Alexandria, VA: Human Resources Research Organization (HumRRO).

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current Armed Services Vocational Aptitude Battery (ASVAB) composites. (Technical Report No. 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Novak, M. A. (1985). A study of leader resources as determinants of leader-member exchange. Doctoral dissertation, University of Cincinnati, 1984). Ann Arbor, MI: University Microfilms International.

Vecchio, R. P., & Gobdel, B. C., (1985). The vertical dyad linkage model of leadership: Problems and prospects. Organizational Behavior and Human Performance, 34, 5-20.

Wakabayashi, M., & Graen, G. B., (1984). The Japanese career progress study: A 7-year follow-up. Journal of Applied Psychology, 69, 603-614.

White, L. A., Gast, I. F., & Rumsey, M. G., (1985, November). Leader behavior and the performance of first term soldiers. Paper presented at the meeting of the Military Testing Association, San Diego, CA.

White, L. A., Borman, W. C., Hough, L. M., & Hoffman, R. G. (1986, August). A path analytic model of job performance ratings. Paper presented at the meeting of the American Psychological Association, Washington, D. C.

Measurement of Leader Effectiveness in a Tactical Environment[1]

Douglas L. Rachford
U.S. Army Research Institute

Ray A. Zimmerman
BDM Corporation

This paper presents the initial stages of ongoing research to examine leader effectiveness and the Platoon Leader--Platoon Sergeant relationship in maneuver platoons in a tactical environment. This research was requested by the Center for Army Lessons Learned (CALL) and is consistent with the goals of the leadership research program at the Army Research Institute (ARI) to develop leader performance measures and feedback systems that will improve the leadership training and lessons learned systems within the U.S. Army. This project has two goals, both at the platoon level: First, to pilot a system for collecting "lessons learned" on leadership and, second, to develop pilot measures of leader performance. The term "lessons learned" refers to the more general information gained by performing a task. For example, in the context of leadership training this means what the trainers learned about training leadership rather than what part of the curricula the student's master.

The tactical environment that provides the context for this research is the National Training Center (NTC). Both CALL and ARI have active research programs at NTC and it is regarded as the premier training site and combat simulation in the U.S. Army. For a training session (rotation) at NTC, two task forces (battalions and all of their support elements) deploy to NTC for three weeks of continuous operations. A resident opposing force--well schooled in Warsaw Pact tactics--fights approximately 10 free-play force-on-force missions against the task forces on a battlefield of approximately 600 square miles. Other, live-fire, missions are conducted on a range with pop-up and moving targets. Multiple Integrated Laser Engagement Systems (MILES)--a simulation technology in which laser bursts are fired with blank rounds and detected by receptors on targets--is used on all individuals and weapons systems in the force-on-force missions to simulate and record firing data and kills. Resident Observer/Controllers (OCs) are attached to each unit to provide on-the-spot feedback and insure that the engagement rules are followed. The length, realism, and size of the exercises coupled with the harsh environment and well trained opposing force makes the NTC an excellent training environment. The excellence of the simulation also makes NTC a valuable laboratory for the Army to do research and examine its doctrine.

Method

The first step in developing the data collection system was to review Army doctrine and training literature relevant to troop leading at the platoon level and the Platoon Leader--Platoon Sergeant (PL-PSG) relationship. This yielded a good general framework for troop leading procedures. The Army's Operational Concept of Leadership (TRADOC Pam 525-28) specifies nine leader competencies: planning, communication, supervision, counseling, professional ethics, decision making, technical proficiency, soldier team development, and management technology. This concept also emphasizes teaching, developing subordinate leaders, and initiative as essential to the leader's role.

---

[1] 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, Department of the Army or the BDM Corporation.

From these materials it was clear that the PL has the responsibility for training of the unit and for all of the operations that the platoon undertakes. What was less clear was the role of subordinate leaders, particularly the PSG. The PSG is to assist the PL in any duty assigned. The PSG must be familiar with all aspects of the PL's job and be prepared to assume the role of PL if the PL is absent.

Armed with this information, interviews with officers and senior non-commissioned officers (NCOs) were conducted to gain a better perspective on the issues of platoon level leadership. Several issues and challenges emerged for constructing measures of effective platoon leadership. First, there is clearly too much to do in leading a platoon for the PL to do it all himself. In garrison the PL (usually a 1st or 2nd Lieutenant) has many tasks within the company which prevent him from completely "running" the platoon. In a tactical environment this is also true, especially when continuous operations make sleep an important task in itself. Second, PSGs have a great deal more experience in the Army organization, in troop leading, and in tactical environments than do PLs. Platoon Sergeants usually have 10-15 years of service while PLs usually have from six months to two years. The PSG, therefore, has a base of expert power and experience to draw on that the PL does not possess. Third, and related to the second, PLs are--more than most officers--receiving on-the-job training. They have a good base of technical knowledge from their schooling but must learn the subtleties of leading a platoon while on the job. A question arose frequently in the interviews as to whether it is the job of the PSG to train the Lieutenant to lead the platoon. Other major issues that arose were the resentment that PSGs feel when turning the platoon over to a new PL, the low motivation of some PSGs, the stereotype that PSGs handle "beans and bullets" (logistics) while the PL takes care of tactics, and the myth that the platoon "belongs" to the PSG in garrison but "belongs" to the PL in the field.

With this background, we decided that detailed behavioral observations of platoons in a tactical environment were necessary to understand what leadership functions were performed and who performed them. Such observations could then be linked to platoon performance (to determine if the behaviors observed were indeed important predictors of unit performance) and provide a basis for understanding more global assessments of the PL-PSG relationship. Two types of observation instruments were seen as necessary, one focusing on platoon leadership behaviors and another focusing on interactions between the PL and PSG. It was also decided that because the nature of platoon leadership requires division of labor between the PL and PSG, any platoon leadership behavior might be observed as performed by either the PL or the PSG.

The first step in developing the observation guides was the identification of the leadership dimensions to be observed. Using the competencies from U.S. Army Operational Concept for Leadership and the "Characteristics of effective leaders at NTC" presented in Endicott and Pence (1985), ten dimensions with labels and definitions were drafted. The draft version of dimension labels and definitions were reviewed and approved by CALL with only minor revisions. The ten dimensions and definitions were:

> PLANNING. Developing a course of action, in detail, to meet a specific purpose or objective; taking into account available resources and situation; looking ahead to anticipate problems and contingencies and coming up with solutions; setting priorities for the order of work; managing time effectively.

COMMUNICATION. Conveying information about plans, intentions, expectations, etc.; making sure that the others understand the information; listening carefully and asking questions to make sure that one understands the information one has received; providing information in a timely manner; providing all of the information the the other person needs.

SUPERVISION. Assigning tasks to be performed; watching, checking, directing, and controlling the work performed by subordinates; coordinating the performance of a number of tasks; rewarding subordinates for good performance.

TEACHING AND COUNSELING. Helping others to learn to do a better job by instructing, coaching, or mentoring; giving accurate and constructive feedback to improve work performance; conducting on-the-spot corrections; giving advice and recommendations when appropriate.

TECHNICAL PROFICIENCY. Knowing the standard; knowing how to use equipment; knowing one's own job and subordinates' jobs.

PROFESSIONAL ETHICS. Leading by example; exhibiting and encouraging honesty and integrity in words and deeds; accepting responsibility for own actions and those of subordinates; demonstrating a willingness to learn; maintaining composure during battle engagement simulations; showing a willingness to take calculated risks.

DECISION MAKING. Using available information to make judgments about the appropriate course of action; considering resources, circumstances and guidance to solve on-the-ground problems in a timely manner; choosing a course of action rather than letting events determine the action.

INITIATIVE. Acting to fill a need or seize an opportunity when such an occasion exists; recognizing a deficiency in planning, communication, preparation, or execution and moving to correct it; acting quickly and aggressively to exploit opportunities to expedite the mission consistent with the Commander's intent.

SUBORDINATE LEADER DEVELOPMENT. Helping subordinate leaders to mature by including them in planning and decision making, and by delegating responsibilities; maintaining a close working relationship with subordinates; giving accurate, timely and constructive feedback about leadership behaviors.

SOLDIER-TEAM/COHESION DEVELOPMENT. Developing horizontal bonds between soldiers and vertical bonds between soldiers and leaders by building: confidence in unit abilities, mutual trust and respect, concern for others welfare, unit identity, shared values and goals (consistent with Army values and goals) and a commitment to high unit performance.

A critical incidents workshop was then held with a group of 16 senior NCOs from various branches. The majority of these NCOs, who were instructors at the U.S. Army Sergeants Major Academy, had recent NTC experience. Participants were asked to provide three incidents reflecting ineffective, average, and effective leadership behavior for each dimension. These incidents were to have occurred at NTC or in combat-like environments. A total of 108 usable incidents were generated during the workshop. Although each of the incidents was related to at least one of the ten leadership dimensions, there were 15 incidents which dealt specifically with the interaction between PLs and PSGs.

Next, all of the incidents were rewritten by project staff to reflect more general behaviors (i.e., those which would be applicable for different branches, different types of missions, etc.). The rewritten items constituted the pool of draft items for the observation guides.

The draft items were then reviewed by a group of 11 Observer/Controllers (OCs) at the NTC. The purpose of this review was to insure that the items were worded appropriately and that the the items were general enough to apply to all branches. The review was also intended to provide OCs with the opportunity to add items if they noticed a lack of coverage for important types of leader behaviors or to identify items which might not be observable. A second review of the items was conducted by individuals from the Center for Army Leadership, the U.S. Army Sergeants Major Academy, and the Combined Arms Training Activity. A number of items were modified and the order in which the items would appear in the observation guides (grouped by mission phase for ease of use) was agreed upon.

Following this review, items which reflected leader behavior on the ten dimensions were categorized by dimension. Several additional items were then written by project staff to increase the coverage of some dimensions. Five-point Likert scales of effectiveness (extremely ineffective, ineffective, marginally effective, effective, and extremely effective) were attached to each behavior while similar scales of frequency (never, seldom, occasionally, usually, and always) were attached to items relating to interactions between PL and PSG.

In addition to the observation guides, which were to be completed for each mission, two instruments were developed which would be completed by the data collectors after observing the same platoon for several missions. The first was designed to obtain overall leadership effectiveness ratings for the PL, PSG and squad/crew leaders on each of the ten dimensions of leader behavior. The second instrument elicited the opinions of the data collectors regarding the PL-PSG relationship. This instrument dealt with issues such as whether or not the PL and PSG had clearly defined roles and responsibilities, whether or not they respected each other's capabilities, etc.

Finally, a set of training materials was developed to prepare data collectors to use the observation guides. The material for the training workshop included instruction on commonly occurring rating errors (halo effects, leniency/severity error, etc.) and a set of 20 sample scenarios to provide the opportunity to practice making the ratings.

Data Collection

Eleven data collectors were sent to NTC two days prior to the units going to the field. Nine of these collectors were senior NCOs and the other two were civilian researchers. The data collectors studied the observation guides overt night (7 data collectors had participated in generating critical incidents or in instrument reviews) and the next day participated in a four hour training

session to alert them to rating errors and to practice rating hypothetical situations. Each data collector then went to the field attached to an NTC Observer/Controller (OC) to stay with a platoon--day and night. The data collectors rode in the OC vehicle and attempted to observe each behavior in the instruments. This meant attending operations orders, observing preparations for combat, directly observing battles, attending the after action reviews, monitoring radio nets, and generally "hanging around" the platoon as much as possible without interfering with its operations. The platoons were observed only in force-on-force missions because the live-fire missions--in addition to being dangerous--were judged to be inappropriate for this task.

The initial observation plan was that each data collector would make behavioral observations on a platoon for three missions, make overall ratings, and then switch to another OC and platoon. Logistical difficulties made it impossible for data collectors to switch platoons in many cases. Other problems and misunderstandings also interfered with the data collection plan with the result that, rather than 33 platoons being observed, 16 separate platoons were observed on a minimum of two missions (additionally, three platoons were observed twice by different data collectors on different missions).

Discussion of Measurement Issues

It is premature to present any of the data collected at the NTC. The analysis of these data is in the preliminary stages. In addition, the size of the sample precludes many of the analyses that were initially planned. We are currently analyzing the items to determine the frequency of use and whether any were unclear or problematic. We are also examining each item's relation to the dimensions they were intended to reflect by comparing them to the overall ratings on the leadership dimensions. We will then revise the observation guides so that--with additional data collection--estimates of internal and interrater reliability can be made. Efforts to validate the instrument will also be made.

A summary of the lessons learned about such a measurement effort at the NTC and discussion of the issues is, however, possible and instructive for future efforts in this or similar tactical environments.

We demonstrated that observation guides can be developed and used to capture important leader behavior in a tactical setting. The data collectors had little problem using the instruments and, in fact, indicated that they were easily able to identify and rate the specific behaviors when they saw them. Many OCs commented that the observation guides were very complete and would be useful in their work (several asked for copies). We were satisfied that, with training, observers could collect detailed data in a field environment. We also learned that collecting data during a large fast-moving combat exercise stretches Murphy's law to its limit. The greatest problem we faced was collecting data on enough platoons to have an adequate sample. Additional data collections of the type described above would be useful, but this solution would not accomplish the long-term goal of developing a feedback system. For further development of a leader performance measurement and feedback system at NTC--including measurement at levels above and below platoons--several problems must be solved. Refinement of the measurement instruments will be important but other issues loom as larger obstacles to overcome in accomplishing the long-term research goals.

We have arrived, in our planning, at an approach to these problems that takes into consideration the goals of the organization and its individuals, the

resources available and the constraints within the context; in other words, a systems approach. In the long-run, to use data from NTC effectively, the NTC Observer/Controllers must collect the necessary measures of leadership, technical and tactical performance. They must be trained, their ratings calibrated, and calibration maintained. The information they collect must then become part of a feedback system that provides information to individuals and units for training purposes and contributes to the Army's "lessons learned" database for decisions about policy, doctrine and curricula. Such a program would also provide data in the quantity necessary for continued research on leadership, command and control, training, and performance measurement.

References

Endicott, S. & Pence, E.C. (1985), Leadership Lessons Learned at the National Training Center (NTC). Unpublished Manuscript, Center for Army Leadership.

United States Army Training and Doctrine Command (TRADOC). (1983), U.S. Army Operational Concept for Leadership. Pamphlet 525-28.

LEADER REQUIREMENTS TASK ANALYSIS[1]

Alma G. Steinberg, Paul van Rijn, and Fumiyo T. Hunter
U.S. Army Research Institute

The U.S. Army Research Institute (ARI) is conducting research to assist the Center for Army Leadership (CAL) and the U.S. Army Sergeants Major Academy (USASMA) in designing and maintaining a sequential and progressive leadership development program for commissioned and noncommissioned officers. This research will provide current, Army-wide data about the leadership tasks required of NCOs (E5 through E9) and officers (Ø1 through Ø6); and the tools and methodology to update this information, as necessary, in the future. It will provide CAL and USASMA with an empirical basis for:

(a) designing leadership development programs that take into account how leadership tasks change from level to level in the Army.

(b) determining needed instructional areas not presently addressed, and the levels for which they may be appropriate.

(c) identifying and addressing similarities and differences in leadership training requirements for different branches of the Army.

(d) determining the appropriate time to be allotted to blocks of leadership instruction.

(e) identifying possible discrepancies between leadership doctrine and what leaders actually do.

The basic approach of this research is an occupational task analysis survey, specially adapted for application to the area of leadership. The task analysis survey approach is especially advantageous because it provides the requisite information in a standardized format suitable for comparisons across groups (e.g., ranks and branches) and can be administered Army-wide. It conforms with format requirements of the Army Occupational Survey Program (AOSP) and provides the Army with a viable avenue to update leader requirements information in the future, as needed. The procedures used to develop the task inventory and the special adaptations for the area of leadership are described below.

---

[1]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.

Procedures for Developing the Leadership Task Inventory

There are two main approaches to developing task inventories (McCormick, 1979). One approach relies on the existing literature - military, non-military, empirical and/or theoretical - to build the inventory (e.g., Clement & Ayres, 1976). The other approach relies on interviews with job incumbents about the tasks they perform. The latter approach was selected for this research as being the one that would most closely meet the requirement to document the leadership tasks that NCOs and officers perform in today's Army.

Small-group interviews (usually about 6 to a group, but sometimes as many as 15 or more) were conducted with several hundred NCOs and officers at a variety of locations including Ft. Hood, Ft. Campbell, Ft. Belvoir, Ft. Carson, Ft. Polk, Ft. Bliss, Ft. Lee, and Ft. Eustis. The interview sessions were approximately 1 1/2 hours in length and focused on present, and sometimes previous, jobs of the incumbents. Interviewees also were asked to indicate the similarities and differences between the leadership portion of their own jobs and the jobs of: (a) soldiers at the same rank in other branches, and (b) soldiers one rank higher and lower than themselves. The task inventory was both developed and reviewed in an iterative process over the course of the interviews. Interviewees were asked to describe what they did in their jobs and then to review the leadership tasks developed from earlier sessions with other incumbents. They were asked also to comment on some tasks derived from other sources, such as instructors in the schoolhouse, retired military personnel, leadership literature and doctrine, instruction manuals, and other task lists.

With each set of interviews, new tasks were added to the preliminary task inventory, existing tasks were clarified, and tasks in similar topic areas were grouped. Headings were selected to reflect the nature of the tasks in each group or duty area.

The completed task list then was reviewed by subject-matter experts (SMEs) at CAL and USASMA to ensure that it was clear, accurate, and complete. Recommendations from these experts guided the final consolidation of the task inventory prior to ARI's internal review.

Special Adaptations for the Area of Leadership

The following discussion highlights some of the ways the traditional task analytic approach was adapted in the current research effort to be more suitable for use in the leadership domain.

Delineation of the leadership portion of the job

In traditional technical task analysis, incumbents have relatively little trouble identifying the technical tasks they do. Identifying the leadership tasks is much more problematic because they are intertwined with the technical tasks and because many different perspectives exist on the boundaries of the leadership domain (Bass, 1981). Thus, the definition of leadership in the Army's field manual on military leadership (FM 22-100, Headquarters Department of the Army, 1983) was used for clarification. Military leadership, as defined in the manual, is "a process by which a soldier influences others to accomplish the mission" (p. 44).

Duty headings

A frequent format for headings of duty areas (groups of similar tasks) is a single word or concept. In the leadership arena this format makes it difficult to determine the appropriate grouping of tasks into duty areas. For example, the Air Force leadership survey conducted in May 1983 (Occupational Survey Branch) contains a duty area heading titled "communicating." Since other duty areas in the survey definitely involve communicating (counseling, maintaining discipline, motivating, training, etc.), it is not apparent what the duty heading "communicating" means. The approach taken in the current survey was to eliminate some of the ambiguity of duty headings by putting them into the same format as task statements. Although this approach does not eliminate all the ambiguity, it communicates the intent of a grouping better (e.g., "maintain 2-way information exchange with superiors").

Task structure

The task statements in the present inventory are constructed as follows: a verb, an object, and, if appropriate, a modifier. Multiple verbs as in the task "OPERATING INSTRUCTIONS, implement and issue" (American Institutes for Research, 1975, p. 18) are avoided. Respondents who only do one of these (implement or issue) find it difficult to respond to these tasks.

Task format

Traditional Army occupational job analyses consist of behavioral tasks which state what individuals actually do while performing a task. The tasks are discrete units of work that have discernible beginnings and endings (McCormick, 1979). Two examples of military leadership tasks which follow this format are: "conduct briefings" and "make on-the-spot corrections."

In order to cover the domain of leadership adequately, however, it is necessary to supplement these behavioral tasks with what Fleishman calls "the behavior requirements approach" (Fleishman, 1982, p. 827). This approach allows reference to inferred processes which may intervene between stimulus events and responses, and includes tasks that are less likely to have a clear, observable beginning and end. Examples of such military leadership tasks are "motivate troops to close with the enemy" and "assess the climate of the unit."

In the leadership area, the distinction between behavioral tasks and behavioral requirements is often blurred. For example, the task "evaluate group performance" may be either. It may be observable with a clear beginning and end on some occasions and not so on others. An example of the former case is the leader visibly timing the group on how long it takes to complete a specific training exercise. An example of the latter case is the leader casually observing the group's performance and making mental notes on the quality of performance.

Task specificity/generality

Although it is frequently recommended that the same level of task specificity be maintained throughout a task inventory, and that the level be not too general and not too specific (Fleishman, 1982), the leader requirements inventory was designed with tasks which vary from general to specific. This practice was followed in order to focus on those aspects of leadership which SMEs hypothesize differentiate leadership tasks of various sub-groups (line and staff, branches, etc.) and various levels within the organization. At the broad end of the spectrum are tasks such as "motivate subordinates" and "delegate decision-making to subordinates." At the other end of the spectrum are much more specific tasks such as "set up command post" and "conduct court martial proceedings."

Another reason for the generality of some of the tasks is the objective of targeting the leadership process rather than the technical portion of the job. For example, generic words such as "vehicles" replace specific categories of military vehicles in order to discover whether the leadership process is the same across different jobs and different ranks.

Task overlap

As a consequence of the varied levels of specificity discussed above and also the nature of leadership, it is not possible to cover the domain of leadership adequately without including tasks that overlap. For example, the task

"supervise soldiers" could be said to encompass many other more specific tasks (e.g., "motivate soldiers who have attitude problems"), but the inclusion of this general task enables the determination of who supervises soldiers and who does not at a single glance. Note that this motivation task also could be performed by soldiers who do not supervise other soldiers.

On the other hand, another sort of overlap is strictly excluded. The following three tasks are an illustration of the type of overlap that is repetitive, pointless, and confusing (American Institutes for Research, 1976, p. 79):

(a) SUBORDINATES, evaluate
(b) SUBORDINATES, interview, consult and counsel
(c) SUBORDINATES, motivate, evaluate and counsel

Task balance

Due to the broad scope of the jobs to be included, it was impossible to keep the inventory to a reasonable length and yet at the same time include all alternative tasks in the domain. Only when the alternative tasks were clearly needed for comparison purposes were they included. Examples of balanced tasks are "motivate subordinates" and "motivate superiors." (Had there been no value in distinguishing who was to be motivated, the task could simply have read "motivate others.")

Conclusion

The Leader Requirements survey is scheduled to be distributed early in 1987. It is anticipated that, in addition to providing empirical data on leader requirements Army-wide and a methodology for updating this data, it will provide (a) a methodology for leadership task analysis that can be applied in other organizational settings and (b) an empirical basis for examining leadership theories which address the leader's role as a function of organizational level. In the leadership area there has been much conjecture and relatively little supporting data.

References

American Institutes for Research (1975). Duty module methodology for officer management system development: Task Data Bank Index, Research Note 79-33. Alexandria, VA: U.S. Army Research Institute.

Bass, B. M. (1981). Stogdill's handbook of leadership: A survey of theory and research. New York: The Free Press.

Clement, S. D., & Ayres, D. B. (1976). A matrix of organizational leadership dimensions. Leadership Monograph Series No. 8. Ft. Benjamin Harrison, IN: U.S. Army Administration Center.

Headquarters Department of the Army (1983). Military leadership, FM 22-100. Washington, DC: Author.

Fleishman, E. A. (1982). Systems for describing human tasks. American Psychologist, 37, 821-834.

McCormick, E. J. (1979). Job analysis: Methods and applications. New York: AMACOM.

Occupational Survey Branch (1983). Officer professional military education (PME) task list: Paygrades 0-1 through 0-6. Randolph AFB, TX: USAF Occupational Measurement Center.

FACTOR ANALYSIS OF COMPOSITE SCORES FROM THE ARMED SERVICES VOCATIONAL APTITUDE BATTERY (ASVAB)

W. P. Dunlap[1], A. C. Bittner, Jr.[2], M. B. Jones[3] and R. S. Kennedy[4]

[1]Tulane University, New Orleans, Louisiana
[2]Analytics Inc., Willow Grove, Pennsylvania
[3]Hershey Medical School, Hershey, Pennsylvania
[4]Essex Corporation, Orlando, Florida

ABSTRACT

An intercorrelation matrix for ASVAB composite scores was calculated from the intercorrelations among ASVAB subtest scores provided by Kass et al. (1982). Composite score reliabilities were estimated in a similar manner. The resulting statistics were factor analyzed and several lines of evidence supported the existence of only a single common factor. First, the eigenvalue of the second factor in Principal Component Analysis fell well below one. Second, when a two-factor solution was attempted by Principal Factor Analysis, the communalities of four composites exceeded their reliabilities, the theoretical upper limit for communalities. This finding implies that use of ASVAB composite scores for personnel assignment represents univariate as opposed to multivariate selection. If ASVAB composites provide the basis of personnel assignment, departures from univariate selection depend on residual variance primarily due to errors of measurement. A similar hazard exists whenever composites of the same subtest scores are used for job assignment

McCormick, Dunlap, Kennedy, and Jones (1982) gave the Armed Services Vocational Aptitude Battery (ASVAB) repeatedly to a group of 57 trainees in a Job Corps Center. A reanalysis of their data showed that if a factor analysis were performed, not on the subtests of the ASVAB, but instead, on the composite scores obtained from those subtests, only a single factor was found. This finding seemed of some importance because it is the composite scores, rather than the original ASVAB subtest scores, that are used for job assignment by the Armed Forces. If in fact these composite scores have but one underlying factor, and if the remaining variance in the ASVAB composite scores is wholly or largely error, then the ASVAB's use in personnel assignment is best understood as univariate selection or placement rather than multivariate (Hunter & Schmidt, 1982). One limitation of the McCormick et al. (1982) study was that the number of subjects used was less than that conventionally deemed sufficient to support factor analysis (Gorsuch, 1974, p. 296). A second limitation was that the subject population might not be representative of the military job applicant pool. A larger and clearly representative sample is required to resolve the issue of the underlying factor structure of the ASVAB composites.

Kass, Mitchell, Grafton, and Wing (1982) have provided a factor analysis of the ASVAB subtest scores for a sample of 98,689 subjects. Their sample represented a randomly selected 20% of the fiscal year 1982 Armed Forces

Paper presented at the 28th Annual Military Testing Association Conference, Mystic, CT, November 3-7. 1986

applicants. Unfortunately, Kass et al. (1982) did not report a factor analysis of the composite scores, the scores that are actually used for job classification. The present study used the Kass et al. data to reconstruct the intercorrelation matrix of the composite scores, and then factor analyzed the composite scores.

METHOD

The Composite Correlation Matrix

The first step in reconstructing the composite correlation matrix was to convert the Kass et al. (1982) correlation matrix among ASVAB subtests to a variance-covariance (V-C) matrix among subtests. This was accomplished by pre- and postmultiplying the correlation matrix (Kass et al., 1982, Table 2) by vectors containing standard deviations (from Kass et al., Table 1). This operation is the reverse of the computation of a correlation from a covariance matrix (Anderson, 1958).

Next, a matrix containing column vectors of 0's and 1's for each of the ASVAB composites was constructed as shown in Table 1. The first column of this table indicates that the composite score for "Combat" (CO) is obtained by adding together ASVAB subtest scores for arithmetic reasoning (AR), coding speed (CS), auto-shop information (AS), and mechanical comprehension (MC); all other subtests are weighted by zero and consequently are ignored. The V-C matrix for the composite scores was then computed by premultiplying the V-C matrix of ASVAB subtest scores by the transpose of the coefficient matrix in Table 1, then postmultiplying the result by the coefficient matrix. This is a standard procedure for computing the V-C matrix of a set of composite scores (Anderson, 1958).

Table 1. Coefficient Matrix for Converting ASVAB Subtests to ASVAB Composites (0 indicates to ignore subtest; 1 indicates to include subtest)

	Composite ASVAB Scores*									
ASVAB SUBTESTS	CO	FA	EL	OF	SC	MM	GM	CE	ST	GT
General Science (GS)	0	0	1	0	0	0	1	0	1	0
Arith Reasoning (AR)	1	1	1	0	0	0	0	0	0	1
Word Knowledge (WK)	0	0	0	1	1	0	0	1	1	1
Paragraph Comp (PC)	0	0	0	1	1	0	0	1	1	1
Numerical Oper (NO)	0	0	0	1	1	1	0	1	0	0
Coding Speed (CS)	1	1	0	0	1	0	0	1	0	0
Auto Shop Info (AS)	1	0	0	1	1	1	1	0	0	0
Math Knowledge (MK)	0	1	1	0	0	0	1	0	1	0
Mechanical Comp (MC)	1	1	0	1	0	1	0	0	1	0
Electronics Inf (EI)	0	0	1	0	0	1	1	0	0	0

*Note: The Composite Score abbreviations are defined in Table 2.

Last, the V-C matrix of composite scores was converted to a correlation matrix by pre- and postmultiplying it by vectors containing the reciprocals of the square roots of its diagonal entries; these diagonal entries are, of course, the composite variances (Anderson, 1958).

Reliabilities

Reliabilities of the composite scores were computed in a manner analogous to that described by Mulaik (1972, p. 72-76). First, it was assumed that the cross-correlations between subtests on successive administrations of the ASVAB would be well approximated by the intercorrelations between subtests on a single administration (Kass et al., 1982, Table 2). The intercorrelations should tend to overestimate the cross-correlations because of the correlation of error components during a single administration (e.g., Thorndike, 1949; Bittner, Carter, Krause, Kennedy, & Harbeson, 1983). Because it is likely that these intercorrelations somewhat overestimate the cross-correlations, the reliabilities computed for composite scores should be slightly overestimated. The diagonal of the correlation matrix between subtests was replaced with the reliabilities of each subtest given by Kass et al. (1982, Table 1). These correlations were then converted to covariances by pre- and postmultiplying with vectors containing the subtest standard deviations. Next, composite covariances were computed by pre- and postmultiplying by the coefficient matrix, Table 1.

Table 2. Intercorrelations of ASVAB Composite Scores

		CO	FA	EL	OF	SC	MM	GM	CE	ST
Combat	(CO)									
Field Artillery	(FA)	975								
Electronics	(EL)	824	834							
Operators/Foods	(OF)	875	838	870						
Surveil/Communic	(SC)	938	927	765	916					
Motor Mainten	(MM)	870	819	846	963	875				
General Mainten	(GM)	797	753	947	880	738	878			
Clerical	(CE)	903	918	711	865	988	813	651		
Skilled Tech	(ST)	957	967	836	898	960	837	793	946	
General Tech	(GT)	817	812	930	901	810	808	861	767	859

Note: Decimal points omitted.

Finally, the covariances were converted to correlations by pre- and postmultiplication with vectors containing the reciprocals of the square roots of the composite variances, obtained during the computation of the composite correlation matrix. The diagonal elements of the resulting matrix are the estimated reliabilities of the composites, presented later in the first column of Table 4. It should be remembered that these estimated reliabilities are likely to be somewhat too high, owing to the use of intratest rather than cross-test correlations.

Factor Analysis

Both Principal Component Analysis (PCA) and Principal Factor Analysis (PFA) were performed. Communalities for the PFA were iterated to stability using the composite reliabilities as initial estimates. The program used was BMDP4M from the BMDP Statistical Software Package (Dixon, 1981).

RESULTS

Table 3 presents the eigenvalues and corresponding percents of total variance explained for both the PCA and PFA. The first factor accounted for 88% (PCA) or 86% (PFA) of the total variance among ASVAB composite scores. The second factor accounted for less than 7% (PCA) or approximately 5% (PFA) of the remaining variance. It seemed reasonable at this point to conclude the existence of only one factor among ASVAB composites, using Guttman's (1954) admittedly conservative rule for lower bounds on rank. The eigenvalue for the second factor in the PCA (0.674) fell well short of the criterion for inclusion of unity. The factor loadings for the first factor identified by PCA and PFA are displayed as the last two columns of Table 3. As can be seen, all loadings (the correlations between the composite scores and the factor) exceed 0.9 except for "General Maintenance," which was only slightly lower.

Table 3
Eigenvalues, Percent of Variance, and Factor Loadings for Principal Components Analysis (PCA) and Principal Factor Analysis (PFA) of the ASVAB Composite Scores.

Factor	PC		PFA		Composite	Factor Loadings	
Number	EV	% VAR	EV	% VAR	NAME	PCA	PFA
1	8.755	(88)	8.620	(86)	CO	.958	.956
2	.674	(94)	.502	(91)	FA	.947	.942
3	.261	(97)	.136	(93)	EL	.914	.901
4	.177	(99)	.033	(93)	OF	.963	.962
5	.060	(99)	--	--*	SC	.955	.952
6	.047	(100)	--	--	MM	.931	.922
7	.021	(100)	--	--	GM	.885	.867
8	.003	(100)	--	--	CE	.917	.906
9	.002	(100)	--	--	ST	.969	.970
10	.000	(100)	--	--	GT	.915	.902

*Note: Eigenvalues less than zero.

A second line of evidence that strongly supports the existence of only a single factor comes from a comparison of the reliabilities of the composites to the communalities produced by PFA when one-factor, two-factor, and three-factor solutions were attempted. Recall that theoretically the true score variance of a measure is best estimated by the reliability of that measure. This follows from the common factor model in which the total variance of any measure is decomposed into three parts: the communality (h^2), which is the variance shared by the measure with the common factor(s); the specific variance (s^2), true score variance of the measure that is not shared with the common factors ($s^2=r_{xx}-h^2$), where r_{xx} is the reliability); and error (e^2), which equals $1-r_{xx}$. The reliability of a measure, therefore, sets an upper bound on the communality of that measure (see Gorsuch, 1974, p. 23-30, for discussion). Under the common factor model, the amount of nonerror variance that may be distributed among common and specific factors is the sum of the reliabilities, which equals 9.403. Therefore, under this model, the first factor actually accounts for about 92% of the true score variance (8.620/9.403).

In Table 4 are presented the reliabilities and error components of the ASVAB composite scores, together with the communalities and specific variances from the one-factor and two-factor solutions by PFA. The

attempted three-factor solution failed to converge because some of the estimated communalities exceeded one. As may be seen in Table 4, the one-factor solution comes close to exhausting the true score variance, leaving only a small amount of specific variance mainly in composites GM, EL, and perhaps GT. It is interesting to note that GM and EL are the only composites that contain the ASVAB subtest GS (General Science). The last two columns of Table 4 show the outcome of a two-factor solution. It can be seen that in four instances the communalities exceeded their theoretical upper limits ($h^2 > r_{xx}$), yielding negative specific variances in those instances. One might still accept this apparent violation of the underlying model of common factor analysis if a great deal of interpretative power were to be gained by a two-factor solution. One must remember, however, that (1) the reliabilities of the composites were most likely overestimated; (2) the PCA eigenvalue for the second factor fell well below one; and (3) the second factor accounts for only 5% of the total variance among all composites.

Table 4
Reliabilities and Error Variances for the ASVAB Composites and the Communalities and Specific Variances for One- and Two-Factor Solutions

Composite Name	r_{xx} Est. Reliab.	e^2 $(1-r_{xx})$ Error	1-Factor h^2 Comm.	1-Factor s^2 $(r_{xx}-h2)$ Spec.	2-Factor h^2 Comm.	2-Factor s^2 $(r_{xx}-h2)$ Spec.
CO	.929	.071	.915	.014	.930*	-.001*
FA	.932	.068	.887	.045	.918	.014
EL	.966	.034	.812	.154	.947	.019
OF	.941	.059	.925	.016	.927	.014
SC	.938	.062	.907	.031	.985*	-.047*
MM	.917	.083	.850	.067	.855	.062
GM	.960	.040	.751	.209	.953	.007
CE	.930	.070	.821	.109	.972*	-.042*
ST	.940	.058	.941	.001	.966*	-.024*
GT	.950	.050	.814	.136	.855	.095

*Note: Cases where the communality exceedes the reliability.

Table 5 presents the unrotated and rotated two-factor solution by PFA. Note that none of the loadings in the second unrotated factor exceed 0.5. Furthermore, when rotated, a number of variables load highly on both factors, suggesting that essentially a single factor is being split.

Table 5. ASVAB Composite Factor Loadings for the Two-Factor Principal Factor Analysis Solution

	Unrotated		Varimax Rotated	
Composite	I	II	I	II
CO	.955	-.138	.794	.548
FA	.941	-.177	.811	.510
EL	.912	.340	.437	.870
OF	.958	.090	.641	.718
SC	.957	-.263	.880	.458
MM	.919	.103	.604	.701
GM	.884	.415	.366	.905
CE	.918	-.360	.918	.360
ST	.969	-.163	.821	.540
GT	.903	.199	.527	.760

DISCUSSION

The major implication of the above findings is that the use of the ASVAB composite scores as multivariate selection may not be justified. Although the 10 subtests of the ASVAB tap several underlying factors, the ASVAB composites tend to blend these factors so as to blur, if not destroy, this multidimensional structure. It appears that, to an overwhelming extent, the composite scores are measuring the same thing; 86% of the total variance (92% of true score variance) among composites is accounted for by a single factor. Of the remaining 14% of total variance, approximately 7% is error variance, as may be seen in Table 3. This leaves approximately 7% of the total variance among composites available for multivariate selection. In itself, a single common factor does not preclude substantial gains for multivariate over univariate selection, even if the correlation among composite scores are high (Brogden, 1951). If, however, the specific variance is error, then the one common factor among the composites is the only source of variation that any of the composites can share with job performance. In such a case, the one common factor is the only available predictor, apart from error. The most defensible procedure, in this case, would be to estimate each individual's common-factor scores using all 10 subscales, and then use the common-factor scores for placement. There is, of course, the 7% of total variance that is not accounted for by the one common factor and that is not error either. If much or most of this 7% is shared with job performance, there would be some basis for multivariate selection. This basis would be severely limited in any event, however, and would depend upon an empirical showing that the ASVAB composites share some variance with job performance apart from the one common factor that they share with one another. In the absence of such a showing, the use of the ASVAB to index specific abilities supposedly congruent with particular military specialties would seem not to be justified, except possibly on nontechnical grounds. It would be simpler and more efficient to use the ASVAB as a univariate placement indicator.

The ASVAB is not, of course, uniquely liable to reduce to a single common factor. In general, composite scores based on the same subtests will be strongly correlated, in part because error components in the subtest scores are shared by all composites that contain them. If several subtests are common to the same composites, the shared error variances mount to sizable proportions. Hence, ostensibly multivariate selection based on composites of the same subtests should always be checked for possible dependence on a single common factor that all but exhausts the composites' true-score variance.

REFERENCES

Anderson, T. W. (1958). An introduction to multivariate statistical analysis. New York: John Wiley & Sons.

Bittner, A. C., Carter, R. C., Krause, M., Kennedy, R. S., & Harbeson, M. M. (1983). Performance Evaluation Tests for Environmental Research (PETER): Moran and computer batteries. Aviation, Space and Environmental Medicine, 54, 923-928.

Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 11, 173-196.

Dixon, W. J. (1981). BMDP statistical software. Berkeley: University of California Press.

Gorsuch, R. L. (1974). Factor analysis. Philadelphia, PA: W. B. Saunders.

Guttman, L. (1954). Some necessary conditions for common factor analysis. Psychometrika, 19, 149-161.

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. D. Dunnette & E. A. Fleishman (Eds.), Human performance and productivity: Human capability measurement. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1982). Factor structure of the Armed Forces Vocational Aptitude Battery (ASVAB), Forms, 8, 9, and 10: 1981 Army applicant sample (ARI Technical Report 581). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McCormick, B. K., Dunlap, W. P., Kennedy, R. S., & Jones, M. B. (1983). The effects of practice on the Armed Services Vocational Aptitude Battery (Technical Report No. 602). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Mulaik, S. A. (1972). The foundations of factor analysis. New York: McGraw-Hill.

Thorndike, R. L. (1949). Personnel selection. New York: John Wiley & Sons.

OPTIMIZING A TEST BATTERY BY VARYING SUBTEST TIMES

Dunlap, W. P.[1], Jones, M. B.[2], Kemery, E. R.[1],
& Kennedy, R. S.[3]

[1]Tulane University
New Orleans, Louisiana
[2]Hershey Medical College
Hershey, Pennsylvania
[3]Essex Corporation
Orlando, Florida

Abstract

Varying the length of a test in terms of either number of items or testing time is known to affect reliabilities, cross-correlations, and validities in predictable ways. Using equations for these effects, a systematic procedure is derived that permits varying subtest times within a battery to optimize the ability of the battery to predict a given criterion. Computer software was developed and tested with normative data from the Armed Services Vocational Aptitude Battery (ASVAB), using selected subscales as surrogate criteria. The optimization procedure results in dropping approximately half of the subtests while allocating differing administration times to the remaining subtests. As a result, the multiple correlation coefficient between the battery and criterion is raised somewhat. The more important finding, however, was that optimized batteries could be developed that required one half or one quarter the administration time of the original battery, and still retain as much predictive power as the original. Implications are discussed in terms of the benefits attainable in terms of costs and throughput in performance selection by such test battery optimization.

Introduction

When designing a personnel selection system, a paramount concern is the allocation of resources for optimal employment decisions. In particular, test batteries must be designed such that the amount of information can be obtained in the most efficient way. This is especially true in organizations such as the U.S. military that screen a large number of job candidates yearly. As noted by Dunnette and Borman (1979), important advances have accrued in the past decade in the areas of statistical thinking and methodology pertinent to the design of personnel selection systems. However, they conclude that a wide range of multivariate prediction studies must be conducted to help provide improved guidelines for developing prediction systems. One parameter in need of investigation is the relative emphasis to be placed on components of a test battery in order to optimize prediction and efficiency.

Paper presented at the 28th Annual Military Testing Association Conference, Mystic, CT, November 3-7, 1986

Elements of a test battery which are to be employed to predict operational or training performance are often designed to be administered in a fixed time period, but could be shortened or lengthened, within wide limits, as desired. Given that subtests in a battery can be modified in length, a critical question that arises regards the optimal allocation of subtest times. Could the validity of the battery as a whole against a given criterion be improved by lengthening some subtests while shortening or dropping others? This question is directly addressable from existing psychometric theory. Holding constant the total amount of time for test battery administration, various possible subtest time allotments can be studied in an attempt to maximize their multiple correlation with the criterion.

Psychometric Theory. The Spearman-Brown Prophecy Formula yields an estimate of a test's reliability when its length is either increased or decreased (Gulliksen, 1950). While often thought of in terms of number of test items, the formula can be adapted for use with timed subtests. If $c=t_n/t_o$, where t_o is the time allotted to the subtest in the original battery, and t_n is the new time allotted, then the reliability of the lengthened (or shortened) subtest will be

$$r_{XX} = r_{xx}c/[1+(c-1)r_{xx}]. \quad (1)$$

Furthermore, the validity coefficient between predictor X and criterion Y adjusted for an increase (or decrease) in a test's length will be

$$r_{YX} = r_{yx}\{c/[1+(c-1)r_{xx}]\}^{1/2}, \quad (2)$$

and the adjusted cross-correlation between any two predictor subtests X and W will be

$$r_{XW}=r_{xw}\{c_1/[1+(c_1)-1)r_{xx}]c_2/[1+(c_2-1)r_{ww}]\}^{1/2}. \quad (3)$$

The importance of these equations is that given the matrix of subtest intercorrelations from the original battery, it is possible to compute the subtest intercorrelation matrix that would obtain if different length subtests were used with Eq. 3. Also, the validities of the subtests from the original battery in predicting the criterion can be recomputed to reflect the modified subtest lengths, using Eq. 2. Last, using the recomputed subtest intercorrelation matrix and the updated validity vector, one can compute the multiple regression coefficient for the new allocation of subtest times. Thus, by studying the behavior of the multiple regression coefficient with different allocations of subtest times, it is possible to mathematically redesign the test battery so as to optimize the prediction of th criterion by the test battery as a whole.

Approach

The purpose of this study is to develop the notion that subtests may be allocated based on psychometric considerations, and this information may be used to design a more compacted personnel selection system. In this regard, a computer program was developed to estimate allocation times to subtests within a test battery with the aim of maximizing its predictive efficiency.

Method

Optimization Software. A computer program was written to accomplish the following: 1) input intercorrelations, reliabilities, validities, and subtest times for a test battery; 2) compute the squared multiple regression coefficient from the original allocation of test times; 3) select a new set of randomly determined allocations; 4) correct the original matrix according to the new subtest time allocation using Eq. 2 and 3; 5) compute the squared multiple regression coefficient for the redesigned test battery; 6) converge on an optimum reallocation of subtest times for maximal prediction of the given criterion using an iterative algorithm.

The convergence algorithm selected was rather crude; the times allocated to any subtest were randomly varied between 0.5 and 1.5 times their previously best assignment using a random rectangular number generator. Each time the resulting multiple R^2 surpassed the previous best R^2, the R^2 and subtest times were stored as the comparison values for the next round of iteration. Iteration was continued until 2000 random adjustments failed to yield an improvement against the previous best time allocation. A higher number of iterations could have been employed, but the obtained multiple R^2's and time allocations from 2000 iterations differed trivially from those obtained from a much larger number of iterations.

Confirmative Testing. The test optimization program was tested using normative data provided by Kass, Mitchell, Grafton, and Wing (1982) for the Armed Services Vocational Aptitude Battery (ASVAB) from 98,689 subjects. The ASVAB has 10 subtests, so each subtest was used in turn, as a surrogate criterion, predicted from the remaining nine scales.

The optimization procedure was done for a battery having the same total test time, approximately 130 min. with nine subtests, as was allocated for those subtests in the original battery. Optimization was also done for a battery one half the total length, about 65 min., and one quarter the original length, about 33 min. The optimized R^2 and allocation time (rounded to the nearest minute) were calculated for each test length.

Results

The results obtained from the optimization program are displayed in Table 1. The first row of Table 1 shows the time allotted to each scale in the original battery; obviously the time required for the remaining nine scales will vary somewhat depending upon the length of the scale selected as the surrogate criterion. The dashes indicate cases where the optimization program assigned a time of less than 30 sec.; however, most of these values were actually set at zero by the optimizing program.

It is readily apparent that the optimization considerably reduces the number of subtests retained in the redesigned battery. With an optimized battery of the original length, the number of subtests retained was about 53% on average; for a redesign battery of one half, the original time 43% were retained, and for the one-quarter length battery, 40% of the subtests were selected. Therefore, as might be predicted, subtests that do not

Table 1

Optimized Subtest Times (in min.) for Each Subtest Predicted From All Other Subtests for Test Batteries of One Times, One Half, or One Quarter the Original Test Length, and Associated Multiple R Squared

	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI	R_2
						Original					
Time	11	36	11	13	3	7	11	24	19	9	
						General Science					.743
1		--	60	--	--	3	6	21	19	24	.760
1/2		--	32	--	--	--	4	9	9	13	.749
1/4		--	18	--	--	--	2	3	4	7	.730
						Arithmetic Reasoning					.709
1	--		7	7	8	--	9	63	15	--	.725
1/2	--		6	--	5	--	5	31	7	--	.704
1/4	--		4	--	3	--	3	15	2	--	.668
						Word Knowledge					.767
1	57	--		74	--	1	--	--	--	--	.816
1/2	30	--		37	--	--	--	--	--	--	.801
1/4	16	--		17	--	--	--	--	--	--	.773
						Paragraph Comprehension					.678
1	--	14	85		14	4	--	--	14	--	.700
1/2	--	6	45		7	2	--	--	6	--	.693
1/4	--	1	25		5	--	--	--	2	--	.680
						Numerical Operations					.536
1	--	31	--	15		61	--	23	10	--	.571
1/2	--	15	--	6		36	--	13	--	--	.559
1/4	--	7	--	3		19	--	6	--	--	.541
						Coding Speed					.466
1	5	10	7	14	100		--	--	--	--	.559
1/2	--	1	5	--	63		--	--	--	--	.555
1/4	--	--	2	--	32		--	--	--	--	.548
						Auto Shop					.647
1	--	--	--	--	--	--		15	49	69	.694
1/2	--	--	--	--	--	--		4	23	39	.677
1/4	--	--	--	--	--	--		--	11	22	.653
						Math Knowledge					.631
1	15	75	--	--	9	--	10		12	--	.647
1/2	8	46	--	--	6	--	--		--	--	.627
1/4	5	22	--	--	3	--	--	--	--	--	.597
						Mathematical Comprehension					.656
1	13	19	--	--	--	--	47	24		21	.671
1/2	8	8	--	--	--	--	24	12		11	.657
1/4	5	3	--	--	--	--	12	5		6	.632
						Electronics Information					.671
1	32	--	11	--	--	--	64	4	23		.694
1/2	16	--	6	--	--	--	33	--	12		.683
1/4	8	--	4	--	--	--	18	--	4		.664

improve prediction much when lengthened tended to be dropped, providing increased testing time for those that improve overall prediction of the criterion. Although more subtests are found useful for longer total battery time, the pattern of relative subtest weightings does not vary dramatically as a function of total battery time.

The squared multiple correlation coefficients, in the right-hand column of Table 1, also tell an interesting story. The first R^2 presented for each surrogate criterion is the coefficient obtained using all nine remaining subtests with times equivalent to those of the original battery, which serves as the comparison prediction index for the "optimized" prediction coefficients which fall beneath it. It is readily apparent that although optimizing the battery while retaining the same total time does improve R^2 (increased variance explained), the increases are not particularly large. In terms of percent increase in variance explained averaged across all criteria, there was a 5.4% increase; the largest change was for Coding Speed (CS) (17.8%), but the change for General Science (GS), Arithmetic Reasoning (AR), Math Knowledge (MK), and Mathematical Comprehension (MC) was only 2.5% or less. Therefore, no tremendous advantage accrues from optimizing a battery while maintaining a constant total battery time.

On the other hand, using the battery optimization system, a test battery can be dramatically shortened and maintain as good, or nearly as good, predictive power as the original battery. When used as surrogate criteria, the subtests Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), and Auto Shop (AS) could be predicted as well with a battery four times shorter, approximately 30 min. as opposed to 2 hours of testing. The remaining subtests could be equally well predicted with an optimized battery only half as long as the original.

Discussion

It was demonstrated that a battery of tests may be streamlined by the strategy of optimization. One way that a battery can be streamlined is by increasing the reliability of some subtests by lengthening them, and then doing away with those subtests which provide redundant information. Another way in which a battery can be streamlined is by decreasing the total amount of testing time required, with virtually no loss in predictive power.

In the analyses presented above, the most dramatic finding was that ASVAB subtests can do the same work, in terms of predicting various criteria, in much less time than it is currently administered. For example, when General Science was used as the criterion, the R^2 for the original "nonoptimized" battery was found to be .743. However, when the total test time was cut in half, the R^2 actually increased to .749. This same result was obtained for WK, PC, NO, CS, and AS. Even more importantly, we found that when predicting some criteria (WK, PC, NO, CS, and AS), the ASVAB may do even better in one fourth of the current testing based on an optimized battery.

Before delving into the implications of using an optimization strategy for developing a personnel selection system, several caveats should be considered. First, the information provided above was for illustrative purposes only. The ASVAB subtest scores were selected because they were readily available to the authors. Because actual job performance criteria were not used, the results of the analyses and conclusions must be considered tentative until further analyses using different batteries, criteria, and samples are conducted.

Seco d, the optimization strategy presented above assigns time allocations to subtests based solely on psychometric considerations. Therefore, social and legal implications of a selection program based upon optimization should also be evaluated in terms of the relative fairness of various strategies. It is conceivable that an optimization scheme could serve to magnify differences in the proportion of majority and minority candidates selected and, therefore, run the risk of increasing the probability of charges of test bias. On the other hand, it is entirely possible that optimization could result in a selection system being more fair in the sense that irrelevant, unfair tests might be eliminated. Further research should be done in this regard.

Overall, the approach described above appears to have some potential benefits. One such benefit involves a decrease in the total time involved in the administration of a test battery. This means that the resultant selection system will be more efficient in that more candidates may be assessed per unit time, with no loss in the amount of information derived from the battery. Also, the time saved in terms of test administrators may be allocated to other personnel-related areas. Even if a decrease in total test time is of no concern, a streamlined battery would free up testing time that may be properly allocated to the development of alternative selection indices.

REFERENCES

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. In M. R. Rosenzweig & L. W. Porter (Eds.). Annual Review of Psychology, 30, 477-526. Palo Alto, CA: Annual Reviews, Inc.

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: 1982 Army applicant sample (ARI Technical Report 581). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Skill Requirement Influences on Measurement Method Intercorrelations

Charlotte H. Campbell
Human Resources Research Organization

Michael G. Rumsey
U.S. Army Research Institute for the Behavioral and Social Sciences

The Army is currently engaged in a project, commonly referred to as Project A, to develop a job-based selection and classification system. The project involves the linking of existing and newly developed predictor measures to measures of performance in the Army. The success of the project will depend in no small part on the degree to which the performance measures accurately and comprehensively reflect actual performance of Army jobs. Toward the end of developing a comprehensive performance measurement system, we have developed four different kinds of measures--ratings, administrative measures, hands-on job performance (work sample) measures, and job knowledge measures.

Here we focus on two of the testing methods--hands-on performance tests and job knowledge tests. It has been suggested that, short of measurement in an actual job situation, a hands-on test has the highest fidelity of any type of measure (Vineberg & Taylor, 1978). Yet, probably because of the enormous expense associated with hands-on tests, they are seldom used. Written tests are less costly to administer and in some cases may be as appropriate as, or more appropriate than, hands-on tests. To use an example presented by Vineberg and Taylor (1972), a knowledge test is better suited to assess an automobile driver's knowledge of driving rules and road signs than a hands-on test.

It is of considerable practical interest to know the extent to which the two testing methods are interchangeable. If it could be shown that both methods provide virtually identical information, then one could be eliminated and considerable savings could be achieved. Otherwise, one must consider the possibility that each type of measure provides a unique, valid contribution to an overall assessment of an incumbent's job proficiency and that both are needed to obtain maximum job coverage.

An investigation by Rumsey, Osborn and Ford (1985) used meta-analytic procedured to examine the relationship between hands-on and job knowledge tests. Excluding investigations which used a language-oriented work sample, they found a mean correlation of .57, adjusted for attenuation, between hands-on and job knowledge tests. This correlation suggests some degree of overlap but not total interchangeability.

Are there factors which might substantially moderate the correlation between the two types of measures? Rumsey, et al. (1985) found some evidence

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

that type of work sample had an impact, as correlations for investigations using verbal performance tests tended to exceed those in investigations using motor performance tests. These investigators also found limited support for the proposition that type of occupation influences the correlation obtained. However, much remains to be learned about potential moderating factors.

Vineberg and Taylor (1972) have suggested that the extent to which a job requires skill is an important consideration in examining correlations between knowledge and work sample measures. They noted that skill, unlike knowledge, can only be acquired through practice. Job knowledge tests are presumably best suited to measure knowledge; performance tests are presumably best suited to measure job skills. For those jobs in which task requirements can be reduced to job knowledge, the correspondence between the two types of measures should be high; for those in which skill is an important requirement, the correspondence should be lower.

The effort reported here involved first identifying the skills that are required to perform hands-on tasks that are tested in nine military occupational specialties (MOS) in Project A. Then, the extent to which these requirements moderate correlations between job knowledge and hands-on test scores was determined.

Method

Occupations (MOS). Performance tests and job knowledge tests were developed for nine Army occupations, or Military Occupational Specialties (MOS). These MOS were selected to be as representative of the full set of entry-level MOS as possible, covering the range of job content, Career Management Fields, and ASVAB Aptitude Area prerequisites. The MOS are shown in Table 1.

Task Selection. For each MOS, selection of tasks from the job domain proceeded according to four criteria: the tasks should cover the job content areas, they should be the relatively more important ones, they should permit variability of performance, and they should not be of very low performance frequency.

Test Development. Fifteen tasks in each MOS were selected for performance testing based on such factors as number of cued steps and degree of skill required. Performance tests were developed to score the soldier on whether each step of the task was performed correctly, and to provide standard conditions and instructions for the testing. Multiple-choice format job knowledge tests were also developed for those tasks in each MOS. All tests were pilot-tested, and later field-tested on 114 to 178 soldiers in each MOS. Results from those administrations were used to revise the tests; in some cases, hands-on tests or job knowledge tests were dropped.

Data Collection. Between June and November, 1985, the hands-on and knowledge tests were administered to over 5000 skill level 1 soldiers in the nine MOS, at 14 sites in the U.S. and Europe. (This was Project A's Concurrent Validation phase.) The numbers of soldiers tested in each MOS are shown in Table 1. Job knowledge tests were administered by project staff; actual scoring of the performance tests was done by NCO, trained in scoring procedures by project staff.

Table 1

MOS Selected for Testing and Numbers of Soldiers Tested

MOS		Number Tested
11B	Infantryman	662
13B	Cannon Crewman	586
19E	Tank Crewman	434
31C	Single Channel Radio Operator	303
63B	Light Wheel Vehicle Mechanic	541
64C	Motor Transport Operator	629
71L	Administrative Specialist	481
91A	Medical Specialist	480
95B	Military Police	638

Knowledge/Proficiency Assignments. Three project staff who had been involved in test development and had served as hands-on test managers during the Concurrent Validation testing independently sorted the hands-on steps into one of the three categories: knowledge, simple motor, or complex motor. The level of agreement among the judges was around 80% across the nine MOS; disagreements were resolved by discussing the assignments among the three judges.

Because each performance test score was the percent of steps performed correctly, we classified the tests as K (Knowledge) if at least half of the steps had been sorted into the knowledge category, and as P (Proficiency) if half or more of the steps were in the two proficiency categories. The P tasks were further categorized as P1 (simple motor tasks where manipulation is trivial, easy to perform, and easily learned) if more steps were in the P1 category than in either of the other two categories, or as P2 (complex motor tasks which require more than two trials to perform well) if more steps were in the P2 category than either of the other two categories. Tasks where the number of P1 and P2 steps were the same, or where neither P1 nor P2 outnumbered the K steps, were held out of analyses that compared those two levels of categorization.

Table 2 shows the number of tasks in each MOS that were tested in both the performance mode and the job knowledge mode, and the number of tasks where the performance test was categorized as K, P1, or P2.

Data Analysis. The nine MOS had between 14 and 17 tasks tested in both the job knowledge and performance modes. For each task, the scores used were the percent of steps performed correctly and the percent of items answered correctly. These scores were then correlated by task across the soldiers in each MOS. After the correlations were transformed to Fisher z scores, they were entered into an analysis of variance, with the nine MOS and the knowledge/proficiency categories as independent variables.

Results

Table 3 presents the means and standard deviations of the correlations between performance tests and job knowledge tests for each of the nine MOS;

Table 2

Number of Tasks Tested in Performance and Job Knowledge Modes and Number of Tasks Assigned to Knowledge/Proficiency Categories for Nine MOS

MOS		Tasks	K	P1	P2	Total P[a]
11B	Infantryman	12	2	7	2	10
13B	Cannon Crewman	17	2	8	7	15
19E	Tank Crewman	14	5	7	1	9
31C	Single Channel Radio Operator	15	10	4	0	5
63B	Light Wheel Vehicle Mechanic	15	4	4	6	11
64C	Motor Transport Operator	14	3	5	5	11
71L	Administrative Specialist	12	4	1	7	8
91A	Medical Specialist	15	6	6	1	9
95B	Military Police	16	8	4	3	8

[a]Includes tasks not clearly P1 or P2; see text.

the statistics are also shown for the groupings of tasks based on knowledge/proficiency category assignments. (The correlations had been transformed, using the Fisher z transformation, before calculating the summary statistics; the results shown in Table 3, however, have been transformed back to Pearson correlations.) In eight of the MOS, the individual task correlations ranged from about .00 to .40; in one MOS, the highest correlation was .19. (Task correlations tend to be substantially lower than correlations for entire jobs; hence, the level of these correlations cannot be meaningfully compared with earlier findings.) With the large number of soldiers tested in each MOS, even small correlations (around .08) are significant at the .05 level. Over two-thirds of the correlations in every MOS were significant at that level.

Two analyses of variance were calculated, using the transformed correlations (Fisher z) as the dependent variable. In the first ANOVA, the nine MOS and the two knowledge/proficiency categories (K and P) were the independent variables. The second ANOVA likewise used MOS, and also the three levels of the knowledge/proficiency categorization (with two levels of proficiency - simple motor (P1) and complex motor skills (P2), as the independent variables. Both ANOVA results are summarized in Table 4.

In both analyses, the main effect for MOS was nonsignificant, and the interaction terms were not significant. In both analyses, the knowledge/proficiency term was significant. Where knowledge/proficiency was considered on only two levels, the difference favored the K tasks, where the performance test had been categorized as predominantly knowledge. In the second analysis, where there were three groups of tasks - knowledge (K), simple motor (P1), and complex motor (P2) - comparisons of the means of those groups revealed that only the difference between K tasks and P1 tasks was significant at the .01 level ($F = 14.33$, $df = 2,95$); K tasks and P2 tasks differed slightly ($F = 6.68$, $df = 2,95$, $p < .10$), as did K tasks and the combined group of P1 tasks and P2 tasks ($F = 7.581$, $df = 3,95$, $p < .10$). The difference between P1 and P2 tasks was not one bit significant.

Table 3

Means and Standard Deviations of Performance x Job Knowledge Test Correlations by Knowledge/Proficiency Category for Nine MOS

MOS			Tasks	K	P1	P2	Total P
11B	Infantryman	N	12	2	7	2	10
		Mean	.17	.26	.18	.09	.15
		S.D.	.37	.14	.16	.02	.14
13B	Cannon Crewman	N	17	2	8	7	15
		Mean	.17	.20	.16	.17	.16
		S.D.	.11	.07	.11	.13	.12
19E	Tank Crewman	N	14	5	7	1	9
		Mean	.14	.23	.09	.12	.10
		S.D.	.13	.19	.07	-	.06
31C	Single Channel Radio Operator	N	15	10	4	0	5
		Mean	.20	.22	.15	-	.15
		S.D.	.14	.17	.03	-	.03
63B	Light Wheel Vehicle Mechanic	N	15	4	4	6	11
		Mean	.10	.10	.07	.10	.10
		S.D.	.04	.04	.02	.03	.04
64C	Motor Transport Operator	N	14	3	5	5	11
		Mean	.15	.26	.11	.09	.12
		S.D.	.12	.20	.09	.05	.09
71L	Administrative Specialist	N	12	4	1	7	8
		Mean	24	.30	.16	.20	.20
		S.D.	11	.13	-	.09	.09
91A	Medical Specialist	N	15	6	6	1	9
		Mean	.17	.17	.15	.33	.17
		S.D.	.13	.18	.08	-	.09
95B	Military Police	N	16	8	4	3	8
		Mean	.15	.18	.10	.10	.11
		S.D.	.11	.11	.06	.17	.10
	Across MOS	N	130	44	46	32	86
		Mean	.16	.21	.13	.14	.14
		S.D.	.12	.15	.10	.11	.10

Table 4

Analysis of Variance Summary Tables for MOS x Knowledge/Proficiency

MOS x Knowledge/Proficiency

	SOURCE	SS	df	MS	F-Ratio	F	p
[1]	MOS	.156	8	.020	[1/4]	1.45	<.25
[2]	K/P	.137	1	.137	[2/3]	18.08	<.01
[3]	MOS x K/P	.061	8	.008	[3/4]	.57	NS
[4]	Within cell	1.499	112	.013			

MOS x Knowledge/Simple Motor/Complex Motor

	SOURCE	SS	df	MS	F-Ratio	F	p
[1]	MOS	.142	8	.018	[1/4]	1.20	NS
[2]	K/P1/P2	.108	2	.054	[2/3]	4.90	<.05
[3]	MOS x K/P1/P2	.161	15[a]	.011	[3/4]	.73	NS
[4]	Within cell	1.388	95	.015			

[a] Reduced by 1 df for missing cell estimation.

Discussion

There is fairly clear evidence here that the differentiation between knowledge requirements and proficiency requirements on hands-on performance tests explains some of the variability in correlations between the two modes of testing. When the steps required on the performance tests are primarily knowledge mediated, and are demonstrations of the acquisition of task knowledge, then the correlations with written tests of the tasks are higher than when most of the performance test steps require demonstration of psychomotor skill, however simple.

Further analyses, already underway, will involve meta-analysis of the obtained correlations, and an examination of the knowledge/proficiency distinction as a possible moderator variable.

References

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985). Comparing work sample and job knowledge measures. Paper presented at the annual conference of The American Psychological Association, Los Angeles.

Vineberg, R. & Taylor, E. N. (1972). Performance in four army jobs by men at different aptitude (AFQT) levels: 4. Relationships between performance criteria (Technical Report 72-23, pp. 17, 19). Alexandria, VA: Human Resources Research Organization.

Vineberg, R., & Taylor, E. N., (1978). Alternatives to performance testing: Tests of task knowledge and ratings (Professional Paper 6-78). Alexandria, VA: Human Resources Research Organization.

Post Differences in Hands-On Task Tests

R. Gene Hoffman
Human Resources Research Organization

One of the major efforts for the U.S. Army's Selection and Classification Project (Project A) has been the development of hands-on performance measures. The effort required preparation of tests to cover approximately 15 tasks for soldiers in nine different job specialties (MOS). Because of equipment differences within certain MOS, it was necessary to create alternate versions of some tests. Thus, 103 different task tests were prepared. Eleven tests were used in more than one MOS with the number of tests per MOS ranging from 14 to 27. As part of the concurrent validation data collection effort, these tests were administered during 1985 to approximately 500 to 600 soldiers per MOS. In order to collect that volume of data, test sites included 13 different Army posts in the United States plus European test sites. At the European sites, approximately 120 soldiers for each MOS were tested. At the CONUS sites, the numbers of soldiers per MOS per site ranged from 9 to 110 with typical numbers being near 30, near 45 or near 60 because of scheduling requirements. The tests were administered in blocks of two to four tasks per test station with typically one NCO in the respective MOS at each site handling test administration for all soldiers at any given station.

Given these "road show" requirements for data collection, considerable effort was made to standardize the hands-on testing procedures. These efforts included attention to test set-up and scoring instructions and to the training of test administrators. Prior to concurrent data collection, test procedures were pilot tested on a small sample of soldiers using four to five test administrators and then field tested on approximately 150 soldiers. Administrator training included five phases: (1) presentation of general testing principles, (2) familiarization with individual test station requirements, (3) practice, (4) review by contractor personnel prior to data collection, and (5) monitoring by contractor personnel during data collection. Further details concerning test construction and administration are presented in Campbell et al., (1985) and Campbell (in preparation).

Given that hands-on testing has a history of being susceptible to scorer differences (e.g., Maier, 1983), this paper examines differences between posts in hands-on test scores and the extent to which any such post differences are not "real" differences, but are, in some way, artifacts of the measurement process. Thus, analyses examined alternative sources of variance in hands-on test scores that could account for any mean differences between posts. Candidate measures for explaining differences available in the Project A data set include: (1) written tests, (2) supervisor and peer ratings of performance, (3) practice, (4) time in service, and (5) ability. Post effects were estimated after variance due to these measures was removed from the hands-on tests (using hierarchical multiple regression) and compared with post effects prior to any adjustment.

Analysis

Analyses were conducted for every hands-on test in all nine MOS. No adjustment was made for tests appearing in more than one MOS. That is, repeated tests were treated as separate observations. Thus, there were 147 observations of post differences where an observation is a test/MOS combination. The first series of analyses estimated unadjusted post effects (percent of variance in hands-on score accounted for by post alone) and post effects adjusted for written test scores (except, obviously, those tasks tested only in the hands-on mode), task ratings by peers and by supervisors, overall performance ratings by peers and supervisors, practice (composite of self ratings of recency and frequency of task performance), time in service (test date minus entry date), and general ability (AFQT). In conducting these analyses, significant reductions in sample sizes between post only and adjusted post analyses were observed for all MOS. The reductions were most attributable to missing ratings. Therefore, an alternative or "reduced" adjustment model was also examined in which ratings were excluded. Thus, for each of the 147 tasks, three different R^2s were calculated between post and hands-on scores: (1) an unadjusted "post alone" R^2, (2) an adjusted R^2 for post after all other variables in the "full model" were controlled, and (3) an adjusted R^2 for post after all other variables in the "reduced (ratings excluded) model" were controlled. Adjusted R^2s were calculated as the increase in R^2 when post was added after all control variables in a hierarchical multiple regression predicting hands-on score. Mean sample sizes for these analyses were 500.21 for post alone, 164.01 for the "full model" (i.e., all variables) and 341.77 for the "reduced model."

The R^2s between post alone and hands-on scores estimate the extent of between post differences in hands-on scores. These were compared to the R^2s for post and hands-on scores after variance due to the other variables in the full and reduced models were controlled. Differences in variance accounted for by post (i.e., differences in R^2s) were calculated as indices of the bias resulting from post differences. Thus, two bias indices for each hands-on test resulted from these analyses: a "full model" bias and a "reduced model" bias. The term bias has been used in response to the question: "Would standardizing hands-on scores by post bias those scores?" Positive values for these bias indices would suggest that any post differences are to some extent real and that standardizing would introduce bias. On the other hand, near zero values suggest that post differences are unrelated to other measurements of performance, therefore may reflect measurement error, and that standardization may be justified.

The above analyses were conducted on a task by task basis. From these analyses it is not possible to tell whether the "post" effects are actually at the post level or are more correctly attributable to scorer differences. Two approaches were used to address this question, neither of which is definitive. First, if "post" effects (within an MOS) were operating consistently for all tasks within an MOS (e.g., motivational differences between posts), then it should be possible to account for post variance in any one task by removing variance associated with the hands-on test scores for other tasks within each MOS. Thus for each task, an adjusted R^2 for post effects were examined after variance associated with other MOS tasks was removed. An "other tasks" bias index was constructed as the difference

between post effects alone and this "other tasks" adjusted R^2. If this index is near zero, the "post" effects are task specific and not consistent across tasks within an MOS.

A second way to partially dissect the task by task post effects is to examine scorer-within-post variance capitalizing on the instances where two or more scorers scored the same test at the same post either by general design (i.e., duplicate equipment and test stations in the test plan) or by local variation (i.e., an early finishing scorer helping at another station).

The series of analyses examining post effects controlling for performance on other hands-on tests occurred some time after the first, and in that interval two 91A tracked tests were merged; therefore 146 separate tasks were analyzed. Again a "bias" variable was calculated as the difference between post effects alone and adjusted post effects.

Results

Results for these analyses are summarized in Table 1 below. All data points were either R^2s (for the Post Only analyses), increases in R^2s (for the full, reduced and other task model analyses), or differences between R^2s (for the bias variables). Thus, table entries are the means, standard deviations, minimums and maximums for these R^2s across the 147 tasks.

Uncorrected post differences account for an average of 19% of the variance in hands-on test scores, indicating the presence of post differences in hands-on scores. Post effects range from 2% to 50%. For only 36 of the 147 tasks is the post effect less than 10% of the hands-on variance. Furthermore, there is no evidence that post differences can be consistently attributed to written test scores, practice, ratings, ability, or time in service. Mean bias from the full and reduced model analyses are both very near zero suggesting that removing post differences by standardization would not bias the hands-on scores.

Table 1

Hands-On Test Variance (R^2) Associated With Post
With and Without Controls and Associated Adjustment Bias

		Variance Associated with Post				Standardization Bias		
		Post Only Model	Full Model	Reduced Model	Other Tasks Model	Full Model Bias	Reduced Model Bias	Other Task Model Bias
Mean	R^2	0.19	0.22	0.18	0.12	-0.03	0.01	0.07
S.D.	R^2	0.11	0.11	0.11	0.08	0.08	0.05	0.06
Min.	R^2	0.02	0.01	0.01	0.01	-0.25	-0.24	-0.04
Max.	R^2	0.50	0.52	0.52	0.34	0.24	0.23	0.33

Results for the "other tasks" model are presented in Table 1. Bias as estimated by this model is somewhat larger than the others and suggests that to some extent post differences for any given task are related to post differences for other tasks. However, certainly not all of the task level post effects are explained.

Table 2 indicates that the 147 tasks are rather homogeneous with regard to reduced model bias. For the 147 tasks, 114 reduced model bias indices are between -.05 and .05. The other bias indices are similarly homogeneous. Thus, the post effects that are present remain so after attempts to explain them are considered and that trend is consistent across all tasks.

Table 2

Distribution of Reduced Model Bias Across 147 Hands-On Tests

Reduced Model Bias	Frequency	Percent
-0.30>	0	.00
-0.25>	1	.68
-0.20>	0	.00
-0.15>	4	2.72
-0.10>	6	4.08
-0.05>	56	38.10
-0.00>	58	39.46
0.05>	18	12.24
0.10>	1	.68
0.15>	2	1.36
0.20>	1	.68
0.25		

The final analysis made use of the duplication of scorers for some tasks at some posts. Because this duplication was not systematically planned, some instances of duplication of scorers were due to a scorer at one post scoring only one or two soldiers. Such cases are not very illuminating. To avoid them, only tasks for which degrees of freedom for scorers-within-post was at least 5 were examined. Forty tasks met this criterion (degrees of freedom ranged from 5 to 23). For these tasks, the mean scorers-within-post effect accounted for 4.6% of the hands-on variance. This number probably underestimates the size of the scorer effect because post effects were still confounded by scorer effects. That is, for all but a few tasks in this analysis, several posts were represented by only one scorer. For the thirteen tasks with 10 or more degrees of freedom for scorers within post (and fewer posts with only one scorer), 6.4% of the hands-on variance is associated with scorer differences. While it is not possible to totally disentangle post versus scorer differences, it is probably safe to conclude that there were consistent scorer differences, and that some of the differences among posts are attributable to scorer differences.

These analyses unfortunately are like trying to show that something does not exist when we can look in only so many places. That is, we are trying to

rule out alternative explanations for the post effects while we are limited in the availability of ways to look. Given the evidence, unwanted post effects at the task level can not be ruled out, and the standardization of hands-on test means by post appears justified.

One may wonder what might be the negative consequences if the decision to standardize by post is incorrect. The most damaging consequence would be an introduction of error leading to a reduction in the predictability of hands-on measures. To shed some light on this possibility, the predictability of standardized and unstandardized hands-on test scores were compared using the reduced model variables (i.e. R^2s for predicting hands-on tests from written tests, experience, practice, time, and ability). Across the 147, the average difference between the two R^2 is .02 with the standardized hands-on scores being slightly less predictable. The standard deviation of the difference across the 147 tasks is .05. Thus, across the tasks standardizing has little effect one way or the other on the predictability of the hands-on scores.

Summary

In summary, post effects on hands-on scores were present and no alternative explanation of those effects was found. This leaves the implication that the post differences reflect error in the measurement process. Second, the post effects seem to be operating idiosyncratically at the task level, i.e., as the post or scorer effects unique to each task, rather than as the post level effects consistent for all tasks in an MOS. Third, while it is not possible to totally disentangle post and scorer, some of the between post differences are probably due to scorer differences. Fourth, post differences should be controlled in further statistical analyses of hands-on test scores. And finally, even if this conclusion is incorrect, statistical corrected by standardizing by post will not have a grave impact on the predictability of the hands-on scores.

References

Campbell, C. H. (1986). Developing basic criterion scores for hands-on tests, job knowledge tests, and task rating scales (In preparation). Alexandria, VA: Human Resources Research Organization.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., and Edwards, D. C. (1985). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H. (1983). Using job performance tests as criteria for validating qualifications standards (Memorandum CNA 83-3123.09). Alexandria, VA: Center for Naval Analyses.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

The Electronic Clipboard: A Central Requirement for Effective Automation of Training Management in Military Units

Nancy K. Atwood and Jack H. Hiller
U.S. Army Research Institute
and
Jane Herman
Perceptronics

Managing home station training of military units is a complex and difficult activity. A typical Army battalion has hundreds of collective tasks on which subordinate companies/batteries, platoons/sections, squads, crews and teams must be trained. Within a battalion, there are hundreds of soldiers with numerous military occupational specialities (MOS's). Each MOS has hundreds of individual job tasks on which soldiers must be trained. Furthermore, units are continually undergoing changes in personnel so that the training needs of units and their soldiers do not remain stable. And, with the ongoing force modernization program, units are acquiring sophisticated high technology equipment which dictates new training requirements calling for higher degrees of soldier skill than ever before.

Planning an effective and efficient training program requires complete and up-to-date evaluation information on the performance of units and individual soldiers in order to identify the collective and individual skills on which units and soldiers need to train. Once training requirements are identified, intricate coordination of leaders (trainers and evaluators), soldiers, and resources (such as ranges, equipment, fuel, and ammunition) is required to ensure that the right people are brought together at the right time with the right equipment.

Training management has traditionally been accomplished using a manual, paper-based system with information that is often incomplete and sometimes unreliable. Recently, automated approaches to training management have received considerable attention in the military training community. For example, the Army Development and Employment Agency (ADEA) is currently developing the Integrated Training Management System (ITMS) using the 9th Infantry Division at Fort Lewis, WA as a testbed. The system is being specifically designed for training management purposes at the division down to company level and will integrate information on training, personnel and logistics via compatible data bases accessed with a relational data base management system.

One of the critical obstacles encountered in designing a workable automated system for training management is the volume of data required. Not only is a large amount of information on training history, personnel background, and available resources required, but it must be accurate, complete, and recent. This

need is paramount when it comes to evaluation data which is central to the problem of determining training requirements and for which there are hundreds of collective and individual tasks on which evaluation data are needed. This data requirement places a heavy burden on personnel, not only to collect the information, but to see that it is entered without error on the computer system. The data collection and entry demands simply cannot be met without dedicated personnel, a scarce or non-existent resource in today's already overburdened units.

This paper describes the development and tryout of the Electronic Clipboard (EC), an innovative concept designed to facilitate data storage and entry to a larger automated training management system. First, the features of the EC are described. Second, the design and results of a field test are presented. Third, conclusion about the viability of the EC are put forth. Finally, lessons learned for technology development and transfer more generally are considered.

Features of the Electronic Clipboard

The EC is a hand-held, field-portable, battery-operated computer unit (see Figure 1). The device serves four major functions; it: (a) receives training guidance and evaluation checklists selected from menus stored in a base-station computer; (b) stores and displays identifying information for units, soldiers, and evaluators; (c) receives and automatically stores ratings for items on evaluation checklists; and (d) sends scores automatically into the unit's training data base stored in the base-station computer for subsequent summary and analysis.

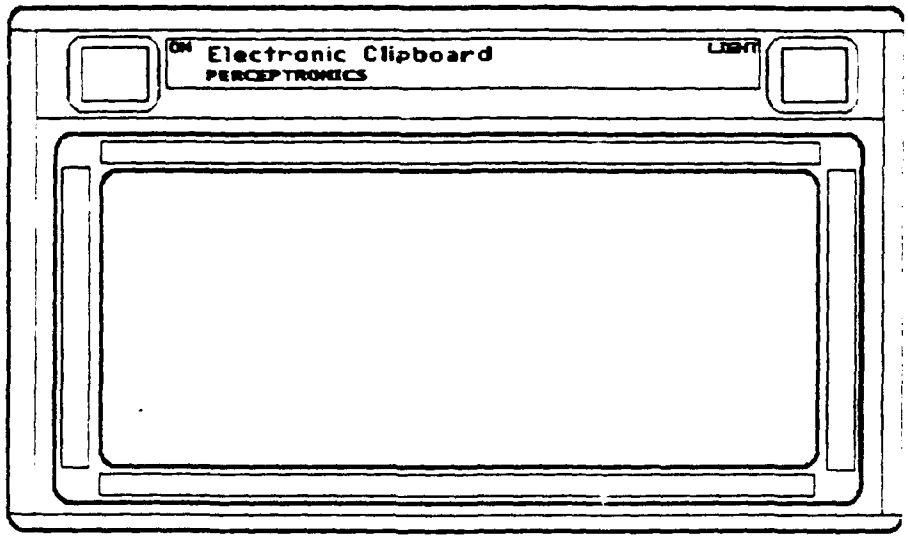


Figure 1

Field Test

A field test of the EC was con ucted at the U.S. Army Armor Center at Fort Knox, Kentucky. The test was designed to assess user's opinions about hardware and software capabilities and to collect their recommendations for improvement.

Information was gathered using two approaches. First, feedback was obtained from 30 instructors/evaluators from the Basic Non-Commissioned Officer Course (BNCOC) and five other USARMC Departments (Training Group, Command and Staff, Weapons, Maintenance and the NCO Academy). Respondents received a tutorial and a hands-on orientation to the system before completing a structured questionnaire. Open-ended comments were also gathered during this process. Second, the BNCOC evaluators also used the EC during Gunnery and Land Navigation exercises under both day and night conditions. They were asked to report their impressions of the strengths and weaknesses of the current design.

Questionnaire responses were overwhelmingly favorable. Of the 45 questions on general functions and specific features of the EC, the vast majority of respondents expressed favorable views on all but two items. Concerns expressed centered on perceptions of potential difficulty in using the EC in a field environment. Many respondents felt the need for a handle or some type of grip, a case with a rougher texture for easier handling, and greater ruggedization.

BNCOC instructors who used the EC for conducting and evaluating training made the following types of recommendations for improvement: provide faster scroll, use non-glare surface on display, provide positive feedback on keyboard (so user can tell when a key has been pressed), develop more ruggedized package, add a numeric keypad, consider making the unit smaller and lighter, and make the unit easier to hold (with a handle or holster and a rougher surface).

Conclusions about the EC

The field test demonstrated the utility of the Electronic Clipboard concept; however, the need for a second generation design was clear. Essential features for the next generation design include: a higher contrast display, a more reliable touch pad, and a ruggedized case.

The TRADOC community followed the field test of the Electronic Clipboard with considerable interest in its potential use, not only at home station, but also at advanced training facilities such as the National Training Center (NTC) and the Joint Readiness Training Center (JRTC). For example, observer/controllers (O/C's) at the NTC observe force-on-force engagements and live fire exercises and must provide quick-turnaround battle summaries and feedback to units as part of the

The EC uses state-of-the-art technology including a liquid crystal display. The display is back lit using an electroluminescent panel for nighttime use. In place of a fixed keyboard, the EC has an infrared-based touch panel for data entry and computer control. The touch areas are configured into two arrays of 4 by 4 "touch keys". Total processing and storage capacity of the EC is 256K bytes which is as much memory as most 16-bit micro-computers.

The unit is powered by rechargeable nickel/cadmium batteries that can support 8 hours of continuous daytime operation within an 80 hour period and 4 hours of continuous nightime use over and 80 hour period. Batteries can be charged to full capacity in eight hours using a field charger or by attaching the EC to the base station computer.

The current version of the EC is compatible with an IBM PC XT as the base station computer. The EC and the base station interface through a full duplex asynchronous commmunication line and an RS-232 connector.

The EC software supports the following functions: (1) receipt of databases from the base station computer including personnel, training drills, and evaluation checklists; (2) diagnosis of the operability of selected system components; (3) interactive identification of evaluators, students, training drills, and evaluation checklists; (4) presentation of training guidance; (5) interactive evaluation of checklist items; (6) maintenance of all scoring information collected in the field; and (7) selected utility functions, including a clock, calendar, stopwatch and low battery power indicator. Figure 2 illustrates a sample display.

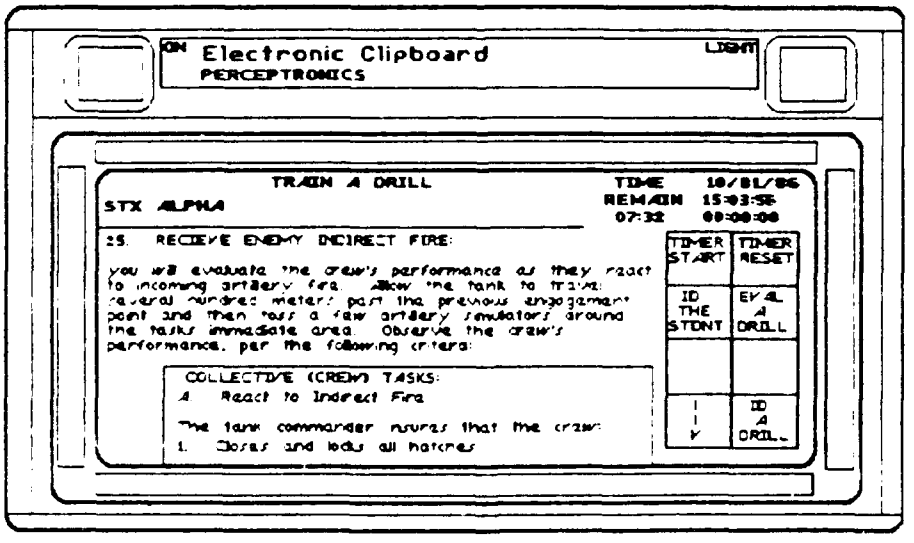


Figure 2

After Action Review (AAR) process. The Electronic Clipboard could serve as a tool for O/C's to facilitate their information gathering and to provide quick summaries of battle events to them in the field. The EC would have the secondary benefit of standardizing the information collected by the O/C's for subsequent use in research and generating lessons learned.

At the request of the Commanding General of Fort Irwin, the Army Research Institute is developing a second generation EC for try-out at the National Training Center. As with the initial prototype, this work will be conducted under contract to Perceptronics. While the requirements have not yet been fully specified, additional features under consideration to meet the particular needs of the NTC training environment include: (1) a mount to the jeep dashboard with attachment to the jeep power source; (2) an audiotape mounted to the EC for use by the O/C's in recording verbal battle commentary; (3) communication via radio from the field to the computer center and back; (4) a free format message capability; (5) applications programs held in the EC to allow summary and analysis in the field; and (6) menu-driven software identifying critical tasks, conditions, and standards for the six most common NTC missions (Defend a Battle Position, Defend in Sector, Movement to Contact, Deliberate Attack, Night Attack, and Hasty Attack) organized by operating system (Command & Control, Maneuver, Fire Support, Mobility/Counter-Mobility, Intelligence, Air Defense Artillery, and Combat Service Support) and echelon (Battalion, Company, Platoon).

Lessons Learned for Technology Development and Transfer

More generally, the EC development effort yielded a number of lessons learned for technology development and transfer. First, it demonstrated the value of rapid prototype development. Hardware and software development was aimed at producing a device that could be "touched, handled, and kicked" as soon as possible in the program rather than creating a fully finished, field-ready device with all of the possible optimization in hardware and software design. This strategy allowed for early evaluation by users of a product rather than a paper specification.

The second lesson highlighted the need to be sensitive to and accommodate the rapid changes occurring in electronics technology. For example, during the course of prototype development massive and significant improvements occurred in LCD technology. These advances impacted not only on the choice of display, but also on the display drivers and the case size. Thus, any design (whether mechanical, electrical, or software) may be subject to significant changes based on advances in technology; sufficient flexibility must be incorporated into the initial design to accommodate anticipated advances.

A third lesson concerned the need for flexible software. An

early design decision was made to use flexible word processing and post-processing programs to create evaluation checklists. This feature proved critical in allowing for easy adaptation to unanticipated changes in the format of data bases required for the field test at the Armor School. Additional flexibility would also have been desirable to accommodate other changes in requirements (i.e., the need for three pass scoring) that arose as part of the field test. A desirable step in this direction would be to incorporate a capability to download applications programs instead of just data bases.

A fourth lesson brought home the point that it is not always cost effective to make all interim hardware and software products deliverable under the contract. For example, during the validation testing of the EC prior to the field test it was determined that design changes were necessary to accommodate a power saver circuit that was needed in order to meet the power requirements of the LCD display. However, it was neither efficient or cost-effective to try to retrofit the existing units with the new circuitry. The practical approach was to insure that all required changes were incorporated into the subsequent field units. In such cases, it is not cost-effective to require delivery of development models since they are suboptimal compared to the final units and contain expensive parts which can be re-used in building final production models.

In sum, the EC development effort illustrated the viability of a field portable computer device as an important component of an automated training management system. Certain hardware design issues remain to be addressed in the development of the second generation EC. However, the critical importance of the software design should not be underestimated. The ultimate utility of the EC will be determined by the extent to which software can be designed to meet the needs of military trainers and evaluators.

AUTOMATION OF ARMY UNIT TRAINING

Dwight J. Goehring
U.S. Army Research Institute Field Unit
Presidio of Monterey, CA 93944-5011

One of the requirements for achieving and maintaining combat readiness is effective management of unit training. In order to enhance unit readiness, training managers must determine individual training needs of soldiers and collective training needs of units, for planning and scheduling, for resourcing and for conducting training and evaluation. These tasks are often extremely difficult to perform.

The efficiency and effectiveness of training management can be improved through automation. The Army Development and Employment Agency at Fort Lewis, Washington, under tasking from the Deputy Chief of Staff for Operations and Plans, is developing a prototype Integrated Training Management System (ITMS) to evaluate the effects of training management automation. ITMS deals with training management from Division through Battalion echelons in long-range, short-range and near-term planning contexts. Long-range planning is projected 18 months into the future. Short-range planning addresses the immediate future, typically three to six months forward. Near-term training management focuses on preparing weekly training schedules three to six weeks in advance.

The current work of ARI in support of the ITMS is focused on the determination of training requirements within the context of the management of training at the Battalion level in the short-term. The analysis conducted by ADEA (June 1985; September 1985) concludes that determination of training requirements consists of two primary processes. First, the training manager develops the Mission Essential Task Requirements (METR) through analysis of missions and tasks that appear on the unit Mission Essential Task List (METL). Next, the training manager formulates the unit Mission Training Plan (MTP) through METR prioritization, reconciliation with the long-range calendar and elaboration of the missions and tasks on the METR. Figure 1 shows a summary of the short-range management process (ADEA, December 1985).

ARI's research on the problem of formulating training requirements is delimited, at the request of ADEA, to the two above described processes. It is assumed that the METL has been developed for the unit (its formulation is based in part upon classified documents and, therefore, is omitted from ITMS which contains no classified information). Further, the resourcing and detailed scheduling of the MTP missions, STXs, drills, collective and individual tasks on the short-range calendar are beyond the scope of the current research effort.

This paper presents a conceptual analysis of the processes involved in development of the METR and MTP for a battalion. A companion document (Goehring, in preparation) lays out a functional design for a computer program module for supporting these two processes in the context of ITMS and reports the results of the development of a computer program which evaluates the feasibility of the approach.

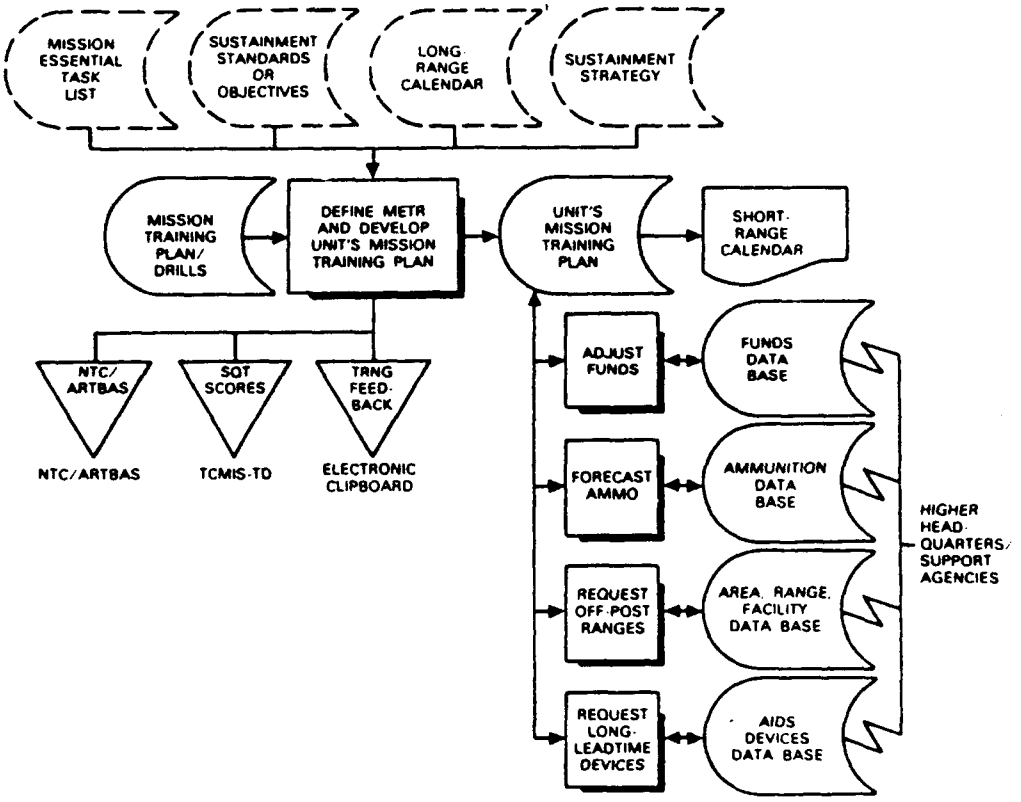


Figure 1. ITMS Short-Range Management Process Summary (ADEA).

Processes in Determining Training Requirements

How training requirements are determined is a complex process. Many different sources and a range of different types of information are combined by the training manager in deciding the activities on which the unit will train. The way information is combined and the logic used in making decisions not only is implicit but also certainly differs between training managers.

Because of these complications, this paper takes a prescriptive approach over a strictly descriptive one. We emphasize procedures which can be automated but ITMS issues and constraints are intentionally excluded here to provide as broad a perspective on the problem as possible.

The analysis of the determination of training requirements was divided into two processes paralleling the findings of the analysis conducted by ADEA (June, 1985; September, 1985). The first process is the development of the METR from the METL. The second process is the development of the unit MTP from the METR. Figure 2 shows this two process sequence with the types of data required for each process. The key in both processes is user judgment.

Development of the METR

While the METL of a unit constitutes all of the mission/tasks, collective as well as individual, which must be performed in the execution of its mission, the METR is the subset of missions/tasks which the training manager determines require training in the short-range.

The training manager systematically and sequentially analyses each of the missions/tasks on the METL, using four related subprocesses. A description of each subprocess follows in the order which their natural interdependencies dictate.

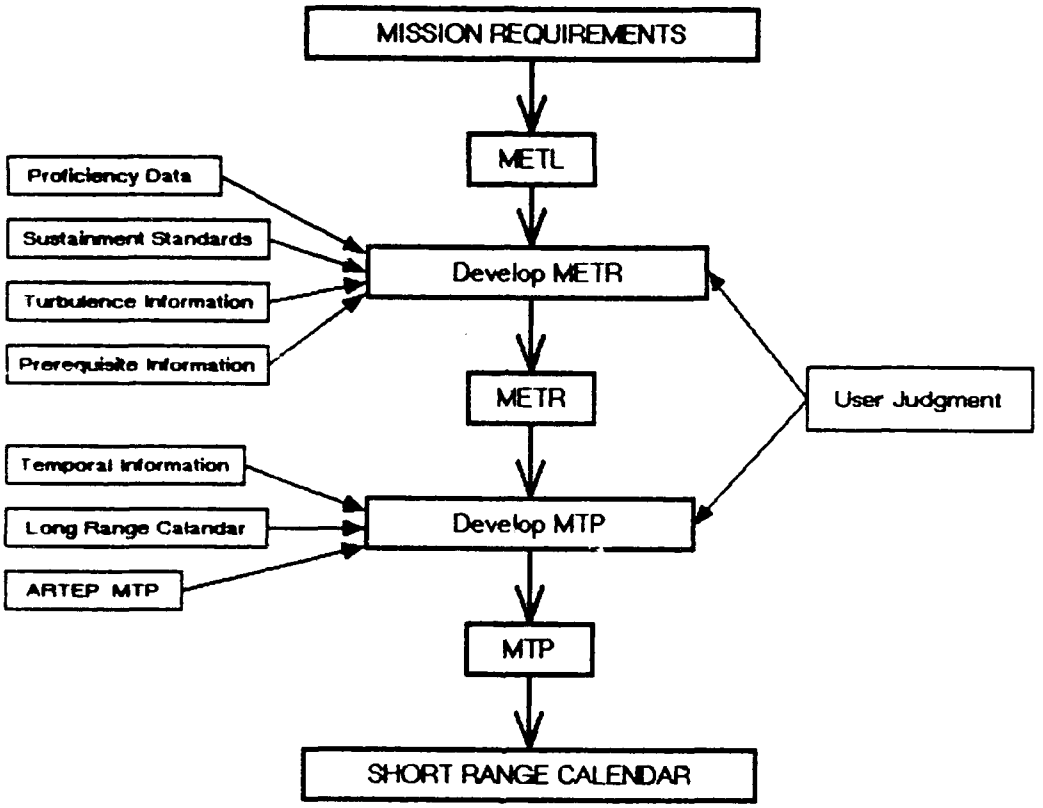


Figure 2. Overview of Training Requirements Function.

Proficiency. The training manager determines if proficiency was achieved for the mission/task in question when it was last trained. If no record or memory exists that it was ever trained, the training manager should identify the mission/task as needing training. If proficiency was not achieved when the last training occurred, the mission/task will need to be retrained.

Proficiency Decay. Where proficiency was achieved during training for a mission/task, the training manager will consider how recently the training occurred. Because collective as well as individual skills deteriorate with disuse as time passes, at some point retraining will be required. The rate of skill decay varies by specific mission and task.

Personnel Turbulence. Although the proficiency of a mission/task may not have decayed based upon sustainment standards and the passage of time, deterioration of proficiency may have occurred if personnel turbulence has been excessive. Personnel turbulence is the phenomenon of movement of individuals into, out of and within Army units.

Analysis of Prerequisite Tasks. Our hypothetical training manager will want to complete the METR by augmenting the missions/tasks identified as requiring training with any prerequisite or supporting tasks which themselves require training. Each identified prerequisite or supporting task needs to be analyzed in terms of its proficiency when last trained, its proficiency decay and the effects of personnel turbulence upon its proficiency. If deficient proficiency is indicated for a task, it may be added to the METR list.

Development of the MTP

The MTP is a subset of the METR mission/tasks which are intended to be trained by the unit within the short-range planning horizon. The process of developing the MTP occurs as the training manager performs four procedures on the METR mission/task list. Each is described below in the order of completion.

Prioritize METR items. The training manager arranges all of the missions/tasks on the METR in order (ties are allowed) according to the importance to train the item within the current planning horizon. The ordering serves as a basis for resolving conflicts among competing tasks in the preliminary assignment of start times on the long-range calendar.

Retrieval of temporal information. Next, the training manager needs information about the duration and any special temporal contraints of the METR tasks/missions. An example of a temporal constraint would be training during winter months for cold-weather survival skills.

Long-range Calendar Reconciliation. The goal of this subprocess is to assign tentative start dates to the missions/tasks which are to be trained during the short-range planning horizon. The training manager uses both the mission/task priorities assigned earlier and the long-range calendar itself. The highest priority missions/tasks are assigned start dates first, followed by missions/tasks of lower priority.

Elaboration. During this subprocess each mission and task which has been assigned a tentative start date within the short-range planning horizon is further specified by the training manager in terms of specific ARTEP Mission Training Plan tasks, STXs, drills and individual tasks to be trained. Both training documents and the training manager's own experience and preferences will affect the specific activities on which the unit will train.

The output of priorization, retrieval of the temporal information, long-range calendar reconciliation and elaboration subprocesses is the unit MTP.

A Decision Support System Approach

The needs (1) for intensive training manager interaction, (2) for

extensive data base access, and (3) for overall system flexibility make the automation of training requirements determination an ideal candidate for the application of Decision Support System (DSS) technology. A DSS is a computer-based vehicle which accommodates decision making for relatively unstructured problems in a flexible and highly user-interactive manner (House, 1983). A DSS is generally oriented toward high level management. It is user initiated and controlled and can support a variety of decision making styles. From another perspective, a DSS can be viewed as a collection of software tools, engineered for a general problem area, which the user then applies to the particular problem at hand in a non-predetermined manner, specific both to the problem and the style of the user.

In contrast to a Management Information System (MIS), a DSS has a less structured information flow, less emphasis on the production of reports as an end goal and greater support for actions resulting from the decisions made by the user. Thus, while a DSS incorporates many of the same functions as an MIS, it exceeds an MIS in flexibility and scope.

The training management problem of identifying training requirements and developing training plans is well suited to a DSS approach. The identification of missions/tasks for inclusion in the METR and the subsequent development of the unit MTP are high level management processes, which can be accomplished according to different styles, but which cannot be developed by completely prespecified algorithms because of the need for user input and judgment. As a result, user interaction and system flexibility are vital to automating the processes involved in METR and MTP development.

User-Centered Computer Environment

The features of the preferred computer environment in which the training manager utilizes the training requirements system deserve description. Characteristics of flexibility of system operation and high information bandwidth interaction are hallmarks of the DSS approach. (A slow teletype terminal exemplifies a low bandwidth system.) Several additional characteristics of the user interface capability are desirable in the training requirements application.

Users of interactive computer systems often need to refer to different sources of data or information simultaneously. An example in training requirements is where the training manager wants to see the training data and the guidance of the higher echelon commander while deciding upon and inputting the training priority of a specific task. One approach to accommodating this need is to employ windows in a multitasking environment. Windows are multiple independent display areas on a cathode ray tube. They can be manipulated independently of one another, each presenting the user with different information from within the database at the same time. New windows can be added or old ones terminated at the discretion of the user. Windows can also display a changeable small part of a large file, using a scrolling capability.

Minimizing the amount of keyboard input is a general specification goal. Often items to be input (such as task identifications or unit designations) already exist within a database and can be displayed to the user for selection by a minimal keyboard input, or by use of a pointing device, such as a mouse. Closely associated with the concept of enabling the user to

choose from item lists rather than having to input detailed data is the approach of enabling the user to invoke various system actions by selection from menus. Menus may be interactive windows that only appear when needed.

It is also highly desirable to have an on-line help facility available to the user. Accessibility to such a facility should be possible no matter where the user is in utilizing the system. In its simplest form and as a minimum the help facility should have a simple hierarchy of descriptions of primary processes and successively subordinate subprocesses. Thus, it becomes the user's responsibility to find the relevant information. An alternative is a help facility which is context sensitive. Embedded training could also be included in the help facility, greatly assisting new users.

The training requirements system needs a range of run modes. At one extreme the user will be able to start the system which will run with minimal or no user intervention. This is the automatic mode. Decisions are based upon internal decision logic and prespecified user decision rules or tables. When running in automatic mode a history of database changes should be maintained to permit rapid user review and easy potential reversal of any actions taken by the system. At the other extreme, the system proceeds only one step at a time, allowing the user to examine all interim results and monitor all subprocesses in detail. Between these two extremes lies intermediate degrees monitoring, selectable by the user.

Conclusion

The determination of unit training requirements is a complex process decomposable into a number of sequential subprocesses. Some subprocesses, such as the computation of turbulence measures and their comparison against criteria can be completely automated. However, for other processes, like the prioritization of training activities on the MTP, the training manager must explicitly make decisions while automation contributes by managing the clerical aspects of the process in a user-centered, highly-interactive, computer environment. Automation will contribute substantially to both the effectiveness and efficiency of the training requirements functional area.

References

ADEA (June, 1985). Functional description for the integrated training management system: System definition. Fort Lewis, WA.

ADEA (September, 1985). Functional description for the integrated training management system: System design (Draft). Fort Lewis, WA.

ADEA (December, 1985) Overview of ITMS. Fort Lewis, WA.

Goehring, D. J. (in preparation). Training requirements determination in short-term unit training management. Research Report, Alexandria, VA: US Army Research Institute.

House, W. C. (1983) Decision support systems: A data-based, model-oriented, user-developed discipline. New York: Petrocelli.

Using CODAP Job Analysis for Training and Selection: Retrospective Considerations

Gerald Fisher, HumRRO
Leaetta Hough, PDRI
Richard Lilienthal, CIVPERCEN

The U.S. Army's civilian workforce currently exceeds 450,000 employees. The Army Civilian Personnel Center (CIVPERCEN) is responsible for designing a personnel system that must select, train, assign, promote and retain civilian employees, supervisors and managers. As the civilian workforce grows, and the need for a more highly skilled and prepared employee increases, new challenges are presented to CIVPERCEN. Among them are the need to make quicker, more defensible, and more objective selection, promotion, and training decisions.

The Personnel Decisions Research Institute (PDRI) and the Human Resources Research Organization (HumRRO) have been assisting CIVPERCEN in conducting a job analysis for three civilian career programs covering 10,000 professional, supervisory and/or managerial employees. The purpose of todays MTA presentation is to trace the procedures and methods used in this effort and how the results of this personnel research can be used to modernize, objectify, and defend training and selection decisions. A paper presented by the authors at last years meeting of the MTA focused on issues related to development of the CODAP* job description inventory used in this study and the qualitative and quantitative elements of inventory development. This presentation focuses primarily on the findings from the CODAP hierarchical cluster analysis (HCA) and how they were derived as well as on the procedures used for subject matter expert (SME) knowledge, skill, and ability (KSA) decisions for selection criteria and training requirements. Examples of findings for supervisors will also be provided.

Theoretical Framework

The Army is in the process of implementing two major civilian personnel programs focusing on a centralized and automated referral and promotion program and a more structured and sequential training planning and delivery system. The Army Civilian Career Evaluation System (ACCES) is the result of a joint effort between CIVPERCEN and the U.S. Office of Personnel Management (OPM) 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 necessary tasks to perform the job are identified, appropriate knowledges, skills, and abilities (KSAs) can be specified for promotion and referral purposes. 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. The current job analysis looks at the problems of developing selection criteria and training requirements in three other career programs within a single effort.

* CODAP is an acronym for the Comprehensive Occupational Data Analysis Programs.

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; Materiel 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 the 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 required tasks are finalized and KSAs are developed both for training requirements and for setting selection criteria.

Method

Since September 1984, PDRI joined by HumRRO has been conducting a CODAP-based job analysis for 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 all 10,000-plus job incumbents at Army installations throughout the world. Following receipt of more than 6,000 completed inventories, CODAP analyses were conducted and pertinent task and duty clusters were developed and validated by SMEs.

Identifying Job Groups through CODAP Cluster Analysis

Twenty-seven job groups were identified by senior SMEs for the 20 job series within the 3 career programs. Separate KSA/SME linkage groups were set up for each job group identified by the data. For instance, in a series such as 1670, the General Facilities and Equipment Series within the Materiel Maintenance Career Program, SMEs identified several different clusters or job groups, including: Field Assistance and Evaluation; Maintenance Support Planning; Provisioning; Integrated Logistic Support; Materiel Evaluation; Weaponry Materiel System Management; Maintenance Operations; and Maintenance Engineering. Separate job descriptions and ACCES and ACTEDS elements were identified by separate SME panels for each job.

In addition to the separate SME panels held for each job group within each career program, a single panel was relied upon to identify first-line supervisory tasks and KSAs. These KSAs and tasks are to be included in training and selection for all three programs with additional technical tasks and KSAs added for specific jobs.

Identifying ACTEDS Training Requirements: Application for Supervisors

SMEs within the Rating and Training Elements Workshops were supplied CODAP job descriptions for each job group. Duties were listed by most to least critical duty area as well as by most to least critical job task within each duty area. So-called Core Tasks were identified for each job and linked to each KSA identified by the SME panel. A core task was identified as:

> "A task that is performed by at least 50% of the group members (i.e., 50% or more of the group) and or contributes to 50% of the criticality."

Criticality was defined as:

> "An equally weighted combination of relative time spent and relative importance. Tasks are ordered by criticality on the job description(within duty category). Thus, the first task (in the job description) was rated as more important and/or time-consuming than the others."

Core Tasks for first-line supervisor are listed in Table 1 on the basis of the performance time frame that SMEs judged the task to be mastered and the most feasible way to teach this task.

Observations

The ACTEDS system uses job task statements as the starting point for identifying training requirements. Related KSAs are subsumed under the task statements in the design and delivery of training for civilian supervisors and professionals. In this study we found that nearly all the key tasks that have to be performed by the first-line supervisor must be learned within the first year of supervisory experience. In fact, the majority of tasks must be performed upon job entry or within the first three months. The conclusion is that training for the first-line supervisor is needed before or immediately after entry to the job on such tasks as: work scheduling and duty and task

assignment to personnel; crediting plans and selection criteria; interviewing and selection; performance appraisal and work evaluation; documentation of work performance; and handling disputes. The methods most often favored for training were guidance and coaching by an experienced worker, job aid (SOP or step-by-step procedure) and classroom learning. Often a combination of training methods were favored.

In terms of ACCES selection requirements for fist-line supervisors the Core Knowledges arrived at by the senior SMEs or those that will be part of all supervisory selections include: EEO, Performance Appraisal, Promotion/ Placement, Management Employee Relations, and Supervisory Methods. Supplemental knowledges or those that the selection official may use for supervisory selection if he/she so wishes are: Information/Materiel Security and Internal Control. Those Core Abilities that will be used (through an accomplishment record) include: Ability to Analyze, Ability to Communicate Orally, Human Relations Ability, Ability to Write, Ability to Plan and Organize, Ability to Innovate, and Ability to Initiate Action. In the ACCES system, applicants numerically rate themselves and are rated by their current supervisor on specific knowledge areas whereas they write an accomplishment record on each ability area used for selection (e.g., problem or objective; what I actually did and when (approximate dates); what the outcome was; verifying person). The implication for supervisory selection is that abilities and the narrative accomplishment records that are used to substantiate them will be relied upon more frequently than specific knowledge areas at least in judging supervisory areas. Tables 2 and 3 present this KSA information.

Conclusion

The preliminary findings from this job analysis for three career programs were that similar job inventories and SMEs can be used for both selection criteria and training requirement analysis. In all supervisor selection and training it is required that basic tasks be mastered upon or soon after entry. In the supervisory selection area the findings using these procedures will see a higher reliance on the accomplishment rating area than on the numerical knowledge rating.

Table 1 — ACTEDS SUPERVISORY TASKS

	Time Frame Rank	Job Level	Performance Time Frame									How Trained?					
			NR	1	2	3	4	5	N	Mean	S.D.	NR	1	2	3	4	5
Interview Candidates for Position Vacancies	8	4	1	1	4	1	0	0	6	2.0	.58	1	0	3	2	0	1
Select Candidates for Position Vacancies	11	4	1	2	4	0	0	0	6	1.7	.47	2	1	2	2	0	1
Develop Personnel Selection Criteria and Crediting Plans	6	4	1	0	5	1	0	0	6	2.2	.37	1	0	2	1	1	1
Determine Staffing Requirements for Operations	1	4	1	0	2	3	0	1	6	3.0	1.00	1	0	2	3	1	0
Assign Specific Duties and Tasks to Personnel	20	4	1	5	1	0	0	0	6	1.2	.37	2	1	4	1	0	0
Allocate Resources (e.g., Personnel, Equipment, Office Space, Budgets) to Subordinates	6	4	1	2	2	1	1	0	6	2.2	1.07	2	1	0	3	1	1
Coordinate Personnel Functions (e.g., Training, Overtime, Leave Requests) with Work Schedule Priorities	15	4	1	5	0	1	0	0	6	1.3	.75	2	1	0	2	1	2
Schedule Work Based on Priorities and Requirements	15	4	1	4	2	0	0	0	6	1.3	.47	1	0	4	1	1	0
Explain Specific Work Requirements, Regulations, or Assignments to Personnel (e.g., Subordinates, Team Members)	13	4	1	3	3	0	0	0	6	1.5	.50	3	2	2	1	0	1
Encourage Subordinates to Participate in Management Programs (e.g., Suggestion, Cost-Reduction, Upward Mobility . . .)	22	4	1	6	0	0	0	0	6	1.0	0	4	3	2	1	0	0
Inform Subordinates of Training Opportunities	15	4	1	4	2	0	0	0	6	1.3	.47	2	1	3	2	0	0
Develop Performance Standards and Individual Development Plans with Subordinates or with Your supervisor	3	4	1	0	3	2	0	1	6	2.8	1.07	1	0	0	2	0	4
Provide Performance Feedback or Counseling to Subordinates	5	4	1	2	1	2	1	0	6	2.3	1.11	2	1	1	1	0	3
Assess Subordinate Capabilities to Make Personnel Recommendations or Decisions (e.g., Career or Promotion . . .)	3	4	1	1	0	4	1	0	6	2.8	.90	1	0	4	1	0	1
Evaluate Work of Subordinates or Team Members for Quality and Timeliness	8	4	1	1	4	1	0	0	6	2.0	.58	2	1	2	1	0	2
Maintain Documentation on Personnel Actions (e.g., Promotion, Disciplinary Actions, Performance Review)	15	4	1	4	2	0	0	0	6	1.3	.47	2	1	0	2	1	2
Write or Revise Job Descriptions	1	4	1	0	2	3	0	1	6	3.0	1.00	1	0	1	2	1	2
Evaluate Requests Made by Personnel (e.g., Leave, Training, Reimbursement of Travel Expenses, Overtime)	15	4	1	4	2	0	0	0	6	1.3	.47	1	0	3	2	1	0
Monitor Participation in or Compliance With Personnel Management Programs (e.g., EEO, Federal Women's Program . . .)	13	4	1	4	1	1	0	0	6	1.5	.76	2	1	0	2	1	2
Monitor Compliance With Union Agreements	11	4	1	2	4	0	0	0	6	1.7	.47	2	1	0	2	0	3
Resolve Complaints or Grievances Submitted by Employees or Union Representatives	10	4	1	3	1	2	0	0	6	1.8	.90	1	0	0	1	0	5
Resolve Disagreements or Conflicts Between Subordinates	20	4	1	5	1	0	0	0	6	1.2	.37	2	1	0	1	0	4

JOB LEVEL

1.a. Determined by (Group majority) What is the lowest level level at which all job incumbents must be able to perform this task at full performance level without assistance?
1=Intern
2=Specialist/Journeyman
3=Non-Supervisory Program Manager (Team Leader)
4=1st Line Supervisor
5=2nd Line Supervisor

PERFORMANCE TIME FRAME

1.b. How soon after entering the job designated in 1.a. must the incumbent be able to perform the task at full performance level without assistance?
1=Immediately upon job entry
2=Within 3 months after job entry
3=Between 3 months & 6 months after job entry
4=6 months to 12 months after job entry
5=More than 12 months after job entry

HOW TRAINED

2. What is the most feasible way to teach this task (consider resource constraints (e.g., funds, facilities, equipment, time)
1=No need to teach (i.e., no training required) (Mentioned for 2 tasks)
2=On-the-job (guidance & coaching by experienced worker) (9 tasks)
3=Job aid (e.g., Standing Operating Procedure, step-by-step procedure) (9 tasks)
4=Correspondence course (e.g., computerized instruction; programmed instruction) (0 tasks)
5=Classroom (9 tasks)

Table 2 — SUPERVISORY KSA RATINGS FOR ACCES

Ability or Knowledge	A Importance of KSA for Overall Job Performance			B Performance Differentiation			Aver. of A&B	C Proficiency			
	N	Mean	S.D.	N	Mean	S.D.		NR	1	2	3
Ability to Analyze	8	4.0	.82	8	3.8	1.34	3.8	2	5	0	0
Ability to Communicate Orally	8	4.3	.47	8	3.5	1.38	3.9	2	5	0	0
Human Relations Ability	8	4.3	.75	8	3.8	1.87	4.1	2	5	0	0
Leadership Ability	8	5.0	0	8	4.2	1.46	4.6	2	3	0	2
Ability to Write	8	4.3	.75	8	3.8	1.34	4.1	2	5	0	0
Ability to Direct Work Activities	8	4.7	.47	8	4.2	1.07	4.4	2	2	0	3
EEO	8	3.8	.88	8	3.3	.94	3.6	2	1	2	2
Supervisory Methods	8	4.8	.37	8	4.5	1.12	4.7	2	1	0	4
Promotion/Placement	8	4.0	.82	8	3.7	1.11	3.8	2	1	1	3
Management Employee Relations	8	4.3	.75	8	3.7	1.37	4.0	2	0	1	4
Ability to Plan and Organize	8	4.8	.37	8	4.2	1.46	4.5	2	5	0	0
Mediating/Negotiating Ability	8	2.8	.37	8	2.8	.37	2.8	2	3	0	2
Ability to Innovate	8	3.5	.69	8	3.7	1.49	3.8	2	5	0	0
Assessment	8	4.2	.80	8	3.3	1.25	3.8	2	1	1	2
Career Management	8	2.7	.94	8	2.3	.47	2.5	2	0	2	3
Position Management	8	3.7	.94	8	3.7	.94	3.7	2	0	0	6
Ability to Initiate Action	8	4.2	.93	8	3.8	1.48	4.0	2	5	0	0
Organizational Restructuring	8	2.2	.37	8	2.2	.37	2.2	2	0	1	4
Internal Control	8	3.1	.94	8	3.0	.82	3.3	2	3	2	0
OSHA	8	2.0	.58	8	1.8	.68	1.8	2	2	0	3
Information/Materiel Security	8	3.5	.78	8	2.5	1.20	3.0	2	4	1	0

A-KSA importance for effective overall job performance.
1 = Unimportant to 5 = Extremely Important

B-KSA importance for distinguishing average from superior performance on the job
1 = Unimportant to 5 = Extremely Important

C-When is the necessary proficiency in this KSA normally acquired?
1 = Prior to job entry
2 = During a brief orientation period (2-3 months) after job entry
3 = More than 3 months after job entry

NOTE: Last 4 KSAs received insufficient scores to be included as final acces selection criteria.

Table 3 — RELATIONSHIP BETWEEN KSA AND CORE TASKS FOR ACCES

ABILITY OR KNOWLEDGE	N	CORE TASKS* >=3	MEAN	S.D.
Ability to Analyze	22	19	3.8	.63
Ability to Communicate Orally	22	18	3.7	.85
Human Relations Ability	22	17	3.6	.91
Leadership Ability	22	17	3.6	.79
Ability to Write	22	17	3.5	.70
Ability to Direct Work Activities	22	16	3.5	.92
EEO	22	13	3.2	.79
Supervisory Methods	22	13	3.0	.79
Promotion/Placement	22	12	2.9	.76
Management Employee Relations	22	11	3.2	1.06
Ability to Plan and Organize	22	11	3.2	.99
Mediating/Negotiating Ability	22	10	2.9	.78
Ability to Innovate	22	10	2.8	.71
Assessment	22	9	2.8	.69
Career Management	22	8	2.5	.95
Position Management	22	7	2.8	.86
Ability to Initiate Action	22	5	2.8	.45
Organizational Restructuring	22	5	2.1	.71
Internal Control	22	3	2.2	.47
OSHA	22	3	1.8	.79
Information/Materiel Security	22	1	2.1	.41

*The [illegible] is "How important is this knowledge or ability for performing this task effectively (i.e., for each core and

IBM CODAP 370 - ALIVE AND WELL IN CANADA

David Owen

Directorate of Military Occupational Structures,
National Defence Headquarters,
Ottawa, Canada.

There are currently four versions of the Comprehensive Occupational Data Analysis Programs (CODAP) being used to conduct occupational analysis. The USAF Human Resources Laboratory (AFHRL) uses the old Sperry (UNIVAC) version which is now being replaced by the completely re-designed ASCII CODAP version. The other two are IBM CODAP 370 which was made available to many users by the Navy Occupational Data Analysis Center (NODAC) in an "export" version, and CODAP 80 designed to be both machine and programmer independent. It was developed as a replacement for CODAP 370.

The Directorate of Military Occupational Structures (DMOS) in the Canadian Forces has been using the IBM version since the early 1970s. We chose IBM because the computer hardware available was IBM and we continue to do so for the same reason. Advances in IBM computer technology have made it necessary to make many modifications to the CODAP 370 package. It continues to work well and has become considerably more efficient through machine and software improvements. We have made many non-CODAP changes to the package to support our continually expanding analysis requirements. The most recent CF improvements include the capability to cluster up to 6000 cases, and the replacement of calls to the Direct Access Input/Output (DAIO) module with FORTRAN I/O statements. The latter change was made necessary by the acquisition of IBM 3380 disk drives.

BACKGROUND

In the 1950's the USAF Human Resources Laboratory (AFSC), Lackland AFB, Texas, Personnel Research Division was actively researching methodologies to define Air Force jobs economically, accurately and reliably. The advent of the computer turned the research project into an operational reality. The embryo of the Comprehensive Occupational Data Analysis Programs (CODAP) was born.

CODAP was originally written for execution on an IBM 7040 computer. Many of the time-consuming outines were written in machine language (MAP) to make them more efficient but, as with any system with modules coded in machine language, it was difficult for agencies not having compatible equipment to make use of CODAP. This problem is still with us today.

In the 1960's the United States Marine Corps undertook a research and development project to determine the organization, procedures, and materiel required to establish a Task Analysis Program. They patterned their program after that of the Canadian Forces. The United States Air Force contributed the CODAP system to the project. The Marine Corps then had the IBM 7040 version rewritten for execution on an IBM 360-65. Subsequently, the Office of the Assistant Secretary of Defense for Manpower and Reserve Affairs made arrangements for major revisions to the programs and procedures to get the system operational on an IBM 370-155 for possible use by all military services including the U.S. Coast Guard. This version remained compatible with the IBM 360 version. It became known as the "export" version of CODAP because it was made available to all U.S. military organizations, other U.S. government agencies, as well as armed forces and agencies of friendly governments. Maintenance and distribution was controlled by the U.S. Navy Occupational Data Analysis Center (NODAC) acting as the Excecutive Agent for IBM CODAP. This is the version which was provided to the Canadian Forces.

In the early 1970's the United States Air Force Human Resources Laboratory obtained a UNIVAC 1108 computer and the CODAP system was rewritten for it. Many of the time-consuming routines were again written in machine language for more efficient processing. This version of CODAP was later expanded to handle task factor analysis. It has undergone continual revision and expansion over the years to meet the requirements of the USAF. This package has been made available to some other agencies having Sperry-Univac hardware. Many of the task factor modules were passed to NODAC for rewrite and inclusion in the IBM 370 version. Unfortunately, they were not incorporated into the production version or the documentation package. We do have the source modules at DMOS, but because there has been no demand for their use they remain in the test library.

In the late 1970's, Texas A & M University was contracted to rewrite the export version of CODAP in ANSI FORTRAN so that it would be relatively independent of machine design and consequently more usable by a wider range of agencies. The rewrite evolved into a complete redesign of the CODAP system with emphasis on flexibility and access to data with little or no programmer interface. The concept is similar to that of the SAS and the SPSS packages. The ANSI FORTRAN CODAP version which became known as CODAP 80 was accepted for distribution in 1985 and support of the "old" IBM 370 version was dropped. In another development, in 1983 MAXIMA Corporation commenced redesign of the Sperry-UNIVAC CODAP system. This version is not yet in full production.

These two updated systems were not suitable for use by the Canadian Forces. Sperry-UNIVAC CODAP is not compatible with our equipment and at about $250,000 is too costly for us to convert to IBM. CODAP 80 does not lend itself to our needs because it cannot cope with the large population sizes of our surveys and it has a costly and time consuming overlap and group process. Faced with not being able to have the new CODAP versions, we decided to continue with the old IBM 370 version, at least for the near future.

PROBLEMS

Having decided to stick with IBM CODAP 370, we soon faced a series of events which caused us considerable anguish and hard work. Our computer centre upgraded their software to VS FORTRAN Release 4.1 as a replacement for FORTRAN IV; making our CODAP procedures inoperable. CODAP calls to system interface modules (IHCxxxx) were not recognized because those modules had been renamed in the new software package. Previous releases of FORTRAN permitted Data Control Block (DCB) parameters such as logical record length, record format, and block size to be specified in the Job Control Language (JCL) for unformatted FORTRAN Input/Output (I/O) operations related to Variable Spanned Blocked (VSB) files. VS FORTRAN considers this inappropriate for these files. Normally, all one would have to do to overcome these two problems is to recompile each FORTRAN source module and sub-routine in the CODAP package using the VS FORTRAN compiler, relink them into new executable modules in the CODAP Load Library, and remove from the JCL any DCB parameters referencing VSB files. It was not quite as simple as that. As we went about the task of re-compiling, the compiler identified errors in some modules that prevented them from executing at all. In addition, some modules compiled and linked correctly but when executed they failed, or produced garbage output, or produced no output at all. All three of these conditions occurred at one time or another in our attempts to get CODAP working under VS FORTRAN. We examined the source modules and discovered instances where source statements were out of sequence, some were missing, and some had been inserted for "debug" purposes. Obviously there were some modules in the CODAP Source Library which were not the same as those compiled and linked into the CODAP Load Library received from NODAC. We had been executing those load modules successfully up to this point.

As we were completing the sorting out of the software problems, our computer centre provided us with another setback by installing new 3380 disk drives and a new 3081 CPU. The result was again quite disastrous. The CODAP Direct Access Input/Output (DAIO) assembler sub-routine failed because it was not compatible with the architecture of the new 3380 disk drives.

Examination of DAIO revealed some major deficiencies in addition to its imcompatibility with 3380 disk drives. At the time it was written it was state-of-the-art; however, today it would be considered a system "hog" because, in handling system I/O, it did not release channels to other programs operating at the same time. A complete rewrite would be necessary. Consultations with our software experts and IBM specialists indicated, given the time and money DAIO could be made operational once more but there was no guarantee that as IBM changed its hardware and software packages DAIO would continue unaffected. We could face the same situation again in the near future.

SOLUTIONS

We were aware that FORTRAN I/O is much more efficient today than it was in the past so our decision on how to solve the DAIO problem was to remove all calls to DAIO and replace them with FORTRAN I/O statements. It was our assumption that IBM would surely take into account the effect of changes to its product line on the FORTRAN software. To allow us the time to make these modifications and still remain operational, our computer centre temporarily retained some 3350 disk drives for our use. For those organizations who have the IBM 370 CODAP source modules and wish to make this change, Appendix "A" contains a list of the modules which make calls to DAIO.

I should tell you that our sister organization in Ottawa, the Public Service Commission, has also been carrying out occupational analysis using the "export" IBM 370 version of CODAP. They execute the load modules on an IBM 4341 under MVS/XA without a FORTRAN compiler nor the FORTRAN Sub-Routine Library. They have been operating successfully for several years but recently faced the same hardware upgrade as we did. They acquired 3380 disk drives with the inevitable result of DAIO failure. As we had already completed our modifications to the CODAP system we supplied them with the new load modules compiled under VS FORTRAN. Unfortunately, Release 4.1 VS FORTRAN requires some run-time modules from the FORTRAN Sub-Routine Library and, of course, the Public Service Commission did not have it. They are presently executing the old load modules using a temporary string of 3350 disk drives while they go through the process of acquiring the FORTRAN Sub-Routine Library. After that they should be operational with the new load modules.

I would like to be able to tell you that FORTRAN I/O is just as efficient in terms of processing speed as DAIO was. Unfortunately, because our computer centre changed to Release 4.1 of VS FORTRAN, a faster 3081 CPU, and new 3380 Disk drives, all at about the same time, the losses and gains in efficiency are somewhat obscured. I can tell you that processing has become considerably faster as a result of all the changes.

Those of you in the programming world can readily understand how much of an impact all these events had on our small programming unit. The problems on more than one occasion seemed insurmountable. However, as we had decided to stay with the IBM 370 version, we made the required modifications and are now moving ahead to enhance our capabilities.

THE FUTURE

We are starting to make continuous use of task factor data such as Learning Difficulty. This has required us to reexamine those modules related to task factors provided to NODAC by the USAFHRL but not incorporated in the IBM CODAP production version. We will examine them closely to see if they can fulfill our particular needs. If not, we will write our own and add them to the CODAP package.

As previously reported to the 5th International Occupational Analysts Workshop at Randolph AFB, Texas in May, 1985, we have added some non-CODAP programs to our analysis package. One of these is the Training Summary (TRGSUM) program now called Job Summary (JOBSUM). This allows presentation of several different types of survey data in a composite way which is easily understood by personnel developing occupational specifications and training.

On the experimental side, we have modified the Overlap and Grouping Programs (OVLGRP) to cluster up to 6000 cases. At this point OVLGRP has been tested successfully on a test sample of 4056. This modification impacts upon a large number of CODAP programs. To date, all but a few have been modified and tested successfully. Testing will be completed in the near future.

CONCLUSION

We in the Canadian Forces consider the IBM 370 version of CODAP, as currently modified, to continue to be a viable product for our purposes. We recognize its limitations, that it is not state-of-the-art, and that it probably could be more efficient. Until we can avail ourselves of a better product for our particular situation we will stay with it. We will continue to modify the system, add programs where necessary, and interact with non-CODAP packages or programs as the need arises. We will also continue to update the User's Manual, which has been provided to us by the US Navy, should others wish to use the modified IBM 370 CODAP package.

ANNEX A

IBM 370 CODAP MODULES WITH CALLS TO DAIO

AVAL14	INPST1	TITLES
AVAL19	JDFPGM	VARGEN
DIAGRM	JOBDEC	BEGPRG
GRPDF1	OVLJDF	PLUCKS
GRPSUM	PRITSK	RESCAN
GRPVAR	PRTVAR	

REFERENCES

Dickinson, R.W. The New CODAP System -- Design Concepts and Capability. Paper presented at 21st Annual Conference of the Military Testing Association, San Diego, California, 15-19 October 1979.

Falle, I.E. Generating Training Summary Reports From CODAP. Paper presented to 5th International Occupational Analysts Workshop, May 1985, Randolph AFB, Tx.

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

Staley, M.R., Weissmuller, J.J., & Phalen, W.J. ASCII CODAP: The Impact of System Design for Emerging Applications. Paper presented to 5th International Occupational Analysts Workshop, May 1985, Randolph AFB, Tx.

Thew, M.C., & Weissmuller, J.J. CODAP: A Current Overview. Paper presented at 21st Annual Military Testing Association Conference, San Diego, California, 15-19 October 1979.

Van Cleve, R.R. Collecting, Analyzing and Reporting Information Describing Jobs and Occupations. Discussion paper presented to 12th Annual Military Testing Association Conference, Oct 1969.

Royal Navy Officers' selection scores and success beyond initial training.

Russell J. Drakeley. Birkbeck College, University of London (UK).

RN officer selection

The Royal Navy's officer selection procedure is known as the Admiralty Interview Board (AIB). All civilian RN officer applicants who meet basic nationality and educational requirements are referred to the AIB without pre-selection except aviator candidates, who must first pass flying aptitude tests. The primary purpose of the AIB was, until recently, to assess candidate's officer potential and predict initial training performance. Validation studies have shown that the AIB is able to achieve this limited objective with some success. Jones (1984), for example, reports average validity coefficients of .56 and .34 against examination results and ratings of officer-like qualities (OLQs) obtained approximately twelve months after entry.

In September 1985 the AIB underwent a series of modifications (including the introduction of scored biographical data) and its purpose was revised to include prediction of later training outcomes. There thus arose a need to determine the validity of the AIB against recent training performance in the Fleet and during specialisation training. This paper therefore presents the results of a follow-up study of approximately 300 officer entrants as far as their first operational appointments (up to four years later). As background, the AIB procedure, and the phases of pre-operational training are described.

The Admiralty Interview Board

The AIB procedure is an example of the assessment centre approach in that candidates are observed both individually and in groups by multiple assessors using multiple techniques. Candidates are assessed in groups of four or five, and arrive at the AIB the evening before the two day procedure. On this first evening candidates complete a biographical questionnaire. On day one candidates complete the AIB psychometric test battery, a general and service knowledge questionnaire, and write a 45 minute essay. At the end of the day all of the written evidence, including any reports or references obtained before the procedure but excluding the psychometric test scores, is collated and forwarded to the individual panel members.

The assessment panel does not meet until day two. It usually consists of four members; the president (Commodore or Captain), a senior Naval officer (usually a Commander), a Personnel Selection Officer, and a civilian school principal. Day two begins with two situational excercises, observed by all members of the panel. The first of these is a "command task" held in the AIB gymnasium. Candidates take turns to lead the group over obstacles using spars, ropes, ladders and similar equipment. Assessments are made of the candidate's performance in command of, and in support of, the rest of the group. Individual panel members marks are them aggregated to produce a single "gym excercise" score. The second exercise is a leaderless group discussion in which candidates are presented with a written problem scenario and required to produce a team solution in the presence of the panel. This is scored in a similar way to the gym exercise. After the exercises, each candidate is interviewed twice, once by the Personnel Selection Officer and once by the remainder of the panel. No marks are awarded specifically for the interviews, and this is the end of the procedure as far as the candidates are concerned.

The final assessment conference now takes place. Through discussion of the available evidence (including, at this stage, the psychometric test scores) the panel arrive at an overall rating; the "Final Board Mark" (FBM). The pre-1985 discussion procedure is described in Herriot and Wingrove (1984). It is worth noting that the FBM is an aggregated, rather than a consensus mark, and it is used to rank order candidates in terms of their suitability to enter training. The actual selection decision, based on the FBM, is made subsequently in the light of manpower requirements.

Initial officer training

Initial training takes place at Britannia Royal Naval College, Dartmouth. With the exception of certain small variations related to specialisation, all entrants follow the same training syllabus, This has two components; professional studies, and leadership training. Professional subjects include navigation, seamanship, operations and warfare, and engineering. Classroom lectures are supplemented by practical instruction in visual and radio communication, and boat handling. Some of this training takes place at sea. The professional syllabus culminates in a series of written examinations, the results of which are summed to produce an overall examination mark.

The leadership syllabus includes a number of practical leadership excerceses, instruction in Naval customs and protocol, and lectures on "functional leadership" (Adair, 1968). The purpose of this training is to prepare the entrants to act and react as officers. Trainee officers' leadership performance is assessed termly by college staff officers using a standard report form. This consists of 22 behaviourally anchored ratings such as "sense of duty", "initiative" and 'power of command". These are summed to produce an overall OLQ mark. In the case of Seaman (Executive) branch entrants (discussed in later sections of this paper), initial training lasts twelve or sixteen months, depending of the length of their chosen commission. If successful at the Naval college they proceed to further training in the Fleet.

Fleet training

Fleet training provides Naval College graduates with an opportunity to put into practice much of the theory learned in the classroom. The aim is to impart sufficient knowledge of the management and operation of HM ships so that the trainee can safely assist with watchkeeping duties both as sea and in harbour. There is less formal leadership training, although trainees are expected to show progress in this area. OLQ ratings therefore continue during this period, using the standard 22 scale report form. The syllabus takes the form of a series of mandatory tasks which involve the trainee in the day to day activities of each of the ship's departments. These tasks are signed off in a "task book" issued to each trainee at the beginning of Fleet training. The majority of Seaman branch trainees spend their Fleet time in at least one major war vessel. The ship's operational program largely dictates the course of Fleet training, hence there is no fixed period, although a minimum of 5 months is specified. Fleet training culminates in the "Fleet Board" examination. The origins of the Fleet Board can be traced to the "Lieutenant's examination" instigated by Samuel Pepys in 1677. The purpose of the Lieutenant's examination was to determine whether a candidate for a commission was able to "judge of and perform the duty of an able seaman and midshipman and his having attained to a sufficient degree of knowledge in the theory of navigation capacitating him thereto" (Pepys, quoted in Ollard, 1974). The examination was conducted orally by three senior officers, one of

whom was of flag rank, and candidates were required to furnish certificates of competence and of good conduct.

The modern Fleet Board is still an oral examination, although the number of assessors has doubled to six. As in Pepys's day, candidates present certificates of competence (in boat handling, radio and visual communications, and watchkeeping) and a number of written journal articles (intended to develop the trainees' report writing and communication skills). The examinations are conducted on a one to one basis and each of the major areas of professional competence is covered individually (see Table 3). The marks awarded in the examinations are combined with those awarded for the journal articles to produce the total Fleet Board examination mark. Candidates must both pass the Fleet Board and obtain satisfactory ratings of OLQs in the Fleet before being confirmed in the rank of Sub-Lieutenant (the "commissioned" rank).

Specialisation training

Specialisation training for Seaman branch officers takes place in various shore establishments in the Portsmouth Naval base area, and lasts sixteen weeks. It is intended to enable officers to carry out bridge watchkeeping and small ship navigation unsupervised, to introduce the principles of co-ordinating the ships weapons and defensive systems from the operations room, and to manage the affairs of a division of enlisted men. Formal leadership training takes place in the Welsh mountains, where they are placed in command of a small company of enlisted men.

The course is examined in three ways, Firstly, students are assessed in navigation during a week at sea. The practical examination includes tasks where the students are required to make rapid navigational decisions under pressure, for example negotiating a narrow channel or dropping anchor at a buoy. In these situations the student, in the role of navigating officer, has virtual control of the ship. Written examinations are held in three subjects; divisional (man) management, operations and warfare, and damage control (including defence against nuclear, biological or chemical attack). Lastly, performance in the leadership excercises is rated on the standard report form discussed above, to produce another OLQ assessment.

On successful completion of specialisation training, officers return to the Fleet to take up operational appointments. In theory it is possible to complete specialisation training within two years of entry. In practice it often takes longer because of disruption to the Fleet program, leave periods and educational appointments such as sponsorship to universities. In the last case the period can exceed four years.

The follow up study

Although candidates for other specialisations attend the AIB, the follow-up study was limited to the Seaman branch specialisation. The sample consisted of 315 Seaman branch entrants who entered initial training between January 1981 and January 1983. Selection and training data were obtained from Ministry of Defence and AIB records. The full data set is too large to be included here. Instead, four selection variables of particular interest, and six training outcomes have been selected as follows:

1. Selection variables
   a. Gym exercise mark. This was the more valid of the two selection exercises designed to detect leadership potential.

b. AIB test battery. A weighted composite of the four psychometric tests.
c. Scored biographical data. A biodata device empirically keyed to predict initial training examination results (see Drakeley, 1984). Scores were calculated by retrospectively applying the key to data held in personnel records since it was not in use at the time the sample entered training.
d. The AIB Final Board Mark. The overall selection ranking scores.

2. Training outcomes
   a. Initial training. Written professional examination total and OLQ ratings.
   b. Fleet Board oral examination total and OLQ ratings.
   c. Specialisation Training. The total mark awarded for the practical examination in navigation and three written examinations, and OLQ ratings.

Table 1 shows the training progress and losses from the sample at each stage.

Table 1. Progress through training: Seaman branch entrants.

Entrants	315		
	↓ ← 1		Transfer in
Initial training	316		
Voluntary attrition	48 ← | → 1		Transfer out
Compulsory attrition	33 ← ↓		
Fleet training	234		
Voluntary attrition	15 ← | → 11		Transfer out
Compulsory attrition	7 ← ↓ → 13		To University
Specialisation training	188		
Voluntary attrition	7 ← | → 2		Transfer out
Compulsory attrition	4 ← ↓ → 12		To University
Available for duty	163		

The left hand side of the table shows the number of individuals lost to the Seaman branch through attrition. It is clear that voluntary attrition exceeds compulsory attrition at each stage, resulting in overall rates of 22% and 13% respectively. It should be noted that the AIB overall ranking score (the FBM) does not predict voluntary attrition (Jones, 1984). A biodata device keyed to predict early voluntary attrition (approximately 50% of which occurs within three months of entry) was implemented in September 1985. Figures to the right show the numbers "lost" to other specialisations (transfers) and to university sponsorships. Most of the latter will re-join the training pipeline at a later stage. A consequence of compulsory attrition and the loss of the more intellectually able entrants to university was increasing range restriction at every stage. Corrections were therefore applied to the correlations between the selection variables and the training outcomes. These are shown in Table 2.

Turning first to the relationship between the selection variables and examination performance, it is apparent that the best overall predictors were the AIB final mark and the test battery. The moderate to high validities against the initial examinations are particularly encouraging considering the aim of the procedure at that time. The biodata device keyed to predict this outcome also showed some generality. Its validity against initial examination performance may be slightly inflated, however, since the sample included some individuals from the key development group. It is noticable, and perhaps not surprising, that the predictive power of the AIB declines as the entrants proceed through training. Although increasing range restriction may have contributed to this result, it is reasonable to suppose that training evens

out some of the differences observed during selection.

Table 2. Corrected correlations: selection variables & training assessments.

Training phase	Initial		Fleet		Specialisation	
assessments	exam (w)	OLQ	exam (o)	OLQ	exam (w/p)	OLQ
sample size	269	248	192	195	164	164
Selection variables						
gym exercise	.25	.26	.38	.32	.20	.15
test battery	.54	.11	.41	.44	.32	.25
biodata	.59	.10	.33	.36	.29	.19
AIB final mark	.55	.16	.45	.51	.33	.28
Initial training						
examination (w)			.48	.38	.42	.26
OLQ			.28	.58	.24	.38
Fleet training						
examination (o)					.35	.28
OLQ					.36	.50

NB For training examinations, w=written, o=oral, w/p=written/practical.

In contrast, validities against initial OLQ ratings were uniformly low (only the gym exercise exceeded .20) whereas Fleet OLQs were easier to predict. It is possible that leadership skills develop slowly and it takes time to aquire a consistent leadership style. Alternatively, initial OLQ ratings are closely based on the practical leadership exercises of which the AIB gym exercise is perhaps a "job-sample". These somewhat contrived situations may not measure the same facets of leadership required in the Fleet where training is much less structured. In the case of the two later stages, the best predictors of OLQs were the ratings made at the previous stage. While this may have been due in part to behavioural consistency, it should be noted that earlier OLQ ratings were available to assessors at later stages. The problem of potential criterion contamination did not apply to the selection scores, which were not forwarded to the training establishments.

The correlation between the initial and specialisation training examination was at least as high, if not higher than that between the latter and the intervening Fleet Board exam (.42 vs .35). Although all three were intended to be measures of professional knowledge, it is possible that the relationships were confounded by the method of assessment; written and written/practical examination in the case of initial and specialisation training, oral examination in the case of the Fleet Board. This possibility, and the foregiong discussion of different facets of leadership performance lead to an interest in the underlying dimensionality of the criteria. Factor analysis results are as yet only available for the Fleet and specialisation training criteria. Table 3 shows rotated factor loadings on the individual examination and OLQ ratings at these stages. Only factors accounting for at least 10% of the total factor variance, and loadings greater than .30 are shown.

Despite a considerable overlap in subject content, no "subject dimensions" appear, rather the professional examination performance dimensions appear to be "written" and "oral" (factors 1 and 4). While this interpretation is intuitively appealing, there is of course the confounding effect of the time of examination. This applies less to the leadership dimensions (factors 2 and 3) which span the two training periods. Factor 2 loads on both sets of OLQ ratings and the practical navigation exam. It was implied above that

success in this examination depends on decisiveness and ability to issue instructions under pressure. This factor could perhaps be labelled "power of command/decisiveness". Factor 3 again loads on both OLQ ratings but also on the mark awarded for the journal articles. Good journal writing requires careful planning and good communication skills. "Organising and communicating ability" might be an appropriate label for this factor. Without the individual scores for each of the rating dimensions, these labels can only be tentative. However, a better understanding of the criteria may lead to better predictors. Drakeley (1984) reported an inability to predict OLQ ratings with biodata. "Power of command" type biodata items might be hard to imagine, but evidence of prior communciating or organising ability might be easier to obtain from personal history items. A device developed to predict only this facet of leadership might be more successful.

Table 3. Rotated factor loadings: later training assessments

Factors		1	2	3	4
	seamanship (incl. navigation)	.42			
	operations and warfare	.34			
	divisional management	.54			
Fleet	damage control	.69			
Training	pay, stores and Naval law	.35			
	weapons engineering	.70			
	mechanical engineering	.53			
	journal articles			.57	
	officer-like qualities		.34	.58	
	officer-like qualities		.63	.32	
Specialisation	navigation		.63		
Training	operations and warfare				.39
	divisional management				.36
	damage control				.36

Conclusion : AIB and later training outcomes

The results presented here are broadly consistent with previous research which shows that the AIB is able to predict initial examination performance at a moderate level (around .50). While validities against later assessments of professional knowledge are less (around .30 up to four years after entry), this may be offset by enhanced prediction of later OLQ ratings. There is thus no evidence to suggest that the new "goal" of prediction, success beyond initial training, will necessitate any fundamental changes in the procedure. Some fine tuning might be achieved, however, through investigation of the dimensions of later training performance.

References

Adair, J. Training for Leadership. London. McDonald and Jones. 1968.

Drakeley, R.J. The use of biographical data in the selection of Royal Navy Officers. Paper, 26th MTA conference, Munich. 1984.

Herriot, P. & Wingrove, J. The consensus discussion in an assessment centre for the selection of Naval Officers. Paper, 26th MTA conference, Munich, 1984.

Jones, A. Royal Navy Officer selection : Developments, current procedures and research. Paper, 26th MTA conference, Munich. 1984.

Ollard, R. Pepys : A Biography. London. Hodder and Stroughton. 1974.

The Effects of Remedial Training on Classroom Performance

at the

U.S. Army Ordnance Missile and Munitions Center and School

by J. W. Illes

Disclaimer The views expressed in this paper are those of the author and do not in any way represent the official views of the United States Army: Nor those of the U.S. Army Ordnance Missile and Munitions Center and School or any of its component elements.

Background. The U.S. Army Ordnance Missile and Munitions center and School (USAOMMCS) at Redstone Arsenal, Alabama is one of the schools operated by the U.S. Army that is charged with Initial Entry Training (IET);i.e., the training of recruits in their basic specialty. At USAOMMCS they are trained as missile or munitions maintenance technicians. They are qualified for entry into their respective courses through their Armed Services Vocational Aptitude BAttery (ASVAB) scores. In the case of the missile maintenance technicians the qualifying score measures an aptitude for learning electronics, while for the conventional munitions maintenance personnel, the ASVAB test of concern in General Maintenance.

Upon arrival at the OMMCS the trainees are administered the Test of Adult Basic Education (TABE) to determine reading level and mathematical ability. In the past, soldiers with ASVAB General Technical scores of less than 100 and who scored at or below the eighth grade level on some part of the TABE were given remedial training before starting their technical training. This could have lasted as long as 240 hours. The program was front-end loaded, so there was no way to establish a control group without depriving a soldier of needed remediation. Consequently, there was no way to measure the effectiveness of this remedial training except through indirect means such as any increase or decrease in ASVAB re-test scores.

In 1985 the U.S. Army Training and Doctrine Command (TRADOC) announced that with the beginning of Fiscal Year 1986 all IET soldiers would receive remedial training on an as needed basis only. The soldier's need was determined by his commander based on instructor class-room perceptions of his performance. This remedial work was to be conducted during on or off-duty hours depending on the soldier's difficulty. The remediation could be as short as an hour or could last over whatever time was required, but not to exceed 240 hours. The TABE is still administered upon arrival so as to provide commanders with a tool to assist them in their evaluation of student problem areas.

Related Literature. A recent U.S. Navy experience reported by H.L. Bowman (29th Annual Meeting of the College Reading Association, Pittsburgh, PA.,October 24-26 1985) leaves little doubt that remedial training in basic academic subjects can be effective. Bowman notes that the Academic Remedial Training (ART) prgram has had a very high success rate in correcting the reading deficiencies of newly recruited personnel during the period from 1981 to 1984.

The ART was the subject of another paper presented by Bowman (Mid-South Educational Research Association 14th Annual Convention at Biloxi, MS, November 6-8, 1985). His purpose was to examine the variables being used to identify U.S. Navy recruits who are in need of remedial basic academic training. Of particular interest was the fact that Bowman considered reading grade level as the determinant for remedial training. While only recruit training was addressed the U.S. Navy considers ART to be highly successful, front-end loaded, full-time prcgram.

Purpose. The purpose here was to determine whether remedial training as conducted at USAOMMCS had any significant effect on class-room performance as measured by End-of-Course (EOC) grades. The null hypothesis to be tested was that there was no significant difference at the >.05 level in the EOC grades of those students who had attended remedial training and those who had not had the benefit of remediation.

Procedure. Data gathering for this study was started on 1 July 1985. The results of TABE testing were entered into a data base management system called the Information Processing Family 2 (IPF2) on the OMMCS mainframe computer, the Control Data Cyber 380. Identifying data, ASVAB scores, and classroom achievement were also entered into each soldiers record. Only IET soldiers were entered into the data base. Prior service personnel were excluded or deleted f: .n the data base when found. The same standard held for members of the National Guard or other reserve components. Soldiers administratively relieved from training were also deleted from the data base. Only those soldiers whose data were complete and had either graduated or were academically relieved remained in the data base for analysis.

The Analysis of Covariance (ANCOVA) procedure was used because the subjects were categorized on the basis of testing rather than at random. Since a post-test comparison was being conducted, the pre-test differences in the covariate TABE scores had to be neutralized. This determined that the ANCOVA procedure be used.

Because missile and munitions maintenance students have different selection criteria and qualifying scores each group was treated to separate analyses. Within each group only those who scored 8.0 or less on some part of the TABE were used in the analysis. For lack of a better term, these soldiers were called "Eligible". Those who attended remedial training for at least four hours became part of the experimental group, while those who did not were in the control group. Other than those with that four hour minimum, no attempt was made to separate soldiers by length of remediation. As a result 152 missile and 312 munitions maintenance soldiers were considered in the analysis. A matrix showing the number of soldiers in each category appears below:

		MISSILE MAINTENANCE			MUNITIONS MAINTENANCE		
		ATTENDED			ATTENDED		
		Y	N	TOT	Y	N	TOT
ELIG	Y	47	105	152	124	188	312
	N	36	464	500	19	521	540
	TOT	83	569	652	143	709	852

Figure 1

The separate variance t-test model showed significant differences (>.05) in the EOC averages of each group in the missile and munitions maintenance tracks with scores of 6.71 and 10.48 respectively. This was done to demonstrate that the two groups were operating at different academic levels.

Two other t-tests were condcted. There was no significant difference (>0.05) in the qualifying scores of the eligible missile and munitions groups with respective t-scores of 1.30 and 0.81; showing that the two groups at least started from similar aptitude levels.

Findings. Figure 2 shops that there was no significant difference (>0.05) in the munitions maintenance group EOC scores achieved by the experimental group and the EOC scores of the soldiers in the control group who did not take the remedial training.

MUNITIONS MAINTENANCE PERSONNEL
ANALYSIS OF COVARIANCE

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	RESIDUALS F
Between	1	29.18	29.18	0.44
Within	306	20196.56	65.57	
Total	309	20225.73		

Figure 2

Figure 3 shows no significant difference (>0.05) in the EOC scores achieved by the Missile Maintenance students in the experimental group and those in the control group who did not take remedial training.

MISSILE MAINTENANCE PERSONNEL
ANALYSIS OF COVARIANCE

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	RESIDUALS F
Between	1	655.18	655.18	3.72
Within	148	26039.34	175.94	
Total	149	26694.52		

Figure 3

Conclusion. The "F" scores resulting from the analyses of co-variance did not demonstrate significant differences at the >0.05 level. The stipulated null hypothesis is therefor valid; the current remedial program does not result in significantly different EOC scores. This contradicts Bowman's implications that remedial academic programs for military personnel are successful. Again, it should be noted that the format of the two programs differ considerably.

Recommendations Generally that the U.S. Army take a page from its U.S. Navy counter-parts and give some serious thought to a front-end loaded remedial program as a means of enhancing IET training.

a. That another study be conducted using other pertinent variables to confirm or dispute these OMMCS findings. Some variables which might be considered are the following; civilian eduction level; mental category as indicated by the Armed Forces Qualifying Test score; whether missile or munitions maintenance was a first enlistment choice; general physical profile; and, length of remediation.

b. A second recommendation is that course re-cycle policies upon completion of remediation be re-worked so as to give the student a "running start" at his problem area; and, re-inforce the remediation through follow-up checks on the progress of remedial students.

Disclaimer. The views expressed in this paper are those of the author only and do not in any way represent the official views of United States Army; nor those of the U.S. Army Ordnance Missile and Munitions Center and School or any of its component elements.

LEARNING STYLES INVENTORIES - THEIR VALUE AND USE IN THE NAVY TRAINING CLASSROOM

ANN M. DICKSON, EdD
NAVAL UNDERSEA MEDICAL INSTITUTE
NAVAL SUBMARINE BASE NEW LONDON

INTRODUCTION

During fiscal year 1986, the Navy enrolled 1,071,904 enlisted personnel and officers in a variety of courses. On any one day, the average number of personnel attending classes was 84,261. Eighty percent of this number were in a group-paced courses. (Niedert, 1986). Since a large number of students are sitting in Navy classrooms, it is only natural to ask questions about what is known about the process by which students learn, i.e., their learning style. For the purpose of this short presentation, we will examine the Experiential Learning Model, the instuments which purport to identify learning styles, and the strategies, based on the research, which may optimize the effectiveness of training in the Navy group-paced classroom.

"Learning Style" has been defined in a number of ways: Claxton and Ralston (1978) define it as "a student's consistent way of responding to and using stimuli in the context of learning;" Riechmann (1978) defines it as a "particular set of behaviors and attitudes related to the ...learning context." Smith & Kolb(1985) define learning style as "how that person deals with ideas and day-to-day situations."

This sampling of definitions suggests that an examination of the elements "that constitute learning style reveals that among educators, psychologists and researchers who have published studies, definitions vary greatly "(Dunn, Dunn & Price, 1977, pg 419).

Research suggests that "students or trainees learn more quickly, efficiently, and comfortably when learning experiences are geared to their learning needs. If the student/trainee does not want the experience, or is unprepared for the difficulty of an unfamiliar learning situation, hands-on-experience may be as threatening for an abstract learner as a lecture is frustrating for an active one" (Smith & Kolb, 1986, pg 2). Eison and Pollio (1985) conducted a research of the literature on learning styles and suggest, "researchers are becoming increasingly interested in the relationship between student learning styles and other issues, such as academic achievement, preference for particular types of learning contexts, and instructor evaluation" (Eison and Pollio, 1985, pg 442). Dunn and Bruno (1986) suggest, "when students are taught through resources and strategies complementing their individual preferences, significantly increased achievement results" (Dunn & Dunn, 1986, pg 43).

LEARNING MODELS

A number of instruments have been developed which purport to identify learning style, and based on their use, a number

of recommendations for changing the classroom or training environment have been suggested. Let us examine the conceptual basis for these instruments. The Experiential Learning Model proposed by Kolb is based on cognitive theory. "The model emphasizes the role experience plays in learning, an emphasis that distinguishes this approach from other learning-process theories" (Smith & Kolb, 1986). Learning is described as a four-stage cycle: "1. Immediate or concrete experience, which is the basis for 2. Observations and reflections. 3. These observations and reflections are assimilated and distilled into a theory or concept--however informal--from which new implications for action can be drawn. 4. These implications can be tested and serve as guides in creating new experiences" (Smith & Kolb, 1986, pg 12). The effective learner needs four different abilities: Concrete Experiences, Reflective Observation, Abstract Conceptualization and Active Experimentation. "A closer examination of the four-stage learning model suggests that learning requires abilities that are polar opposites, and that the learner must continually choose which set of learning abilities he or she will use in a specific learning situation" (Smith & Kolb, 1985, pg 13). Based on these four abilities, Kolb identifed four types of learners: Accommodator, Converger, Diverger, and Assimilator. He suggests a learner is predominantly of one type, and that each type requires a different type of learning environment. Marshall and Merritt (1986) suggest that the model has validity for assessing learning styles, but that utilization of the model for research and practical application has suffered because of inadequate instrumentation.

Keefe (1979) "conceptualized a learning style as comprised of three types of behaviors: cognitive, affective, and physiological/physical. A cognitive behavior is viewed as one resulting from a preference for a given type of information processing or cognitive style. An affective behavior is the result of a given attitude or opinion. Physical/physiological learning style behaviors are of two types: environmental factors that impinge on learning and biological factors in the makeup of the individual that have an impact on the learning situation" (Ferrell, 1983, pg. 33).

Dunn, Dunn & Price (1977) suggest their research data yielded 18 categories which suggested learners are affected by their (a) immediate environment: sound, temperature, light and design; (b) emotionality: motivation, responsibility, persistence; (c) sociological needs: self, pairs, peers, teams; (d) physical needs: perceptual strengths and/or weaknesses, time of day, intake of food and fluids, and mobility.

LEARNING STYLES INVENTORIES

Learning style instruments have been used primarily in the fields of education, business, and medicine. They have been used to examine the relationship between learning style and age, educational level, undergraduate major, creativity, personality, occupation, career choice, career choice influence, and

preference for a particular instructional method of learning situtation. In the area of instructional methodology, Smith & Kolb (1986) and Dunn & Dunn (1978) suggest that when students' learning styles are matched with instructors' learning styles, learning is the most effective. The most widely used Learning Style Inventories are the following: Grasha & Riechmann Student Learning Style Scales, Kolb's Learning Styles Inventory, Dunn Learning Style Inventory, and the Johnson Decision Making Inventory. The Navy uses Kolb's Learning Styles Inventory in its Leadership Management Education and Training courses at the Naval Submarine Base New London. Ferrell (1983) compared these instruments to determine if cognitive, affective, and physiological/physical behaviors were being measured by these instruments. Each of the instruments tapped only one or two areas of behavior that comprise a wider range that make up the construct of learning style. Of the four instruments, Kolb's Learning Styles Inventory and Johnson Decision Making Inventory tapped cognitive behavior. Dunn's tapped cognitive and physical/physiological; and Grasha-Riechmann's tapped cognitive and affective behavior. "No one instrument stood out as better than the others, and they vary in the degree of factor analytic support for their conceptualization and the amount of variance accounted for. The implication is that either the instrument or the paradigm is lacking, perhaps both" (Ferrell, 1983).

A sampling of validity and reliability studies of Kolb's Learning Styles Inventory has presented conflicting data and analysis. (Plovnick, 1975; Freedman & Stumpf, 1978, 1980; Lamb & Certo, 1978; Geller, 1979 ; Wunderlich & Gjerde, 1978; Fox, 1984; Marshall & Merritt, 1986; Smith & Kolb, 1986)

Fox (1984), whose research does not support the construct validity of Kolb's Learning Styles Inventory (LSI), summarizes the implication of his conclusion: "(1) either the LSI is not an adequate indicator of learning styles, (2) the descriptors Kolb used to characterize the four quadrants of the learning styles matrix inaccurately reflect the real attributes of each learning style, or (3) learning style is not the basis for either preferences for, or evaluation of, educational activities" (Fox, 1984, pp 83-84).

Although models describing the learning process lack construct validity and instruments designed to measure learning styles lack validity and reliability, educators intuitively know that students do learn differently-according to their own preferred learning styles. Does the process, however, measurably affect the product or outcome? In general, those researchers who say "yes" are those who have written the instruments. (Grasha, 1972; Riechmann & Grasha, 1974; Riechmann, 1978; Dunn & Dunn, 1978; Smith & Kolb, 1986) Why should the Navy be interested in knowing how students perceive new information or experiences and how they process what they perceive?

THE NAVY GROUP-PACED CLASSROOM

The Navy sends 84,261 students into the classroom each day. Let us examine briefly the characteristics of the Navy

group-paced, training classroom:

1. Courses are standardized. There is little or no deviation from the topic, time frame, or instructional delivery system used from group-to-group. All students have the same instructional experience.

2. The courses are group paced, allowing little time, if any, for individualizing the instruction by changing the rate of instruction. Courses are taught to the "average-rate" learner.

3. The courses utilize a lecture style delivery. Those students lacking listening skills and/or notetaking skills are at a disadvantage.

4. Instruction is by personnel who are content experts, but not professional educators. Training is a secondary task, not the primary task for which the instructor has been trained. Instructors may lack the ability to adapt to changing populations, the means, via the lesson topic guides, to change, or the knowledge to guide their students in appropriate study techniques.

5. Students have little freedom of action. Absenteeism, tardiness, flexible scheduling are not allowed. Behavior must conform to established standards.

6. Courses are developed to "train", not "educate". Training is lean. The trainee is taught only what is needed to do a job or task. The conceptual basis for behavior is not always presented. The abstract learner, the one who needs to know "why", is at a disadvantage over the trainee who wants to know only "how".

7. Training is intensive. Eight-hour days are common. Little assimilation time is allowed.

8. Instruction in study skills by specialists is usually not available.

The general characteristics of the Navy group-paced classroom demonstrate the learning environment is not learner centered. The research, however, stresses the importance of placing more emphasis on the learner, assessing learning styles, and adapting training accordingly.

RECOMMENDATIONS

What then may the educational managers do to increase the effectiveness of learning?

1. Research should be encouraged to assess the validity and reliability of learning styles instruments, to implement a learning styles assessment program, and to revise training programs accordingly.

2. Each command should train counselors in advising and study skills so that programs may be developed for those seeking or needing assistance. Research studies have demonstrated that students are neither adequately trained in a systematic way in various study skills, nor do they intuitively know how to learn

or what is important to learn. (Christen & Murphy, 1984) Although there are many reasons why students fail, improper study habits and inattention to school work are two factors often cited in the research. (Hart & Keller, 1979) However, if you have ever attended a Navy Academic or Student Review Board,you probably have not heard these reasons cited by too many enlisted personnel. If, however, you ask the students to describe their study habits, you will note the inadequate study skills. Using commercially produced or command developed instruments, trained counselors could survey study attitudes and then determine the content of a study skills class or individual counseling session. Students need to be taught that success may be achieved by adopting those skills which work best for the individual.

3. Students need to be considered as active learners. It has been useful for the Navy to refer to students as "trainees". The term reminds the managers of the mission of the training - to provide practical, on-the-job skills. However, the use of "trainee" detracts from a consideration of the complex process by which students perceive new information or experiences and how they process what they perceive. It is the process of learning in the Navy classroom which needs to be understood better.

SUMMARY

When an organization trains 84,261 students each day, it is important to consider the needs of the learner as well as the needs of the organization.

BIBLIOGRAPHY

Claxton, C. S. & Ralston, T. Learning Styles: Their Impact on Teaching and Administration (AAHE-ERIC/Higher Education Research Report No. 10.) Washington, D.C.: AAHE, 1978.

Cotterell, J.L. Matching teaching to learners: A review of a decade of research. Psychology in the Schools, 1982, 19, 106-112.

Christen, W.L. & Murphy, T.J. Learning how to learn: How important are study skills?" NASSP Bulletin, October 1985, 83-88.

Dunn, Rita & Bruno, Angela. Dealing with learning styles. The Education Digest, 1986, 51, 43.

Dunn, Rita & Dunn, Kenneth. Teaching Students Through Their Individual Learning Styles. Reston, VA: Reston Publishing Co., 1978.

Dunn, Rita, Dunn, Kenneth & Price, Gary. Diagnosing learning styles: A prescription for avoiding malpractice suits. Delta Kappan, 1977, 58, 418-410.

Eison, James & Pollio, Howard R. A multidimensional approach to the definition of college students' learning styles. Journal of College Student Personnel, 1985, 26, 434-43.

Ferrell, Barbara G. A factor analytic comparison of four learning styles instruments. Journal of Educational Psychology, 1983, 75, 33-39.

Fox, R.D. Learning styles and instructional preferences in continuing education for health professionals: a validity study of the LSI. Adult Education Quarterly, 1984, 35, 72-85.

Freedman, Richard & Stumpf, S.A. Learning style theory: Less than meets the eye. Academy of Management Review, 1980, 5, 445-447.

Freedman, Richard D. & Stumpf, Stephen A. What can one learn from the learning style inventory?" Academy of Management Journal, 1978, 21, 275-281.

Geller, Lester M. Reliability of the learning style inventory. Psychological Reports, 1979, 44, 555-561.

Grasha, A.F. Observations of relating teaching goals to student response styles and classroom methods. American Psychologist, 1972, 27, 144-147.

Hart, D. & Keller, M.J. Self reported reasons for poor academic performance of first-term freshmen. Journal of College Student Personnel, September 1979, 529-534.

Keefe, J.W. Learning Style: An overview. In Student Learning Styles: Diagnosing and Prescribing Programs. Reston, VA: National Association of Secondary School Principals, 1979.

Lamb, Steven W. & Certo, Samuel C. The learning styles inventory (LSI) and instrument bias. Academy of Management Proceedings, 1978, 28-32.

Marshall, John C. & Merritt, Sharon L. Reliability and construct validity of the learning style questionnaire. Educational and Psychological Measurement, 1986, 46, 257-263.

Niedert, D.P. Naval Education and Training Program Management Support Activity, Pensacola, FL. Telephone interview, October 21, 1986.

Plovnick, M.S. Primary care career choices and medical student learning styles. Journal of Medical Education, 1975, 50, 849-855.

Riechmann, S. Learning Styles: Their role in teaching evaluation and course design. Paper presented at the 86th annual meeting of the American Psychological Association, Toronto, 1978.

Riechmann, S.W. & Grasha, A.F. A rational approach to developing and assessing the construct validity of a student learning style scales instrument. Journal of Psychology, 1974, 87, 213-223.

Smith, Donna M. & Kolb, David A. User's Guide for the Learning Style Inventory: A Manual for Teachers and Trainers. Boston: McBer and Company, 1986.

Stumpf, Stephen A. & Freedman, Richard. The learning style inventory: Still less than meets the eye. Academy of Management Review, 1981, 6, 297-29.

Wunderlich, Roger, M.D. & Gjerde, Craig. Another look at learning style inventory and medical career choice. Journal of Medical Education, 1978, 43, 45-54.

INTEGRATING COGNITIVE LEARNING STRATEGIES INTO TRAINING

George M. Usova, Ph.D.
U. S. Dept. of the Navy

The author conducted a needs assessment of Naval shipyard training and program offerings to determine if there existed a need to develop instruction in the area of Learning - Study Skills efficiency to support academic and trade theory instruction in the Apprentice Program. The data received strongly support a need for developing shipyard-wide instruction in this area.

The overall mission of the Shipyard Training Modernization Program is to modernize instruction in skills trade training. That modernization effort includes using the Instructional Systems Design (ISD) approach for either developing instruction where none exists or in improving existing instruction to conform to sound principles of educational technology. Those sound principles ensure that all instruction developed consists of valid, well-defined and clearly stated objectives, consistent and sound test items, and instructional lessons which clearly support the objectives and contain the lesson components recognized in high quality instruction, such as motiviation, demonstration, student practice, reinforcement, and feedback. In sum, the ISD approach to training ensures that the essential information to perform the job is presented so that student learning can occur and be adequately measured.

Even though Shipyard Training Modernization Program instruction is developed in a sound fashion and in accordance with the ISD principles, students themselves may have difficulty or inefficiencies in learning the material presented, particularly in the area of knowledge acquisition. Student learning efficiency and study habits are essential to success in achieving knowledge and skills in trade instruction.

Successful training depends on successful learning. Students need to know how to learn, process, interpret, and remember information; they need specific information on essential learning strategies, such as effective notetaking, listening, memory techniques, concentration, time management, reading rate adjustment, and others.

An answer to enchancing learning potential is to integrate these cognitive learning strategies into the training of trade content. Job trainees may either have forgotten how to learn or never acquired learning study skills in the first place. According to Diekhoff (1982), the following findings indicate a need for learning efficiency instruction: (1) Test results indicate that between 15% and 30% of 12th-grade students read at or below the 9th-grade level; (2) Surveys of technical trainees in the armed services show little variation in approaches to learning from technical manuals; (3) Most students report that they learn by reading and learning essentially by rote; and the percentage of the population in the 18-24 age range (in which learning skills are declining and from which many technical trainees are recruited) is projected to decline to only 8% in 1995 (compared to 13% in 1975). There are fewer applicants to choose from and the quality of the pool is decreasing.

Of the many organizations that rely on training to fill highly technical jobs, none has the resources or experience of the Department of Defense (DOD). In recent years, various DOD research and development agencies - most notably the Defense Advanced Research Projects Agency, the Air Force Human Resources Laboratory and the Army Research Institute for Behavioral and Social Sciences - have recognized a need for learning skills pretraining and have targeted increasing portions of their training budgets for designing and evaluating "learning strategy training programs."

The idea is that an individual's ability to benefit from training depends not only on the training provided but on a set of training learning skills. As the program systematically improves ways of presenting information through curriculum development, it is equally advisable to prepare trainees to learn the information already being presented. Research conducted by DOD agencies in coordination with researchers in universities and in private industry has shown that learning strategy pretraining helps speed learners toward competency levels in technical training programs. In addition, these trainees view the learning process positively, develop greater confidence in their learning capabilites, and are more likely than others to learn that success and failure are determined by one's own effort.

Cost-effectiveness makes learning strategy pretraining an inviting prelude to many technical training programs. Mastery of the strategies then becomes an integral part of technical training as trainees use their strategies to learn technical materials. Learning accelerates as trainees become increasingly proficient in applying the new learning strategies. In other words, effective learning strategies are self-reinforcing.

It must be emphasized that learning strategy training programs are not solely designed for slow learners; on the contrary, such programs are designed to assist all students, regardless of ability, in methods of learning and study efficiency. In fact, these programs provide human learning information processing skills that are developmental and enriching. Skills learned from these programs can be applied to all academic and trade level theory courses to improve comprehension and information retention. One further distinction needs to be made; learning strategy training programs must not be confused with remedial instruction (usually manifested in disabilities in Reading and Mathematics). Remedial programs in the basic skills address individuals who have more serious learning difficulties in receiving, integrating, and expressing verbal and written information.

Method. In conducting this shipyard survey to determine if a need existed for Learning-Study Skills instruction, the following areas were investigated.

A. A survey of training administrators in all eight naval shipyards to determine whether programs existed, and if not, whether they are needed.

B. An interview conducted with 13 apprentices in Norfolk Naval Shipyard to determine the extent of learning-study skills instruction received, how, when, where, and under what conditions.

C. A review of selected findings of the Training Information Survey (1984)--a shipyardwide survey conducted to determine beliefs held by Naval shipyard training personnel about apprentice need for learning-study skills instruction.

D. A review of library and periodical literature on trade related learning and study instruction from 1978-1984.

Findings:

A. Survey of training administrators. Training administrators and Instructional Technologists in each of the eight Naval shipyards were surveyed: Six yards felt that a need for some type of study skills training exists. Two yards felt they have already satisfied any need that may have existed by conducting courses for first year apprentices. From the responses received, it appears that:

(1) Reaction to a need for study skill training is positive and a need does exist.

(2) Three yards already have some kind of a course and/or instruction in place.

B. Apprentice interviews. Thirteen apprentices were interviewed. The apprentices varied by shop and year in the program; three were in the second year and the remaining in the third year of the program.

Three had received some sort of instruction in either text outlining or notetaking.

Ten were positive there had been no instruction or training of any type provided. These apprentices were either members of informal study groups or received tutoring from their peers.

Ten were positive there had been no instruction or training of any type provided. These apprentices were either members of informal study groups or received tutoring from their peers.

One apprentice personally felt no training was needed. The comment was made that "Some people need help, others don't".

One apprentice preferred a self-paced book for learning. The others were fairly evenly divided between books, tip sheets and handouts. Four felt a combination of all types of material was best.

NOTE: Samples of each study-type medium were provided and the apprentice was given the opportunity to examine the materials.

Most of the apprentices felt the intruction would be better given at the beginning or before Academic Theory instruction. They also thought the training should be optional rather than a mandatory class.

Most apprentices said that note-taking was a difficult skill. Some instructors presented so much material so fast that it was impossible to take notes. To them, training is definitely needed. The comment was made by several that older apprentices needed more help since they had been away from school longer. Those apprentices entering training soon after high school had less trouble learning the material.

C. Training information survey. A Training Information Survey was conducted in 1984 by the Shipyard Training Modernization Program to assess attitudes and actual practices toward training among Group Superintendents, Shop Superintendents, Supervisory Training Instructors, and Trade Theory Instructors in all eight Naval shipyards. Based upon an analysis of a sample of this population (N=256) surveyed revealed the following findings about the beliefs that training personnel hold about apprentices:

(1) 79.5% of apprentices sometimes or rarely demonstrate skill in efficient learning strategies or effective study skills.

(2) 74% of apprentice hires lack basic skills in either reading, mathematics, and communication skills, or a combination of all three.

These two convincing statistics demonstrate a strong need for apprentice learning-study skills instruction. The need is based upon beliefs actually held by Naval shipyard personnel involved and familiar with the apprentice program.

D. Literature search. A study of the literature was conducted along with the investigation of external sources and expert opinion to determine the existence and availability of trade related learning-study skills material and information.

(1) A library search of learning study skills curricula materials in Research in Education (ERIC system) and Current Index to Journals in Education during the time from 1978 to present.

(2) The National Center for Research in Vocational Education.

(3) U.S. Navy Recruiting Command

The literature and external resources survey yielded the following: (1) study guide handouts on how to study, (2) research information on the effectiveness of study skills programs and (3) specific articles on the teaching of selected study skills. All the materials received and reviewed focused upon study skills in the general application sense. There were no trade specific study skills material available. Leading academic and trade resource personnel confirmed and reinforced this finding. To their collective knowledge, there exists no trade-related or trade specific learning study-skills material.

Trade-related rationale

The rationale behind using trade-specific or trade-related learning study skills material to increase apprentice school students learning efficiency rests upon the premise of educational meaningfulness. While study skills are general in nature and can be applied to any content-area, that application needs to be performed in the trade area. It is more meaningful for a student to learn notetaking skills for a lecture on welding (if he is a welder) than it is to practice that skill in a general and isolated sense. It is more important for a student to see the value in mnemonics (memory aid) applied to remembering the process steps of flushing and charging a system (Air Conditioning and Refrigeration) than it is to practice that skill in a general manner. It is also more meaningful to show an apprentice how to adjust his reading rate when reading trade related material with practice examples of trade material itself rather than practicing on general category reading matter. In sum, study skills can show their greatest application if developed within the trade-material contexts; then apprentices can develop these skills with the optimum amount of meaningfulness.

Conclusions. In view of all the data available the following conclusions are made:

A. Learning-study skills programs should be front-loaded; i.e., taught during the first year, and preferably, prior to Academic Theory.

B. Group-paced instruction is preferred by apprentices as the delivery mode.

C. Shipyards vary in their preferences of instructional delivery, ranging from formal training to module development.

D. Apprentices believe that early learning strategy training is necessary.

E. Apprentices vary in their preferences of type of instruction, ranging from textbook to handouts.

F. There is a dearth of published trade-specified or trade-related study skills material.

G. A review of the research on the value behind learning strategy training is strongly supported by the training community.

H. Attrition rate of apprentices is higher during the first year = 33% (Source: OPM Apprentice Study, 1981).

In sum, a need for Learning-Study Skills instruction has been demonstrated. Shipyard training personnel have stated through survey and questionnaire responses that apprentices can profit from such instruction. Apprentices, themselves, have indicated a need for such instruction; finally, the research literature has shown that students in the training arena and in DOD agencies who have participated in effective learning strategies have benefitted through higher achievement gains.

Recommendation. The following recommended course design has been undertaken and development is underway. That a Learning-Study Skills course be developed with the following characteristics:

- Group-paced
- Modular format
- Taught during first year
- Implemented in all Naval shipyards
- Modules to range 30-60 minutes in instructional time

Modules to be developed in each learning strategy area (4-8 hours of total instructional time) to include the following topics:

- Notetaking
- Memory strategies
- Study techniques, e.g., SQ3R
- Concentration techniques
- Listening
- Time Management
- Test-taking
- Reading rate

Each of the eight modules to follow a uniform format and to be developed using trade-related material. Example:

- Module Topic
- Pretest
- Background (theory, value, and importance of)
- Examples (actual trade passages)
- Demonstration (by instructor, as applicable)
- Practice (actual trade passages)
- Posttest

References

Diekhoff, George M., How to teach how to learn, Training, 1982, 19, 36-40.

Some Conditions Affecting Assessment of Job Requirements

Elizabeth P. Smith[1]
U.S. Army Research Institute
for the Behavioral and Social Sciences

Paul G. Rossmeissl[2]
Hay Systems, Inc.

As an adjunct to the Army Research Institute's Project A to improve the selection and classification process, research was initiated to develop and test a rating scale method to assess (Eaton, et. al., 1984) human attributes (e.g., abilities, interests, etc.) that are needed for success in a particular Military Occupational Specialty (MOS) (Smith, 1985). The work followed from the ability taxonomy and rating scale work by Fleishman and his associates (see Fleishman & Quaintance, 1984). Within Project A, a taxonomy of human attributes that affect performance was developed from expert judgments of validity (Wing, Peterson, & Hoffman, 1984). The taxonomy included 21 clusters of cognitive/perceptual, psychomotor, and noncognitive (temperament and interests) variables. Smith (1985) constructed a set of scales corresponding to 20 of these attributes plus physical strength and stamina. This set of scales, the Attribute Assessment Scale (AAS), which was designed to use work supervisors as Subject Matter Experts (SMEs), contains primarily Army-specific behavioral anchors. Several problems were uncovered during preliminary tests of the instrument with two different samples (Smith & Rossmeissl, in process). The research which is presented here attempted to address those issues. As with the earlier research, the goal was to demonstrate that the scales can produce reliable, differential profiles of attribute requirements that discriminate across MOS. These profiles then could be matched to measures of an individual's attributes for selection and classification purposes.

In the first test of the AAS (Smith, 1985), senior noncommissioned officers (NCOs) from two MOS provided ratings of the requirements for entry level work in their own MOS for three performance levels (15th, 50th, and 85th percentiles). Two types of Intraclass Correlation Coefficients (ICCs) were calculated over all attributes. The first (r_1) provides a point estimate of interrater reliability or the reliability of a single rater. The second (r_k) indicates the reliability of the mean rating. These coefficients were extremely weak. There was very little interrater agreement and at least 30 raters were needed to obtain moderately reliable means--a number higher than would be practical in operational use. An ANOVA indicated that attribute profiles for the two MOS were not significantly different.

There appeared to be three major problems related to the instrument and the research. First, the inclusion of three performance levels may have had a strong, negative impact on the results. The demands of the task appeared

---

[1]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.

[2]Affiliated with U.S. Army Research Institute at the time this research took place.

to impose a unique kind of restriction in the range of possible ratings, plus it took considerable effort. Second, the multiple levels added more confusion to a performance criterion which was already very broad -- all work within all duty positions -- allowing for considerable variance. The third problem centered on the scale anchors. This included SMEs' frustration with their content and/or difficulty in using them as reference points for evaluating duties within their MOS.

The second test of the AAS (Smith & Rossmeissl, in process) considered two of these issues. SMEs were a small number of officers and NCOs from three MOS. We provided a written job description from Army Regulation 611-201, and SMEs gave a single rating of the level of each attribute required for "average" performance of entry level work in their own MOS. An important aspect of the research was a post-rating discussion period during which SMEs provided information about problems that they had in completing the task, specific issues related to interpretation of "average" performance, confidence in their responses, and ways to improve the procedures.

With the exception of one MOS for which procedural problems were noted, the results were promising. Overall, the magnitudes of the ICCs were better than those obtained in the original research. Reliabilities of mean rating (r_k) equal to .73 and .84 with only 4 and 9 raters respectively were encouraging. ANOVA results indicated no significant differences in profiles across MOS, but given the small sample sizes this was not surprising. Our post-rating discussions indicated that use of the criterion "average" performance may have reduced MOS differences as well. Problems with this terminology included some tendency a) to describe the average soldier rather than, e.g., the average Administrative Specialist, b) to confuse average performance with average level of requirements, and c) to view average performance as actually substandard. The discussions also confirmed there were still problems related to the anchors and the ambiguity/enormity of the "whole job" criterion.

Given these outcomes, we decided to test the rating scales again under different conditions. In this research we examined ratings of attribute requirements for the whole MOS versus ratings of important, representative component tasks using two sets of scales with different anchors.

METHOD

Sample

One hundred fifty-nine NCOs from three MOS (Cannon Crewman: 13B, Light Wheel Vehicle Mechanic: 63B, and Single Channel Radio Operator: 31C) at two posts served as SMEs.

Procedure

Within MOS and posts, SMEs were assigned in blocks of 12 or less to one of 4 condition groups. Group I rated the job as a whole, using the original, behaviorally-anchored AAS. Group II rated the job as a whole, using scales with generic anchors (1=very low, 4=moderate, 7=very high). Groups III and IV rated the attribute requirements for 15 component tasks of their MOS. The tasks were those used in the hands-on testing portion of Project A. Group III used the behaviorally-anchored scales; Group IV, the generically-anchored ones. SMEs estimated the levels of the 22 attributes

which are required for "successful performance" of Skill Level 1 work for their own MOS. SMEs in the previous research favored this choice of performance criterion. In addition to the written instructions, we provided SMEs with brief training in how to use the scales to derive ratings.

Analyses

To determine reliability, we calculated ICCs (r_1 and r_k) from Attribute X Rater ANOVAs by group for each MOS. To compare reliabilities based on same sized groups, we estimated reliability of mean ratings based on 6 raters (r_6) using the Spearman-Brown formula. We performed an MOS X Attributes X Anchor (Generic vs. Behavioral) X Criterion (Whole Job vs. Tasks) univariate repeated-measures ANOVA to examine differences in profiles among MOS and any effects due to anchor or criterion conditions. The single, highest rating assigned to any task within each attribute was used in the ANOVA.

Results

The ICCs (r_1, r_k, r_6) for the four conditions by MOS are given in Table 1. Overall, estimates of interrater agreement are low. The best r_1s are for Radio Operators across all 4 conditions, yet there still are large between-subjects variances for all MOS. Across MOS, no particular condition yielded higher r_1s or r_ks than another.

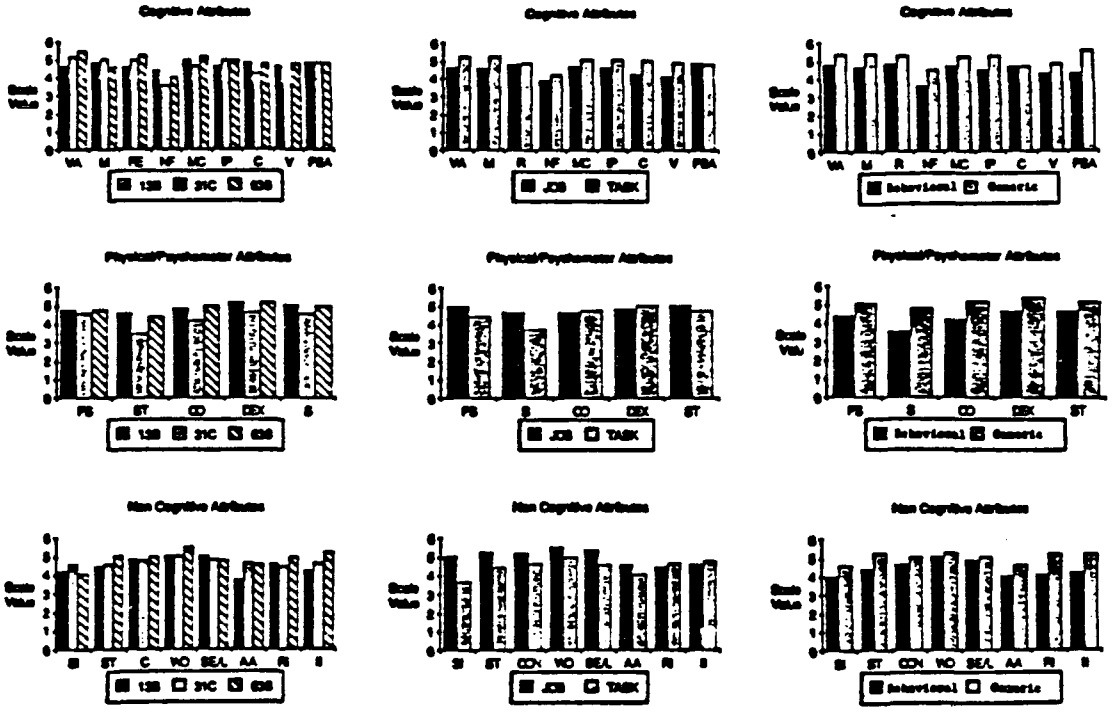
Table 1

Reliability estimates for a single rater, mean of k raters, and mean of six raters of three MOS by experimental conditions

MOS	Anchor Type	Criterion	k	r_1	r_k	r_6
Cannon Crewman	Behavioral	Task	19	.08	.63	.34
		Job	12	.22	.77	.63
	Generic	Task	25	.07	.67	.31
		Job	12	.04	.36	.20
Radio Operator	Behavioral	Task	9	.28	.77	.70
		Job	12	.22	.77	.63
	Generic	Task	12	.17	.71	.55
		Job	17	.19	.80	.58
Mechanic	Behavioral	Task	22	.11	.74	.43
		Job	7	.12	.48	.45
	Generic	Task	6	.07	.32	.32
		Job	6	.21	.61	.61

The MOS X Attribute X Anchor X Criterion ANOVA indicated there are significant differences in attribute profiles across MOS and that these differences were affected by the experimental conditions. Although the 4-way interaction is not significant, two 3-way interactions (Attribute X MOS X Anchor and Attribute X MOS X Criterion) and all 2-way interactions involving attribute are significant with a Geisser-Greenhouse $p<.05$. That is, mean

ratings vary across attributes within and across MOS. Means also differ as a function of the type of anchors or the criterion. Collapsing over type of criterion, generically-anchored scales yielded higher mean ratings for all attributes. On the other hand, the effects of criterion condition (job vs. tasks) were dependent on the type of attribute. For the most part, across MOS, we found higher means for evaluations of the whole job for the cognitive/perceptual attributes, some of the psychomotor attributes, and of the noncognitive attributes, realistic and investigative interests. The opposite was true for physical strength, stamina, and the other noncognitive (temperament) attributes. Figures 1(a-c) graphically depict the three 2-way interactions.



a b c

Attributes:

Cognitive		Physical/Psychomotor		Non-Cognitive	
VA	Verbal Ability	PS	Physical Strength	SI	Social Interaction
M	Memory	S	Stamina	ST	Stress Tolerance
R	Reasoning	CO	Coordination	CON	Conscientiousness
NF	Number Facility	DEX	Dexterity	WO	Work Orientation
MC	Mechanical Comprehension	ST	Steadiness	SE/L	Self-Esteem Leadership
IP	Information Processing			AA	Athletic Ability
C	Closure			RI	Realistic Interests
V	Visualization			II	Investigative Interests
PSA	Perceptual Speed & Accuracy				

Figure 1. Comparison of profiles of attribute means by a MOS, b Criterion (Job vs. Task), and c Anchor (Behavioral vs. Generic).

DISCUSSION

As with initial tests of the AAS, the interrater agreement found here is relatively low. For most purposes, however, we are more interested in the reliability of the mean ratings which were moderate for most of the conditions. Use of generically-anchored scales did not improve the reliabilities as our previous research had suggested, but the behaviorally anchored scales were no more reliable than the generic anchors. In effect, however, using behavioral anchors tended to lower mean ratings, perhaps by reducing a "more means better" tendency toward inflating estimates of requirements for good performance. These findings suggest that in similar situations the impact of using behavioral based anchors may not merit their increased developmental effort and cost.

Similarly, to the degree it was tested here, having SMEs rate components of the job did not increase agreement among raters either. In our analyses we used only one of the 15 ratings made by SMEs in the task rating conditions. Perhaps we would find better interrater reliability if we focused on each task individually. The choice of criterion did affect magnitude of ratings, but not in the same way for all attributes. Differences in means, as well as lack of agreement among raters, may well have been a function of the comprehensiveness or representativeness of the tasks. Some SMEs argued that the specific tasks we used required little or none of some attributes (especially temperament attributes), but that these attributes are required for other aspects of the job. A few SMEs indicated they gave high ratings on the tasks for this reason, thus ignoring our instructions to rate only the 15 tasks provided.

Although we were unable to increase reliability by altering the conditions of the administration of the AAS, the data were sufficiently reliable to yield meaningful results. The key interaction of MOS and attribute was statistically significant: We did attain significantly different requirements profiles across MOS mean ratings. Also significant were the comparisons investigating the effects of anchor type and level of analysis (job versus task). In other words, while the reliabilities were low, they were sufficient to provide valuable information. Given this and the other findings, the AAS, while not producing results which advocate its use for selection and classification purposes, still may have some potential. For example, it may be useful for identification of the two-three top high-driver attributes for an MOS, or for evaluation of a narrowly defined task, such as a particular kind of mission. Our debriefings with SMEs lead us to believe that any future use of the AAS or similar kinds of instruments really should involve an intensive training session. SMEs should be given thorough explanations, with examples, of what the attributes entail and helped to see how they relate to various aspects of the job.

REFERENCES

Eaton, N. K., Goer, M. H., Haris, J. H., & Zook, L. M. (October, 1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1984 Fiscal Year. (Technical Report No. 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Fleishman, E. A., & Quaintance, M. K. (1984). Taxonomies of human performance. Orlando, FL: Academic Press Inc.

Smith, E. P. (1985, November). Developing new attribute requirements scales for military jobs. Proceedings of the 27th Annual Conference of the Military Testing Association, San Diego, CA.

Smith, E. P., & Rossmeissl, P. G. (in process). Attribute assessment: Initial test of scales for determining human requirements of military jobs. Technical Report. U.S. Army Research Institute for the Behavioral Sciences, Alexandria, VA.

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.

COMPUTERIZED MEASUREMENT OF VIGILANCE ABILITIES

Charles H. Cory[1]
Navy Personnel Research and Development Center

Background

A major area of promise for the use of computers in personnel classification testing is for expanding the number of abilities which can be measured: for measuring abilities which are not measurable by paper-and-pencil tests. One such ability is vigilance. Vigilance--the ability to remain alert even in very boring situations--is obviously important for military jobs which require watch-standing. Although the vigilance phenomenon has been studied extensively, tests which measure vigilance and are suitable for personnel selection/classification have not been developed. The use of computer terminals as personnel testing instruments permits vigilance to be measured more easily than was formerly the case. This paper describes exploratory research for the evaluation of VIG, an experimental test of vigilance which is administered on an IBM PC.

Vigilance was defined by Mackworth (1957) as "a state of readiness to detect and respond to certain specified small changes occurring at random time intervals in the environment." Previous research (Mackworth, d330) has found this state of readiness to decrease over time and under conditions of boredom. The classic vigilance research has used a clock watching task in which the task is to detect jumps in the sweep second hand. In other research, vigilance tasks have involved the detection of variations in light intensities and in auditory pitches.

In order to facilitate the manifestation of performance decrements over time, vigilance tests have required long administration times. Mackworth tested her subjects for two hours; others such as Dobbins, Tiedemann, and Skordahl (1961) tested subjects for seven and a half hours. These time limits were clearly too long to be feasible for Navy classification tests.

This paper describes results from administration of VIG, a computerized test of vigilance which is being evaluated for supplementing the ASVAB for classifying Sonar Technicians. Personnel in Sonar Technician jobs operate sonar equipment to detect and classify enemy targets. A major part of the job involves scanning CRT screens which reflect the images of sonar sweeps--sorts of surrealistic images which are continually changing. Sweeps show filmy cloudlike images periodically interspersed with regularities which indicate targets. The test was designed to be an analogue of the Sonar Technician detection task in that it requires detection of targets presented at infrequent intervals within a background of heavy noise. The challenge of this research is to produce a vigilance test which has both validity and face validity for the sonar detection task and which has a short enough time limit to permit its use in personnel classification.

---

[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.

Description of VIG

VIG emits a series of randomly placed upper case letters across the screen. If the example (Figure 1) were being displayed, the letters would be emitted rapidly and would be scrolling up one line at a time. At random intervals a lower case letter would be emitted. The task is to respond by pressing a key on the computer as soon as a lower case letter is detected. In Figure 1 the target letter is the small "w" center right in the distribution. Total time for the test is 26 minutes, during which 18 target letters are displayed, approximately one every 87 seconds. The sample consisted of 234 enlisted personnel who were administered VIG at the beginning of "A" school, and who had taken the ASVAB prior to enlisting in the Navy.

Measures collected for VIG were number of targets detected(VGTR), number of targets for which false positive responses were made(VGTW), and mean detection latencies for targets correctly detected (VGMLC). For each variable totals/means were calculated for the first and second halves (each based on nine targets) and a difference score (VGMRD, VGMWD, or VGMLD) was calculated by subtracting the statistic for the second half from that for the first.

Analysis

T tests were calculated for the difference scores and the results are shown in Table 1. They show that for the second half of VIG on average about 8 percent fewer targets were detected, detection latency increased about 33 percent, and about 50 percent fewer false positives were emitted. All findings are statistically significant and are consistent with findings reported in the literature for vigilance tests (Mackworth, 1970).

The vigilance effects shown in Table 1 were determined from group data: they are based on differences between means. But the value of vigilance as a predictor depends on its characteristics for individual cases. It depends upon the distribution characteristics of vigilance variables and on the answers to such questions as "Does the vigilance effect occur because a few individuals have large decrements, or because many individuals have small decrements, or because of some combination of these conditions?" Therefore, the inter-relationships of vigilance effects within individuals were studied. For ease of communication a vigilance effect will be defined as a score on VGMRD, VGMWD, or VGMLD in the same direction as the differences shown in Table 1.

Table 2 shows a tabulation of the binary patterns of vigilance effects formed by coding 1/0 for the presence/absence of vigilance effects. Table 2 shows that 92.1% of the sample had vigilance effects on one or more predictors, 64.9% had vigilance effects on two or more predictors and 17.5% had vigilance effects on all three predictors. Looked at another way, 53.2% of the sample had a vigilance effect for rights, either alone or in combination with other criteria; 49.8 an

effect for false positives; and 71.5% an effect for latency to rights. Thus at least some form of vigilance effect occurred for 92.1% of the sample, and the most common vigilance effect was an increase in detection latency (i. e., a decrease in speed of detection).

The bivariate distribution of total latencies and latency differences for VIG, shown in Figure 2, shows another aspect of the vigilance relationships. The distribution shown in this figure is shaped like an 85 degree angle which has been rotated 45 degrees and its point placed at 0 on the latency difference scale. Greater latency differences on the ordinate are associated with greater total latencies on the abscissa. Persons in the lower part of the angle are exhibiting the vigilance effect: their average detection time is slower for the second half than for the first half. But the interesting aspect of Figure 1 is the cases in the upper part. These persons performed counter to the vigilance effect--their performance improved in the second half. If vigilance is an important selection variable, persons in the upper part of the distribution are the ones whom we will be selecting. Difference scores for rights and for false positives (not shown) also exhibited these effects. We examined the relationships of the three difference scores on the only criterion available during this early stage of the research: "A" school final grade. The correlations of all three vigilance difference scores with this criterion were very small and were not statistically significant.

In order to examine the relationships among vigilance and the ASVAB subtests, a principal components analysis was performed on a matrix of intercorrelations among the 10 ASVAB and the six vigilance scores. Extraction was stopped when the eigenvalue fell below 1.00. The factor loading matrix resulting from a varimax rotation of these components is shown in Table 3. The six factors extracted accounted for approximately 64% of the variance. The first four components in Table 3 are essentially the same factors that have been reported from previous factorings of the ASVAB (Moreno et al., 1984; Ree et al., 1982). Component 1 is technical knowledge, defined by significant loadings on the Auto Shop, Mechanical Comprehension, and Electronics Information subtests of the ASVAB. Component 2 is the ASVAB verbal component, defined by the Paragraph Comprehension, Word Knowledge, and General Science scores. General Science also loads significantly on technical knowledge--a finding which is consistent with previous research. These findings indicate that the vocabulary and sentence structure complexities of GS result in its being as much a measure of verbal skills as it is a measure of technical knowledge. Component 3 is the speed component of the ASVAB (Coding Speed and Numerical Operations) plus VGTR, the total rights score for VIG. VGTR is not considered a measure of perceptual speed as are CS and NO, but the perceptual ability measured by VGTR must be closely related to perceptual speed. Component 4, defined by high loadings on Math Knowledge and Arithmetic Reasoning, is the ASVAB mathematics factor. The two remaining components are defined exclusively by scores from VIG. Component 5 is defined by vigilance effects for rights and for false positives, plus the total false positives score. This is the closest

that the data come to defining a factor which combines different elements in the vigilance response. Component 6 measures speed of target detection. It is defined by the total latency and the latency effect scores.

From the standpoint of clarity of interpretation, it is disappointing that vigilance abilities encompass two factors, rather than one. Clearly the most parsimonious interpretation of the three effects is that they are manifestations of a single syndrome and therefore should load on a single factor. The fact that VIG variables defined two factors indicates that their inter-relationships are more complex than was presumed.

Summary

In summary, VIG, a computerized test designed to measure the vigilance abilities of personnel being considered for assignment as Sonar Technicians, was developed. Although administration time for VIG was much shorter than that of vigilance tests previously reported in the literature, the vigilance effects found were similar to those previously reported for vigilance tests. They were: decrease over time in number of targets detected, false positives, and in speed of target detection. Because of its relatively short administration time, VIG could be used for personnel classification testing, whereas the tests used in previous research could not. Vigilance effects in some degree occurred for 92.1% of the cases. The most common vigilance effect was a decrease in average speed of target detection. Principal component analysis of the ASVAB-VIG intercorrelation matrix found the abilities measured by VIG to be relatively independent of those measured by ASVAB and to consist of two separate abilities.

REFERENCES

Dobbins, D. A., Tiedemann, J. G., & Skordahl, D. M. (1961). Field study of vigilance under highway driving conditions. A.P.R.O. Technical Research Note 118.

Mackworth, J. F. (1970) Vigilance and Attention. Middlesex, England: Penguin Books Ltd.

Mackworth, N. H. (1957). Vigilance. The Advancement of Science 53, 389-393.

Moreno, K. E., Wetzel, C. D., McBride, J. R. & Weiss D. J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. Applied Psychological Measurement, 8, 155-163.

Ree, M. J.,Mullins, C. J., Mathews, J. J., & Massey, R. H. (1982). Armed Forces Vocational Aptitude Battery: Item and factor analysis

of forms 8, 9, and 10(Technical Report 81-55). Brooks Air Force Base TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

VIGILANCE

F G E U F M Y K L K A Y
W R H J I S S W H F S W
E A G R E E S A A w S D T T T
R T Y YU U E E D O F J J LL
FJ i E D E E E I F F L D
Q Z ADR Y U U I PO P D W S D F H

MEASURES: PERCEPTUAL SPEED
VIGILANCE

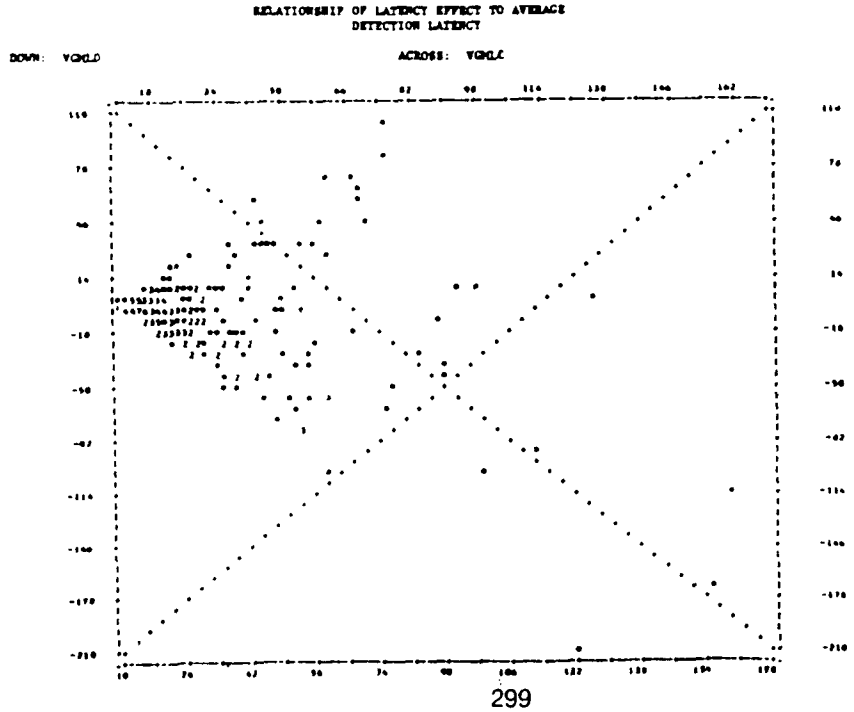


Table 1

T-Tests of Mean Pairwise Differences of Rights, Wrongs, and Latencies

Variable	N	M	- Difference	T Value	2-tail Probability
Total Right, 1st Half	234	5.71	.70	4.91	0.0
Total Right, 2nd Half		5.01			
Total Wrong, 1st Half	234	1.62	.76	7.75	0.0
Total Wrong, 2nd Half		.86			
Mean Latency, 1st Half	228	3.0	-1.0	-4.64	0.0
Mean Latency, 2nd Half		4.0			

Table 2

Percent of Sample with Vigilance Effects

Vigilance Effect	%
RWL	17.5
RW Only	11.4
RL Only	20.3
WL Only	15.7
R Only	4.0
W Only	5.2
L Only	18.0
None	7.8
Total	99.9

Table 3

Principal Components Analysis of ASVAB-VIG Scores

Variable	Component 1	2	3	4	5	6	Communality
AUTO SHOP	[81]	16	-01	-08	-03	-01	69
MECH COMP	[76]	11	06	19	-03	17	66
ELEC INFO	[65]	29	-11	-16	01	-16	57
PARA COMP	13	[76]	03	05	04	-08	61
WORD KNOW	20	[74]	15	-02	-05	10	63
GEN SCI	41	[52]	-13	14	-02	-14	50
CODNG SPD	-03	14	[78]	10	-02	-11	65
NUM OPER	-21	02	[75]	28	-06	-18	72
VGTR	23	-09	[61]	-16	05	35	56
MK	-08	-01	07	[84]	-13	02	74
AR	07	12	13	[79]	07	03	67
VGTW	-01	-05	07	-14	[83]	-13	73
VGMWD	-30	27	-04	-08	[68]	27	71
VGMRD	17	-17	-12	30	[51]	-13	44
VGMLC	06	-01	-07	-07	28	[77]	68
VGMLD	06	-06	-16	-01	14	[76]	63
Eigenvalue	2.09	1.64	1.63	1.62	1.74	1.49	10.2
% Variance	13.1	10.3	10.2	10.1	10.9	9.3	63.9

Varimax Rotation: Extraction stopped at Eigenvalue = 1.

APTITUDE SELECTORS FOR AIR FORCE OFFICER NON-AIRCREW JOBS

1Lt Thomas O. Arth and M. Jacobina Skinner
Air Force Human Resources Laboratory

I. INTRODUCTION

The Air Force Officer Qualifying Test (AFOQT) has been used since 1951 to select individuals for commissioning as Air Force officers and for selection of pilots and navigators. Although studies by Finegold and Rogers (1985) and Arth (1986) have shown that scores on the AFOQT predict success in non-aircrew technical training courses, the AFOQT is not used to classify officers into non-aircrew specialties. This research was undertaken to determine the potential utility of the AFOQT for differential prediction of success in non-aircrew jobs.

The predictive utility of AFOQT Form O, the test version in operational use since September 1981, was evaluated. AFOQT Form O is a paper-and-pencil instrument containing 380 multiple-choice test items distributed among 16 subtests. The subtests are designed to assess verbal, quantitative, spatial, and specialized ability areas. All subtests are defined as power tests, but completion rates among examinees in the standardization sample suggest that the majority of the subtests have a speeded component (Rogers, Roach, & Wegner, 1986). Administration time for the entire battery is about 4.5 hours.

Five composites - Verbal (V), Quantitative (Q), Academic-Aptitude (AA), Pilot (P), and Navigator-Technical (N-T) - are derived from combinations of the subtests. The V and Q composites each contain three subtests. These composites and subtests will be referred to as non-aircrew composites and subtests hereafter. They are then combined to form the AA composite. The remaining subtests are used exclusively in either or both the P and N-T composites, which will be referred to as the aircrew composites. Additionally, the N-T composite incorporates the Q subtests, and the P composite include the Verbal Analogies subtest found also in the V composite.

Form O contains many features in common with earlier forms of the AFOQT. A complete description of the development and standardization of Form O, together with a comparison with its predecessor (Form N), is available in Rogers, Roach, and Wegner (1986). Information on earlier forms and the history of the officer aptitude testing program in the Air Force has also been reported (Gould, 1978; Rogers, Roach, and Short, 1986).

As their names imply, the P and N-T composites are used to select candidates for pilot and navigator training through the use of minimum scores. The V and Q composites are used to select individuals into the Air Force Reserve Officer Training Corps (AFROTC) and Officer Training School (OTS) also by using minimum scores. However, the research cited above found in certain instances that the P and N-T composites had higher zero order correlations than the other composites with success in non-aircrew courses. This finding would seem to imply that some of the subtests used for the aircrew composites (P and N-T) could be used to predict success in non-aircrew courses.

The effectiveness of the subtests used in the P and N-T composites was shown by Hunter and Thompson (1978) for pilots and by Valentine (1977) for navigators. This type of analysis has not been done for any non-aircrew Air Force specialty (AFS). Prior to AFOQT-O, only individuals applying for pilot or navigator training were required to test on all parts of the AFOQT.

However, all individuals currently applying for a commission are administered the entire AFOQT. This change in administration procedure provides the first opportunity to assess the predictive validity of the subtests and to evaluate the potential of using unique combinations of the subtests beyond the current use for pilot and navigator specialties.

II. METHOD

Subjects for the study were 1,025 active duty officers assigned to eight AFS requiring entry-level technical training. Course identification numbers, sample sizes, and titles are shown in Table 1 below. Only OTS subjects were used due to the unavailability of AFROTC AFOQT-0 scores.

Table 1. Officer Technical Training Course Samples

Course Number	Title	N
1741X	Air Weapons Controller Fundamentals	147
2001	Space Environment and Operations	100
3051	Communications and Electronics Engineer	156
4021	Aircraft Maintenance Officers	215
4051A	Munitions Officers	81
7000	Administration Officers	126
8000	Fundamentals of Intelligence	122
8121	Security Police Officer	78

Subjects were identified from historical data files containing technical training and AFOQT records on Air Force officer applicants and commissionees maintained by the Technical Services Division, Air Force Human Resources Laboratory. The sample was restricted to OTS commissionees who had tested on AFOQT Form 0 prior to entering service and whose technical training records reflected successful completion of training with a valid final school grade between September 1981 and June 1985. A total of 2,025 officers in 58 different technical training schools were identified. Of these, the eight courses shown in Table 1, which had 75 or more cases, the minimum number judged to be sufficient for planned analyses, were selected for the final study.

The demographic composition of the eight training courses by race and gender was mixed with Caucasians and males forming the majority groups. About 82 to 90 percent of the students in each course were Caucasian and about 63 to 93 percent were male. The largest groups of female students were assigned to Administration (37%) and Intelligence (26%) job specialties.

The primary predictor variables were scores obtained on AFOQT Form 0 subtests and composites. Subtest scores were "number right scores" indicating the number of items answered correctly. For each of the five composites, scores reported in both raw (total number of correct answers for subtest groupings) and percentile (01 - 99) form were used as predictors.

Final grade earned in each training course was used as the performance criterion. The grade reflects academic achievement level in training and is reported in percentages ranging from 60 to 99. Grades were available only on course graduates.

Table 2. Validity Coefficients (Uncorrected) for AFOQT Aptitude Scores and Final Grade in Eight Courses

AFOQT Predictor	Course							
	Air Weapons Controller	Space Env & Ops	Comm & Elec Engineer	Aircraft Maint	Munitions	Admin	Intelligence	Security Police
Non-aircrew subtests								
Verbal Analogies	.31**	.15	.14	.18**	.10	.29**	.27**	.31**
Arithmetic Reasoning	.20*	.30**	.25*	.28**	.19	.25**	.33**	.20
Reading Comprehension	.24**	.35**	.13	.20**	.19	.33**	.29**	.42**
Data Interpretation	.22**	.25*	.14	.26**	.17	.21*	.22*	.31**
Word Knowledge	.20*	.22*	.12	.13	.14	.24**	.24**	.37**
Math Knowledge	.29**	.22*	.11	.21**	.23*	.34**	.30**	.29*
Aircrew subtests								
Mechanical Comprehension	.20*	.19	-.02	.28**	.39**	.11	.23**	.29*
Electrical Maze	.10	.20*	-.18*	.21**	.15	.20*	.10	.07
Scale Reading	.24**	.22*	-.02	.21**	.22	.20*	.26**	.27*
Instrument Comprehension	.15	.24*	.01	.17*	.29**	.07	.15	-.04
Block Counting	.09	.10	-.14	.20**	.18	.06	.15	.18
Table Reading	.21**	.05	.02	.19**	.02	.25**	.26**	.19
Aviation Information	.11	.28**	-.02	.10	.20	-.08	.20*	.00
Rotated Blocks	.05	.08	-.06	.27**	.17	.02	.18*	.21
General Science	.13	.37**	.04	.11	.40**	.15	.12	.20
Hidden Figures	.20*	.23*	-.03	.25**	.30**	-.09	.20*	.09
Verbal Raw Score	.28**	.28**	.15	.19**	.17	.32**	.30**	.42**
Quantitative Raw Score	.29**	.29**	.20*	.29**	.24*	.32**	.35**	.31**
Academic-Aptitude Raw Score	.34**	.33**	.19*	.27**	.24*	.36**	.38**	.41**
Pilot Raw Score	.29**	.26**	-.05	.30**	.30**	.22*	.32**	.25*
Navigator-Technical Raw Score	.29**	.29**	.01	.33**	.33**	.26**	.35**	.33**
Verbal Percentile	.27**	.28**	.15	.18**	.17	.32**	.31**	.41**
Quantitative Percentile	.27**	.30**	.19*	.27**	.23*	.32**	.35**	.27*
Academic-Aptitude Percentile	.33**	.28**	.19*	.26**	.24*	.36**	.38**	.39**
Pilot Percentile	.29**	.26**	.06	.29**	.31**	.23**	.32**	.23*
Navigator-Technical Percentile	.29**	.30**	.00	.32**	.33**	.27**	.36**	.31**

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

Relationships between final grade and each AFOQT measure were evaluated using the Pearson product-moment correlation coefficient (r). Additional statistics were obtained from multiple regression analyses. The significance levels of the R^2 differences were computed by F-tests.

III. RESULTS

Bivariate validities for AFOQT aptitude predictors and final grades are shown for each course in Table 2. The majority of the correlation coefficients were positive and significant. This trend was more clearly apparent for the aptitude composites than for individual subtests. Coefficients ranged from -.18 to .42, with the majority between .20 and .35.

Inspection of coefficients for subtests in the Verbal and Quantitative composites, the first six subtests listed in Table 2, revealed that most related significantly to training performance in six of the eight courses. The Communications and Electronics Engineers course and Munitions course were exceptions. Only one of the subtests in the Q composite, Arithmetic Reasoning in the former course and Math Knowledge in the latter, predicted final grade.

Additional significant validites were found among the ten aircrew subtests used in the P and N-T composites. In seven courses, positive relationships were obtained between two or more subtests and academic outcome in training. The highest number of significant aircrew subtests was eight in the Aircraft Maintenance course, followed by Space Environment and Operations and Intelligence courses with six each. Findings for the Communications and Electronics Engineers course merit discussion. Only the Electrical Maze subtest correlated significantly with course grade, and the direction of the aptitude-performance relationship was negative.

Correlations for AFOQT composites were significant for the most part. In most courses (six of eight) all five aptitude composites predicted final grade in training. As a group the composites were least predictive in the Communication and Electronics Engineers course; only the Q and AA composites were significant. Results were highly consistent for the composites expressed on both raw and percentile scales. This finding was not unexpected; students' relative ability standing on the raw score scale is retained in the conversion to the percentile scale.

Results of the multiple linear regression analyses are shown in Table 3. The R^2 of the total sample for each predictor set is given in the table. F tests were computed among the models to test for levels of significance.

When the restricted model 2 was compared with the full model 1, no significant difference was found. However, a significant difference did exist between full model 5 and restricted model 6. This indicated that the AFOQT could not differentially predict performance among the eight courses by using the six non-aircrew subtests but could do so by using all sixteen subtests. Confirmation of these results was uncovered when a significant difference was found between models 5 and 1. That difference showed the addition of the ten aircrew subtests added substantial predictive power to the AFOQT.

Other comparisons on the composite level replicated these findings. No significant difference existed between models 3 and 4 which meant using the non-aircrew composites alone could not differentially predict performance in the eight courses. Furthermore, the addition of the aircrew composites to the non-aircrew composites was able to account for a significantly greater

amount of the variability than the use of the non-aircrew composites alone (full model 7 vs. restricted model 3). All models' R^2 were significantly greater than that of model 8.

Table 3. R^2 of Regression Models

Model	Predictor Sets	R^2
1	Group membership*, 6 non-aircrew subtests, group membership by 6 non-aircrew subtest interactions.	.37
2	Group membership, 6 non-aircrew subtests	.34
3	Group membership, 2 non-aircrew composites, group membership by composite interactions.	.35
4	Group membership, 2 non-aircrew composites.	.34
5	Group membership, 16 subtests, group membership by 16 subtest interactions.	.46
6	Group membership, 16 subtests.	.35
7	Group membership, 4 composites**, group membership by 4 composite interactions.	.38
8	Group membership.	.27

* Group membership is a dichotomous variable referring to membership in one of the eight technical training courses.
** The AA composite was not used in these models as it is redundant with the V and Q composites.

IV. DISCUSSION AND CONCLUSION

The AFOQT was found to be a valid indicator of success in officer technical training and to have potential as a differential prediction instrument for non-aircrew jobs. The predictability of academic achievement in entry-level technical training by non-aircrew composites and subtests was clearly demonstrated. Results on the V and Q composites affirm earlier findings by Finegold and Rogers (1985) and Arth (1986). More importantly, by capitalizing on the availability of subtest scores for officer commissionees - data not readily accessible for research purposes until the implementation of AFOQT Form O - it was possible to demonstrate empirically that the constituent six subtests are valid as well. Together, the results lend support to the use of V and Q composites as aptitude selectors for officer training programs.

Additionally, validites of aircrew aptitudes at the composite, and especially at the subtest level, point to the opportunity to improve the selection of Air Force officers for non-aircrew jobs. Aptitudes assessed by the AFOQT and currently used exclusively for pilot and navigator selection were also shown to be valid indicators of academic achievement level in training for non-aircrew jobs. The differential prediction capability of non-aircrew and aircrew subtests in combination for non-aircrew jobs indicate that the potential exists to reconfigure subtests in unique combinations to optimize preditability of job training success in different AFSs.

Continuation of this research stream is recommended. The current study must be viewed as exploratory. Its scope was too narrow and the sample sizes too small in most courses to defend global and definitive statements

about the utility of the AFOQT for job classification decisions about non-aircrew jobs. Nonetheless, the results are encouraging and additional research to verify and extend the findings to other AFSs is warranted.

V. REFERENCES

Arth, T. (1986) Validation of the AFOQT for non-rated officers (AFHRL-TP-85-50) Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Finegold, L. & Rogers, D. (1985). Relationship between Air Force Officer Qualifying Test scores and success in air weapons controller training (AFHRL-TR-85-13). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Gould, R. (1978). Air Force Officer Qualifying Test Form N: Development and standardization (AFHRL-TR-78-43). Brooks AFB TX: Personnel Research Division, Air Force Human Resources Laboratory.

Hunter, D. & Thompson, N. (1978). Pilot selection system development (AFHRL-TR-78-33). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rogers, D., Roach, B., & Short, L. (1986). Mental ability testing in the selection of Air Force officers: A brief historical overview (AFHRL-TP-86-23). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rogers, D., Roach, B., & Wegner, T. (1986). Air Force Officer Qualifying Test Form O: Development and standardization (AFHRL-TR-86-24) Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Valentine, L. (1977). Navigator-observer selection research: Development of new Air Force Officer Qualifying Test navigator-technical composite (AFHRL-TR-77-36) Brooks AFB, Tx: Personnel Research Division, Air Force Human Resources Laboratory.

MILITARY TESTING ASSOCIATION
PAPER

TROUBLESHOOTING PROFICIENCY EVALUATION PROJECT FOR THE NATO SEASPARROW SURFACE MISSILE SYSTEM

Presented by
Dr. Harry B. Conner
Navy Personnel Research and Development Center

PROBLEM/BACKGROUND

The Navy faced the problem of being able to objectively measure the technical proficiency of the troubleshooting technician and his ability to contribute to operational readiness. There was no way to evaluate the success of technical training or the effect of hands-on training in the Navy "C" schools i.e., System Training. As a result of this need the Troubleshooting Proficiency Evaluation Project (TPEP) initiated development of a computer-based troubleshooting proficiency evaluation test to measure the technical troubleshooting proficiency of system technicians on the NATO Seasparrow Surface Missile System (NSSMS).

OBJECTIVE/DESCRIPTION

As part of this feasibility study/research effort, the objective was to design, develop, implement, test and evaluate TPEP as an alternative to the present "C" School troubleshooting, training and evaluation technique. The "C" school technique is characterized by group training/evaluation and test administrator involvement. The major objective of the TPEP effort was to evaluate EPICS (Enlisted Personnel Individualized Career System) personnel. These individuals were compared to the Conventional Personnel System (CPS) NSSMS graduates at the three points over a growth period of one year. They were also compared to experienced conventionally trained NSSMS technicians with 1.5 to 4.5 years of on-the-job experience. All three groups were evaluated on the following measures captured during the fault diagnosis exercises: identification of the fault, time required, number of incorrect components replaced, proof of faulty component, test points relevant and irrelevant to diagnosis, measurement relevant and irrelevant to diagnosis, and illogical approaches with the given symptoms. Administration/performance of each troubleshooting problem consisted of receiving the symptoms from the computer program, entering the reference designation of the test points desired, making appropriate measurements, and finding the solution by identifying the faulty component down to the lowest replaceable unit or part (as directed by the system maintenance philosophy).

STUDY QUESTIONS

1) Are EPICS System Technician Training (STT) graduates, FT eligible and ineligible, as proficient in troubleshooting NSSMS faults as NSSMS C-school graduates at three points in time; a) completion of NSSMS school, b) four to eight months after graduation, and c) 11 to 12 months after graduation?

2) Does EPICS STT graduate and C-school graduate troubleshooting proficiency improve at a similar rate over time?

3) How do EPICS STT graduates and C-school graduates compare in troubleshooting proficiency with NSSMS technicians having greater than 18 months on-the-job experience?

IMPLEMENTATION

The Kaypro II micro-computer was originally selected for and utilized in the initial testing (current RDT&E program utilizes IBM-PC and Zenith-248s). NSSMS System troubleshooting scenarios and supporting software were developed and verified by NPRDC staff and NSSMS System Technician Training (STT) course and "C" school instructors. The subjects under test were EPICS STT school graduates (categorized as eligible or ineligible based on Navy selection factors), NSSMS school graduates, and NSSMS experienced technicians. All groups are assigned to Spruance-class destroyers, aircraft carriers, and auxiliary ships in the Atlantic and Pacific fleets which had the NSSMS aboard.

ANALYSIS/RESULTS

Demographics. Demographic characteristics of the comparison groups were investigated and the results of the investigation of the paygrade of the personnel in the test across the three testing episodes showed that the EPICS and CPS Inexperienced personnel were about the same paygrades but both of these groups were less than the CPS experienced personnel (as would be expected).

Time in service for each comparison group was also gathered and showed that the time in service for the EPICS and Inexperienced personnel was about the same (with EPICS being slightly greater) and time in service for these groups was less than tne experienced personnel (also, as would be expected).

Also school attended was determined and results showed that almost all of the EPICS and Inexperienced personnel attended the same school (CSTSC, Mare Island) whereas the Experienced group was split across two school locations (Mare Island and Dam Neck).

As can be seen from the previous presentation the paygrades of the groups were as expected i.e., the EPICS and NSSMS inexperienced were relatively the same but less than the experienced at graduation with the difference decreasing over the time of the test as promotions occurred; the time in service of the EPICS personnel was greater than the inexperienced and less than the experienced, both differences expected, due to the design of the EPICS career plan; most of the subjects attended the Mare Island course, except for the NSSMS experienced who were split equally between the two schooling options. What differences there were in the demographic makeup of the groups when inspected, did not indicate a confounding factor in the resultant analyses.

Troubleshooting Proficiency. Correlations were computed between the performance factors for each test group and scenario. This indicates whether a group performed the fault diagnosis process with a consistently different method. It also identifies those individual performance factors which are accounting for essentially the same proficiency score variance. This (not shown here) indicated there was no reliable difference in subject results in their approach.

Two independent variables were manipulated. The first, group membership, refers to the two EPICS subgroups (eligible and ineligible) and the two NSSMS groups (inexperienced and experienced). The second variable, consisting of the three measurement points is repeated on all the test groups. The dependent variables to be analyzed included the composite proficiency score and a System Proficiency score. This latter score is the average of an individual's proficiency scores across all nine scenarios. The individual scores of each test group was determined and group

means were computed.

The scores for each group were accumulated for each set of tests (i.e., test scenario 1, 2, & 3 = set 1; 4, 5, & 6 = set 2; 7, 8, & 9 = set 3; 1-9 = composite) and weighted means were determined for each testing episode.

Table 1
WEIGHTED Z SCORE BY TEST SET

Groups	(Set #1)		(Set #2)		(Set #3)		Composite	
EPICS	N	$\bar{X}$	N	$\bar{X}$	N	$\bar{X}$	N	$\bar{X}$
Eligible	39	.05	28	.12	21	.09	88	.09
Ineligible	25	-.36*	14	-.24	12	-.14	51	-.05
NSSMS								
Experienced	11	.07	10	.03	8	.20	29	.04
Inexperienced	61	.10	44	-.04	38	-.07	148	-.08

*Mean is reliably different from both EPICS eligible and ineligible and NSSMS experienced at the .01 level

To address Evaluation Questions 1 & 2, an analysis of variance (ANOVA) was computed on a 3 X 3 mixed design. Table 1 presents the data utilized for the analysis. As can be seen from the Table there were statistically reliable differences in the performances of the EPICS Ineligible as compared to the other two groups of EPICS Eligibles and Conventional Personnel (i.e., NSSMS inexperienced personnel), at the first testing episode (at the end of training i.e., graduation). There were however no reliable differences over the next two testing, or composite episodes. A time versus performance chart was plotted of the weighted Z scores as presented in Table 2, the result is shown in Figure 1.

Testing for a main effect for Groups determined if they were consistently different. Testing for a main effect for Measurement Points determined if there was a growth in troubleshooting skills. A significant interaction term indicates a differential growth in troubleshooting skills between groups. Multiple range tests were employed to determine specifically which groups, at which measurement points, were contributing to any significant main or interaction effect (Table 2, Figure 1). As can be seen in the repeated measure computations, there was no reliable differences within or between the test results. Even though updated measures were not available for the NSSMS experienced group (due to limited number of personnel taking all sets over time).

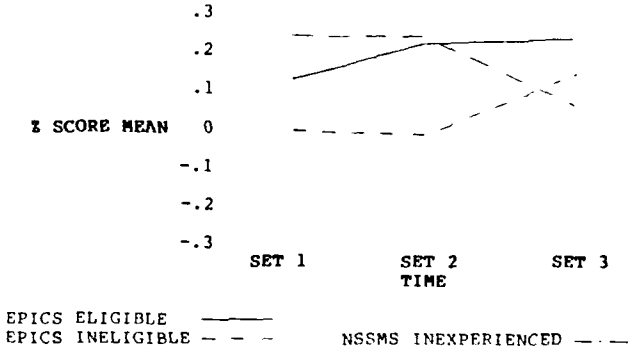
Two other ANOVAs were computed to answer Evaluation Question 3. A 3 X 3 ANOVA was computed with the substitution of the NSSMS experienced technician group for the NSSMS C-school graduate group (see Table 1) data. It should be remembered that this group was measured at three points in time. It actually included measurements of three different subgroups of experienced technicians on the same scenario sets as those administered to the comparison groups at different points. To determine if subgroup differences contributed to confounding interaction effects, a simple F test was computed across them. This ANOVA (including multiple comparison tests) should indicate the degree to which EPICS groups improve over time compared to a "standard" determined by the experienced technician group performance, and indicate if they

ever achieve parity within a year's time after graduation. This has important implications for extrapolating the average EPICS sailor's contribution for a six year enlistment.

Table 2
WEIGHTED Z SCORE DATA
ONLY THOSE WITH COMPLETE DATA (REPEATED MEASURES)

Groups	(Set #1)			(Set #2)			(Set #3)		
EPICS	N	$\bar{X}$	STD	N	$\bar{X}$	STD	N	$\bar{X}$	STD
Eligible	17	0.02	0.80	17	0.12	0.55	17	0.03	0.80
Ineligible	8	-0.09	0.50	8	-0.10	0.58	8	0.03	0.71
NSSMS									
Inexperienced	32	0.14	0.39	32	0.13	0.56	32.	-0.05	0.50

Figure 1
Weighted Z Score



The second analysis for Evaluation Question 3 employed a 2 X 3 ANOVA design. In this instance, NSSMS inexperienced graduates were compared with NSSMS experienced technicians. The same comparisons with the NSSMS Experienced Technician group were conducted as in the previous analysis. Table 1 presents the data utilized for this analysis.

Analysis of variance for composite groups of EPICS vs. NSSMS inexperienced and experienced were completed, in terms of the original three research questions and there was no statistically reliable difference in the results.

As can be seen in Tables 3 and 4 there was a significantly reliable difference in the performance of the EPICS and NSSMS inexperienced personnel at testing on Set 1 and Set 3. Interestingly the EPICS did worse at Set 1 but better at Set 3. In the comparison of the EPICS vs. EXPERIENCED, there was no reliable difference in the performance of the EPICS and NSSMS experienced groups.

Table 3
T TEST TABLE
EPICS vs. NSSMS INEXPERIENCED

ITEM	GROUP	N	X	SD	T	p
SET 1	EPICS	64	-.11	.76		
	NSSMS Inexperienced	72	.10	.41	-1.9642	.0001**
SET 2	EPICS	42	.003	.58		
	NSSMS Inexperienced	54	-.03	.67	.24	.29
SET 3	EPICS	33	.007	.73		
	NSSMS Inexperienced	37	-.07	.51	.49	.04*

* p significant @.05
** p significant @.01

Table 4
T TEST TABLE
EPICS vs. NSSMS EXPERIENCED

ITEM	GROUP	N	X	SD	T	p
SET 1	EPICS	64	-.11	.76		
	NSSMS Experienced	11	.07	.53	-.72	.47
SET 2	EPICS	42	.002	.58		
	NSSMS Experienced	10	.03	.83	-.11	.91
SET 3	EPICS	33	.007	.73		
	NSSMS Experienced	8	.20	.57	-.68	.5

CONCLUSIONS

TPEP Feasibility. The program that was developed for this evaluation effort appears to present a feasible approach. That is, the troubleshooting scenarios developed were considered by experts in the field, for purposes of the limited application, to be representative of the troubleshooting requirements of the work environment. The evaluation criteria, i.e., the factors and weighing of same, was also considered to be credible by the subject matter experts. It was concluded that the approach has a high degree of "face validity" and seems to be a feasible approach in addressing the question of evaluation/discrimination in one of the higher order cognitive skills required in the high technology maintenance occupational community. As was indicated earlier this effort was developed utilizing a logical and rationale approach but was not subjected to as

stringent a development process as would be necessary if the approach were to be considered for wider use and application.

EVALUATION of EPICS vs. NSSMS CPS. Based on the test and evaluation of the TPEP on the EPICS (eligible and ineligible) and NSSMS conventional personnel system (CPS) personnel (inexperienced and experienced) it was determined that the answers to the original three questions were:

Question 1. Were EPICS System Technician Training graduates, FT eligible and ineligible, as proficient in troubleshooting NSSMS faults as CPS NSSMS C-school graduates (inexperienced) at three points in time; a) completion of NSSMS school, b) 4 to 8 months after graduation, and c) 11 to 12 months after graduation?

Answer 1: There was a statistically reliable difference in the performance of the EPICS ineligibles at graduation as compared to the EPICS eligible and CPS NSSMS (both experienced and inexperienced). There was however no reliable differences in performance of the groups at the other points of evaluation.

Question 2. Did EPICS STT graduates and C-school graduates troubleshooting proficiency improve at a similar rate over time?

Answer 2: The EPICS performance (particularly the ineligibles) showed an increasing performance trend whereas the NSSMS C-school performance showed a negative trend albeit none of the changes were reliably different over time.

Question 3. How did EPICS STT graduates and C-school graduates compare in troubleshooting proficiency with NSSMS technicians having greater than 18 months on-the-job experience?

Answer 3: There was a statistically reliable difference at two of the evaluation points, i.e., the NSSMS experienced had higher proficiency scores on the graduation set of scenarios (although not reliably different to the EPICS eligibles when compared separately) and lower proficiency scores at the last comparison point.

RECOMMENDATIONS

In that it appears the approach is feasible and appears to have significant utility both as an evaluation and training assessment process the following recommendations are made:

- o Investigate the procedure in greater detail.
- o Improve the Test and Evaluation Process and Application.
- o Develop, test and evaluate approach for other occupational areas.
- o Apply results in Training Assessment and Improvement.

CURRENT STATUS

The Deputy Chief of Naval Operations for Manpower Personnel and Training has requested the Navy Training Laboratory (Code 52) of the Navy Personnel Research and Development Center, San Diego, California (92152-6800), to pursue the recommendations as presented above with a start date of the new/expanded research effort of November 1986. For further information on the project contact the project manger Dr. Harry B. Conner at the above address or by phone at (619) 225-6721 or on Autovon at 933-6721.

DEVELOPING OUTLINES FOR SPECIALTY KNOWLEDGE TESTS USING OCCUPATIONAL SURVEY DATA

2LT Kathleen M. Longmire
Mr. William J. Phalen
Air Force Human Resources Laboratory

Mr. Johnny J. Weismuller
Texas Maxima Corporation

In 1983, the Air Force Human Resources Laboratory initiated a contract to develop a viable procedure for integrating occupational survey data into the Specialty Knowledge Test (SKT) construction process. This paper reviews the Laboratory's major research findings and reports on recent accomplishments. Topics include the grouping of job inventory tasks into meaningful test content areas and the establishment of "testing importance" weights. Covered will be exploratory work to develop a suitable, cost-effective criterion to predict content area weights using currently available task factors. Recent work has focused on development of an automated procedure to make the survey data more responsive to the needs of SKT development teams. An interactive, micro-based approach has been developed which allows survey data to be downloaded into a microcomputer for easy access and manipulation. Results obtained through the application of this approach will be reported along with recommendations for continued development and refinement of available software.

Complete paper is available from: 2LT Kathleen M. Longmire
AFHRL/MODM
Brooks AFB, TX 78235-5601
Autovon: 240-3551

Estimates of Task Parameters for Test and Training Development

R. Gene Hoffman and Patrick Ford
Human Resources Research Organization

The Army's Project A is a large scale effort to validate the ASVAB and a battery of new selection and classification tests for enlisted soldiers. The effort requires comprehensive job performance measures as validation criteria. In the early stages of the project the domains of nine selected MOS were described to allow the selection of performance variables which could be translated into reliable and representative samples of those performance domains. The problem was to narrow down large domains. The problem is a familiar one in the military context in both the testing and training arenas. That is, job analyses have already been conducted and doctrinal directive written which specify at great length the tasks which soldiers in each MOS are supposed to be able to perform. Far too many tasks are designated as part of the job than any particular training or testing program can cover.

To reduce the task domains for Project A, five task parameters were identified as potentially significant for the selection of sets of representative tasks. These include (1) the relative importance among the tasks, (2) the similarities among the tasks, (3) the performance frequency of each task, (4) the difficulty of each task, and (5) the variability in performance for each task. Details concerning all of these parameters and how they were used in task selection is reported elsewhere (HumRRO & AIR, 1984) and will not be repeated. Our focus is retrospective. Performance measures have been constructed and administered to approximately 400 to 650 soldiers in each of the nine MOS. This provides the opportunity to examine the validity of the task selection data for three of the task parameters: (1) task difficulty, (2) task variability, and (3) task frequency.

Data Base

The "population" for this analysis is tasks rather than people, and the sample is the overlap between the set of tasks for which hands-on performance tests were administered during Project A's concurrent validation phase and the AOSP task list as refined for task selection uses (Campbell, et al., 1985). Some adjustments were necessary because equipment variation necessitated the use of alternative test forms whereas AOSP statements were equipment generic. Thus, 135 tasks spanning the nine MOS were included in the analysis.

Difficulty and variability task parameters were estimated during task selection using a single rating scale. For each AOSP task within their respective MOS, subject matter experts (SME; Ns ranged from 10 to 26 for the nine MOS) were asked to describe the performance distribution of soldiers. They were asked to indicate: "Out of 10 soldiers, how many can do the task: (1) All of the time?, (2) Most of the time?, (3) About half of the time?, (4) less than half of the time?, or (5) Never?" SMEs were also given an escape option of "Not observed." Each set of SME responses therefore represented a

frequency distribution of task performance. By assigning performance values (1 to 5) to the response intervals, a performance mean and standard deviation was computed for each task for each SME. For each task, these individual SME means and standard deviations were averaged across SME, excluding SME who responded with "not observed." Thus, the average SME mean and average SME standard deviation became the difficulty and variability parameters used in the task selection process. Interrater reliabilities within each MOS were in the .70s and .80s for task difficulty and in the .50s and .60s for task variability for the nine MOS. SME (generally E-6 to E-7) rated approximately 150 to 300 tasks within their MOS. Further details are presented in HumRRO and AIR (1984).

Task frequency data used in task selection were taken directly from the AOSP survey results for skill level one soldiers. The specific index was the percent of soldiers reporting that they performed each task.

On the criterion side of this validation, actual test statistics from the concurrent validation data collection provide task difficulty and variability estimates. Performance on these tasks was assessed using four modes: (1) hands-on tests, (2) written tests, (3) peer ratings and (4) supervisor ratings. Means and standard deviations for all four measurement modes were used as criteria against which SME derived estimates were compared. Hands-on and written test scores were percent correct for either steps or items. Performance ratings were given by both peers and supervisors on a 7-point scale ranging from "among the very worst" to "among the very best" at the end points with "about the same as others" at the midpoint.

Project A concurrent validation also included a job history questionnaire completed by each soldier. For each task in the hands-on test sample, the questionnaire asked soldiers to describe on a five point scale how recently they had performed the task and how frequently in the past six months they had performed the task. These responses, averaged across soldiers, provide an independent assessment of task experience for validating AOSP frequency data.

Convergence between task selection data and concurrent validation measurement data was assessed with simple correlations. Correlations within each MOS and across all MOS are reported.

For MOS level correlations for task difficulty and variability estimates, Ns range from 13 to 17 tasks for hands-on and ratings measures, and 12 to 16 for written measures. Not all MOS had the same number of hands-on tests and for six tasks there was no matching written test. One task had no matching rating. Across the nine MOS, the total numbers of tasks were 135 for correlations involving hands-on data, 129 for correlations involving written tests and 134 for correlations involving ratings. Since AOSP frequency data were not available for all tasks, MOS level correlations of task experience were based on Ns which ranged from 10 to 15, with a total of 108 tasks across all MOS.

Results

Table 1 presents correlations between SME estimates and data-based estimates of task difficulty. At the MOS level the correlations fluctuate from -.04 to .95 and given the small Ns on which these correlations are computed such large fluctuations are expected. Confidence interval estimates depend on sample size, size of the observed correlation and are not symmetrical. For simplicity however, it is useful to use one central confidence interval for reviewing a set of correlations. Thus, the 95 percent confidence interval, using the lowest N (12) and an average $\underline{r}$ near .50 is $\underline{r}$ = -.10 to $\underline{r}$ = .84 which is not very different from the range observed in Table 1. Across all MOS, SME ratings of task difficulty are more predictive of rating means as given by peer and supervisors than written and hands-on test score means. The .95 confidence interval for total sample correlations using the lowest N (129 for written tests) and an average $\underline{r}$ = .50 is $\underline{r}$ = .36 to $\underline{r}$ = .62. Thus, the variation among the correlations is not greater than chance.

Table 1

Correlations Across Tasks Between SME Means and Measurement Mode Means For Each MOS and Total Sample

MOS	Hands On	Written	Peer Rating	Sup. Rating
11B	0.50	0.21	0.69	0.80
13B	0.92	0.70	0.81	0.82
19E	0.54	0.47	0.95	0.93
31C	0.58	0.13	0.83	0.86
63B	-0.04	0.07	0.69	0.56
64C	0.34	0.51	0.49	0.65
71L	0.71	0.66	0.36	0.30
91A	0.30	-0.11	0.65	0.74
95B	0.21	0.15	0.29	0.31
TOTAL	0.43	0.33	0.59	0.62

Table 2

Correlations Across Tasks Between SME Standard Deviations and Measurement Mode Standard Deviations For Each MOS and Total Sample

MOS	Hands On	Written	Peer Rating	Sup. Rating
11B	0.62	0.34	0.86	0.68
13B	0.75	0.37	0.77	0.77
19E	0.51	0.52	0.28	0.17
31C	0.60	0.28	0.17	0.54
63B	0.16	0.14	0.07	-0.02
64C	0.26	0.12	0.87	0.82
71L	0.22	0.39	0.70	0.30
91A	0.25	0.06	0.18	0.68
95B	0.50	0.39	0.33	0.48
TOTAL	0.35	0.26	0.42	0.48

Table 2 presents the analogous correlations between SME estimates of task variability and data based estimates. Again at the MOS level the correlations fluctuate from -.02 to .86. Again, however correlations do vary more than expected by chance.

For reference, intercorrelations among task means and among task standard deviations are presented in Tables 4 and 5 in an Appendix.

Table 3 presents correlations between Project A frequency and recency and AOSP task experience estimates, as well as correlations between an unweighted linear composite of the frequency and recency with AOSP frequency.

Looking at the composite, correlations range from .02 to .90 for the within MOS data (.95 confidence interval for an average $\bar{r}$ = .56 is $\bar{r}$ = -.10 to $\bar{r}$ = .88). Across all MOS, frequency and recency means for the 108 tasks each correlate .46 with AOSP frequency (.95 confidence interval is $\bar{r}$ = .31 to $\bar{r}$ = .58). Frequency and recency means correlated .91 with each other, so that using a composite of the two does little to strengthen the relationship between the two sets of experience data.

Table 3

Correlations Across Tasks Between AOSP Frequencies and Job History Responses for each MOS and Total Sample

MOS	Frequency	Recency	Composite
11B	0.85	0.90	0.88
13B	0.55	0.46	0.52
19E	0.53	0.43	0.50
31C	0.14	0.09	0.13
63B	0.00	0.05	0.02
64C	0.65	0.81	0.76
71L	0.11	-0.08	0.02
91A	0.88	0.92	0.90
95B	0.49	0.58	0.53
TOTAL	0.46	0.46	0.47

Discussion

Results indicate that, in the absence of hard performance data, SME estimates can provide reasonably valid, though certainly not perfect, estimates of difficulty and variance. Given validity coefficiencies in the .40 to .60 range, SME estimates of task difficulty can be useful for making gross judgments differentiating particularly hard or easy tasks. In essence, that was the use made of the SME difficulty estimates during task selection with the very hard and the very easy tasks generally not selected for testing. Thus, there is some degree of range restriction in the SME ratings used in the present analysis and the validity of the SME estimates may be understated.

The strength of the relationship between SME task difficulty and performance rating means is interesting in light of the performance rating scale. Theoretically the scale should have led to means near the mid-point for every task, with near zero variance across tasks. Realistically, our knowledge of common rating errors led us to hedge our bets here. Thus, we analyzed the performance rating means expecting to find convergence with SME means. Even though the standard deviation across tasks of the rating means were restricted to .28 and .36 for peers and supervisors, respectively, the variance in task means that did exist was strongly associated with SME task

difficulty estimates. Raters apparently had a hard time making purely normative judgments. That is, raters may have been reluctant to give average or below average ratings on tasks that almost all soldiers perform well.

Validities for the SME estimates of performance variability are lower. Intercorrelations among all estimates of task variability show a similar reduction (compare Tables 4 and 5 in the Appendix). Thus, relative differences among tasks in variance seem more affected by test mode than do their relative differences in difficulty. This makes SME estimates of task performance variability less useful for task selection.

Project A and AOSP estimates of task frequency show modest but perhaps more limited convergence than might be expected from two self-reports of essentially the same phenomenon: participation in various tasks. There are, however, several differences between the two which may have reduced their convergence. First, they provide different experience indices (percent of soldiers who do a task from AOSP data versus average number of times a task is done from Project A data) which may have distorted the relative distributions for tasks done as a daily part of the job (e.g., type a DF for 71L clerks) versus tasks practiced only during set training periods (e.g., load, reduce a stoppage and clear an M16). Second, the surveys were conducted at different times (several years apart for some MOS), and any instability over the intervening time periods would reduce convergence. This was the case for two MOS with low experience convergence (31C and 63B) where preparation of task tests was more cumbersome than other MOS because of the variety and continuing evolution of equipment. Finally, AOSP estimates were based on a sample of the entire first tour, while Project A estimates were based on soldiers representing a more limited range of one to two years time in service. As soldiers increase in time in service, their job duties may expand and change. The distinctions between the two surveys are important caveats for interpreting either set of experience data.

References

Campbell, C. H., Campbell, R. C., Rumsey, M. G., and Edwards, D. C. (1985). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO) and American Institutes for Research (AIR) (1984). Selecting job tasks for criterion tests of MOS proficiency (ARI Working Paper RS-WP-84-25). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Appendix

Table 4

Intercorrelations Among Measurement Mode Means

Mode	Hands-On	Written	Peer Rating	Supervisor Rating
Hands-On	1.00			
Written	0.52			
Peer Rating	0.58	0.40	1.00	
Supervisor Rating	0.53	0.37	0.93	1.00

Table 5

Intercorrelations Among Measurement Mode Standard Deviations

Mode	Hands-On	Written	Peer Rating	Supervisor Rating
Hands-On	1.00			
Written	0.40			
Peer Rating	0.48	0.17	1.00	
Supervisor Rating	0.42	0.20	0.70	1.00

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

ROYAL AIR FORCE NAVIGATOR SELECTION: RESOLVING AN OLD PROBLEM.

Eugene F. Burke
United Kingdom Exchange Psychologist
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

INTRODUCTION. This paper provides a brief summary of recent research aimed at improving selection and allocation to the role of Royal Air Force navigator. This ongoing research program has two broad objectives:

(1) **Development of New Selection Tests.** Based on job analyses conducted at training and operational squadrons, computerised measures of spatial and dual-task abilities have been constructed for inclusion in the RAF's automated testing system (Burke, 1983; D'Arcy, 1986). Reliability, validity and normative data are currently being collected for these new measures.

(2) **Validation of Current Selection Tests.** An estimate is obviously required of the improvement in prediction offered by these new tests. Current selection tests have therefore to be validated so as to establish a baseline against which the validity of new measures can be compared. The validation of existing RAF navigator selection tests is the subject of this paper.

PROBLEMS CONFRONTING THE VALIDATION STUDY. Meeting the latter of the above objectives was found to be problematic due changes over time in selection conditions, recruitment and training policies. With regard to selection conditions, an earlier study had shown that ab-initio selection (direct selection at the point of applicant assessment) to the role of navigator was dependent upon the selection and hiring ratios for pilot (i.e. the proportion of applicants above the minimum cut-off for pilot aptitude and the pilot recruitment target). Selection to navigator was found most likely to occur either when an applicant had achieved scores above the minimum for navigator but below that for pilot, or when scores were above both minima but there were a large number of qualified applicants in relation to the pilot recruitment target. However, these selection conditions were identified during a period in which the ratio of applicants to recruiting targets was high, a situation which has not continued to be as favourable during the 1980's. Indeed, review of the recruiting targets and achievements for the financial years 1979 to 1982 showed consistently larger shortfalls in ab-initio navigator recruitment in comparison to such recruitment to pilot.

These shortfalls in ab-initio recruitment had led to an increase in navigator recruitment from within existing RAF personnel, principally from those failing pilot training at either Flying Selection School (FSS: a short flying screening course reintroduced by the RAF in 1979) or pilot basic flying training (BFT: the first formal stage of pilot training). Interviews conducted with instructor staffs indicated a general attitude that ex-FSS candidates performed badly in navigator basic flying training, an opinion that was subsequently born out by analysis

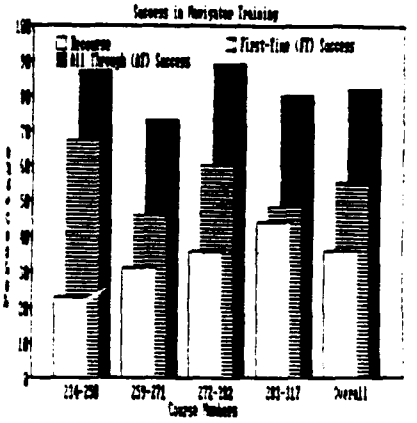
TRAINEE TYPE	PASS WITHOUT RECOURSE	PASS AFTER RECOURSE	FAIL TRAINING	N
EX-PILOT BFT	62%	30%	9%	108
AB-INITIO NAVIGATOR	57%	31%	12%	176
EX-FSS PILOT	49%	29%	22%	110

CHI-SQUARE SIGNIFICANT AT THE 0.05 LEVEL

of success rates.

RAF navigator training comprises two stages: basic flying training in which trainees are introduced to the principles of air navigation and during which simulator and flying sorties are used to assess potential for later fast-jet or multi-engine training; advanced flying training in which trainees are taught the skills relevant to specialisation within fast-jet and multi-engine roles. In accord with previous validation studies, success in navigator basic flying training was taken as the criterion for validation of existing selection tests.

Course results were obtained for 537 navigator trainees for the period October 1978 to May 1985 (the point at which the validation study was undertaken) providing data for three success rates: all-through success given by a trainee achieving a pass irrespective of whether that trainee underwent recoursing and additional training due to poor performance on his original training course; first-time success given by a trainee achieving a pass without being recoursed and undergoing additional training; and recourse. Overall, this data showed a high level of all-through training success (82%), but this was undermined by a much lower rate of first-time success (55%). Both of these rates were further found to follow distinct trends corresponding to changes in recruitment and training policy. All-through success rates were found to decline on initial recruitment of ex-FSS candidates; to increase with closer screening by instructor staffs of these candidates prior to entering training; and a further decline with the introduction of revised basic flying training assessments aimed at reducing failure at the advanced stage of training. Similar trends were also found for first-time success. The most notable trend was that for recourse rates which showed a monotonic increase over time. As such, all-through success rates became increasingly dependent on higher levels of recoursing with all-through success rates showing an eventual decline for later courses.



A final validation sample of 401 navigator trainees was obtained who had complete selection test and training success data (this sample included an additional seven trainees who did not fall within the ab-initio and ex-pilot groups presented in the table on the previous page). As the study was limited to a dichotomised criterion, first-time success (pass without recourse) was chosen as reflecting the training load upon instructor staffs and facilities.

A NEW MEASURE OF NAVIGATOR TRAINING POTENTIAL. Aptitude measurement for the roles of pilot and navigator currently comprises the following six tests: MATAB1 (Algebraic and Arithmetic Reasoning), MAT62 (Non-verbal Reasoning), MATF (Speed and Accuracy in the Use of Numerical Tables), INSB2 (Speed and Accuracy in the Interpretation of Aircraft Instruments), CVT (Single-axis Pursuit Tracking) and DMA (Dual-axis Compensatory Tracking). Three validations of these tests had been conducted against navigator basic training success since 1973. Each of these studies had derived a different composite with the only constant elements in this changing aptitude profile being the tests MATAB1 and MATF. An initial data run in the current study using the usual approach of allowing the tests to vary individually in the regression yielded a further best weighted composite excluding MATAB1. Although the content of navigator training has been modified following the introduction of new aircraft systems, there are obvious problems in interpreting the results of these validations as reflecting

meaningful changes in the aptitude requirement for navigator.

The problems inherent in changes in the composition of the training population and the process of trainee assessment were found to be further compounded by the intercorrelations between current selection tests, suggesting at least a moderate degree of multicollinearity. One solution to this problem is given by entering factor rather than individual test scores into the regression analysis (Kerlinger and Pedhazur, 1973; Dobie, McFarland and Lang, 1986). Test intercorrelations were obtained from an applicant sample of 12,306 tested between 1977 and 1982 (the period during which the validation sample were selected for RAF service). These correlations were entered into program 4M of the Biomedical Data Package (Dixon, 1981) to yield a two factor varimax solution. The first of these factors is defined by tests MATAB1 and MAT62, and was named for convenience as Aircrew Reasoning (AR) given the content of these tests and those of MATF and INSB2 which also loaded significantly on this factor. The second factor is defined by the psychomotor tests CVT and SMA and was named as such (PM).

It should be noted that factor analysis was applied in the present instance to seperate out distinct components of test variance. The stability of this two factor solution has yet to be further tested by increasing the set of measures included in the factoring, which may suggest that alternative models provide a better fit than the one employed here. Although five of the current aptitude tests fall quite neatly into the two clusters obtained, INSB2 yielded non-trivial loadings on both factors. Comparison of the validities of INSB2 for the current navigator sample (0.105 uncorrected for restriction of range) and a comparable cohort of pilot trainees (N = 787, r = 0.245 uncorrected for restriction of range) shows a much lower contribution to the prediction of navigator success. Regression analyses found no substantial gain by including INSB2 in the computation of AR and it was therefore calculated from standardised scores on tests MATAB1, MAT62 and MATF. Given the small differences in the factor weights and the equivalence in reliabilities of these tests, AR was computed using unit weighting.

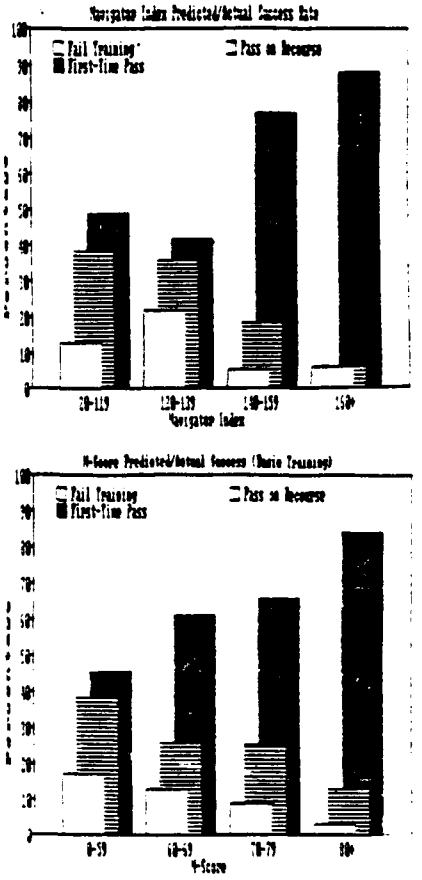
The application of factor analysis is one of two differences from previous navigator validations in the methods used to determine a navigator aptitude composite. In preference to the least squares models applied in the past, a non-linear logistic model was employed to regress the dichotomised criterion of first-time pass/fail on the AR factor (Aldrich and Nelson, 1984). Logistic regression had been used in revising the pilot aptitude composite and had been found to achieve a better fit to the dichotomised criterion of pass/fail in pilot basic flying training. RAF selection staffs were found to prefer the revised pilot aptitude composite which gives a direct indication of training success in terms of probabilities ranging from 0 through 100% (Walker-Smith, 1984). Previous pilot and navigator aptitude composites had used linear weighting of stanine scores to obtain scales ranging from 20 to 180. This revised pilot aptitude scale is referred to as P-Score.

Logistic regressions were carried out using the BMDP program PLR. In addition to the AR and PM factors, biographical variables of age, level of education and previous flying experience at time of selection were also included. Only AR was found to offer significant prediction of first-time success in navigator basic flying training and to be consistent across periods of high and low wastage. This new logit based model was called N-Score to parallel the P-Score scale.

CONSISTENCY IN SELECTION AND ALLOCATION. The process of test validation usually follows a narrower focus upon predictive accuracy. In the present context, validation was directed towards the resolution of problems in the selection and

recruiting systems within which test information is embedded. The content of the tests themselves has not changed for some time, and the purpose of the analyses described was to utilise the constancy represented by existing tests to derive an aid for managing and optimising allocation to the roles of pilot and navigator.

Brown (1981) has noted that serious distortions of test validity may occur when tests are used without systematic controls to safeguard the quality of test administration, and when decisions to select are made despite scores below cut-offs. In a recent review of computer based test interpretation, Reid (1985) cites research on the "base-rate fallacy" which occurs when base-rate information for a measure is ignored in preference for more specific information concerning the particular individual under assessment. In providing reminders of actuarial data, Reid (1985) emphasises the need for diagnostic aids which reduce assessor variability in the use of base-rate data and control the temptation to ignore such data in fixating upon case specific information. Since the apparent validity of a test may be conditional upon a number of implicit factors and that military testing is frequently undertaken by lay personnel having little appreciation of the subtleties of psychometrics, the minimisation of classification errors is of obvious importance.



The probabilities of pilot and navigator training success given respectively by the P- and N-Scores have been integrated in a simple matrix so as to reduce the complexity of the allocation decision to be made by selection staffs. The use of this matrix is further simplified by reducing it to four areas defined by cut-offs for each role: Accept Pilot or Navigator (above cut-offs for both P- and N-Scores); Accept Pilot Only (above cut-off for P-Score only); Select Navigator Only (above cut-off N-Score only); and Reject Pilot and Navigator (below cut-offs for both P- and N-Scores). Using the aptitude scores as coordinates, an individual applicant's position on the matrix can be plotted to give an immediate impression of the appropriate selection/allocation decision. Thus, the parameters of this simple interpretative aid are clearly defined and allow for consistent presentation of actuarial data across selectors; provide a standardised basis upon which selectors' decisions are to be justified; a policy tool from which recruiters can identify priorities (over/under selection to each role)

P-SCORE

ABOVE N-SCORE CUT-OFF ONLY SELECT FOR NAVIGATOR	ABOVE CUT-OFFS P- AND N-SCORE SELECT FOR PILOT OR NAVIGATOR
BELOW CUT-OFFS P- AND N-SCORE REJECT FOR PILOT AND NAVIGATOR	ABOVE P-SCORE CUT-OFF ONLY SELECT FOR PILOT

N SCORE — C2 — C1

C1 = P-SCORE (PILOT APTITUDE) CUT-OFF

C2 = N-SCORE (NAVIGATOR APTITUDE) CUT-OFF

and problems (over recruiting from bad cells in the matrix); and a policy tool from which trainers can anticipate the need to adjust future training patterns. Whilst intended to minimise the additional costs incurred by later recruitment from those failing pilot training, the matrix allows for such recruitment by identifying those with a high likelihood of first-time success in both pilot and navigator basic flying training. Accordingly, a clear function is also apparent for selection test information in decisions concerning the reassignment of those failing pilot training.

REFERENCES.

Aldrich, J. H., and Nelson, F. D. (1984). Linear probability, logit and probit models. Beverly Hills: Sage Publications.

Brown, S. H. (1980). Validity distortions associated with a test in use. Journal of Applied Psychology, 64, 460-462.

Burke, E. F. (1983). Ability requirements for RAF navigator. Chief Scientist (RAF) Report No. 18/83, UK Ministry of Defence.

Burke, E. F., and Schmit, V. (1986). N-Score: An improved method for the selection of RAF navigators. Directorate of Science (Air), UK Ministry of Defence (In preparation).

D'Arcy, S. H. (1986). Aptitude testing in the Royal Air Force 1939-1986. Air Clues, August.

Dobie, T., McFarland, K., and Long, N. (1986). Raw score and factor score multiple regression: An evaluative comparison. Educational and Psychological Measurement, 46, 337-347.

Dixon, W. J. (1981). BMDP - Biomedical computer porgrams. Berkely: University of California Press.

Kerlinger, F. N., and Pedhazur, E. J. (1983). Multiple regression in behavioral research. New York: Holt, Rhinehart and Winston.

Reid, G. H. (1985). Computer-based test interpretation: The potential of quantitative methods of test interpretation. Computers in Human Behavior, 1, 207-219.

Walker-Smith, G. (1984). P-Score: An improved method of pilot selection. Directorate of Science (Air), UK Ministry of Defence (Unpublished manuscript).

Conditional Logit Analysis for Personnel Selection

Donald H. McLaughlin
American Institutes for Research
Palo Alto, California

Estimating the contribution of quantitative factors to a dichotomous outcome, such as "success" or "failure," is a frequently encountered data analysis problem. For example, one may want to determine whether a particular factor contributes to the likelihood of completion of a training course, the likelihood of being promoted, the likelihood of re-enlistment, the likelihood of answering a particular question on a test correctly, or in biomedical applications, the likelihood of dying.

Although linear regression is often used to address this kind of problem, the linear model is not appropriate for outcomes that can take on only two values. A more valid approach is to model the probability distribution of the dichotomous outcome, expressing that probability as a function of quantitative factors. For this approach, it is necessary to postulate a family of functions that map the set of all real numbers onto the set of numbers that can be valid probabilities (i.e., numbers from 0 to 1). Requiring monotonicity and symmetry have led statisticians to focus on two such families: normal ogives (probit analysis) and logistic curves (logit analysis), which are known to be practically interchangeable for fitting real data. The logit function is:

$$(1)\quad \text{Prob}(\ Y_i=1\ ) = p_i = e^{B'Q_i}/(1+e^{B'Q_i}) = w_i/(1+w_i),$$

where Y=1 indicates a "success", B'Q is a linear combination of component predictors factors, and w is short for $e^{B'Q}$.

A data set of Y_i's and Q_i's can be analyzed to estimate the vector of weights, B', and to test whether the components of B' are significantly different from zero. The accepted approach is to estimate the likelihood of the data set, given B' and Q, and to choose B' that maximizes that likelihood. The likelihood of a particular set of independent successes and failures is, of course, the product of the individual event probabilities.

Frequently, the data set will consist of a series of "pools," each representing an independent replication of the effects; and frequently it is reasonable to suppose that Q determines not how many in a pool are successful but rather which ones in each pool are successful. In such cases we condition the logit analysis on prior information about the total number of successes in a pool, s, as well as the total size of the pool, n. When such a condition is appropriate, it improves the analysis by removing one parameter for each pool from the number of parameters to be estimated. The conditional logit function for one success in a pool of n elements is:

$$(2)\quad \text{Prob}(\ Y_i=1\ ) = e^{B'Q_i} \Big/ \sum_{j=1}^{n} e^{B'Q_j} = w_i \Big/ \sum_{j=1}^{n} w_j.$$

If there are several successes per pool, there are two approaches for conditional logit analysis: a method combining (1) with the formal definition of conditional probability, and a method combining (2) with the concept of sequential sampling of single items without replacement.

According to the first approach, which has been investigated by Cox (1972), the conditional likelihood of the data is expressed as a ratio:

$$(3)\quad \text{Prob}(\text{ data } \mid s\ ) = \text{Prob}\ (\text{ data }) \;/\; \sum_{r \in C(s,n)} \text{Prob}\ (\text{ data}_r\ )$$

$$= \frac{\prod_{k=1}^{n} (p_k)^{Y_k}\ (1-p_k)^{(1-Y_k)}}{\sum_{r \in C(s,n)} \prod_{k=1}^{n} (p_{r(k)})^{Y_{r(k)}}\ (1-p_{r(k)})^{(1-Y_{r(k)})}},$$

where r(k) is an abbreviation for the index of the k-th element in combination r, $Y_{r(k)}$ indicates which s items are successes in combination r, and C(s,n) is the set of all the different combinations of s successes in a pool of n items. For the logit function defined by (1), this can be simplified to :

$$(4)\quad \text{Prob}(\text{ data } \mid s\ ) = \frac{\prod_{k=1}^{n} (w_k)^{Y_k}}{\sum_{r \in C(s,n)} \prod_{k=1}^{n} (w_{r(k)})^{Y_{r(k)}}}\ .$$

To simplify the presentation, we introduce abbreviated notation for the denominator in (4):

$$\text{Prob}(\text{ data } \mid s\ ) = \prod_{k=1}^{n} (w_k)^{Y_k}\ \;/\; R(s,n).$$

The alternative conditional logit model, investigated by Kalbfleisch & Prentice (1980), considers the successes to have arisen from a sequence of individual successes, each sampled without replacement, from the pool remaining after the previous successes. In this case, the conditional likelihood of the data is expressed as a sum:

$$(5)\quad \text{Prob}(\text{ data } \mid s\ ) = \sum_{r \in P(s)} \prod_{k=1}^{s} w_{r(k)} \;/\; \left(\sum_{j=1}^{n} w_j - \sum_{j=1}^{k-1} w_{r(j)}\right),$$

where P(s) is the set of all permutations of the selected subset. Each term in the sum is a product of s terms like the righthand side of (2), the denominators of which include the terms remaining after the preceding successes in the particular permutation.

Either of these two formulations requires very significant amounts of computation when s is larger than a handful and n is larger than a dozen. The computations for the Cox model are available in the SAS Supplemental Library as PROC MCSTRAT, but, even though the authors of PROC MCSTRAT (Smith et al. 1981) have optimized the computation of the denominator in (4), the cost is prohibitive for all but the smallest pools. For example, the number of terms in the denominator, for s=10 and n=50, is more than 10 billion. Kalbfleisch & Prentice (1980) and others have pointed out the difficulty of these computations as a reason for neglecting the Cox model for practical

purposes. The computations for the Kalbfleisch & Prentice model are similarly difficult and are available in standard packages, such as BMDP2L, only through approximations.

Two approximations to the conditional logit analysis are recommended in the literature. In one, proposed by Breslow, equation (5) is simplified by supposing that the sampling is really with replacement and just happens not to include any item twice. In this case, the probability of each permutation is the same, so the denominator in (5) is just the denominator of (2) to the power s. A second approximation to (5), proposed by Efron, retains the notion of sampling without replacement but still equates the terms in the denominator so that only one "permutation" need be considered. Efron proposed to approximate the subtracted terms in the denominator of (5) by the average of the terms for all the actual successes.

An exact computation for conditional logit analysis

In the course of analyzing factors affecting promotions among pools of applicants for a promotion, I recently examined the Cox model computations and found a shortcut that dramatically reduces the computational effort required. There is an exact repesentation of the sum in the denominator of (4) that involves a few hundred terms, rather than billions. I have programmed this computation into a FORTRAN module that makes conditional logit analysis feasible without need for approximations; and using this exact form, I have examined the accuracy of the approximations using Monte Carlo data.

The sum in question, R(s,n), is the sum of products of s factors, w_k, for all combinations of s factors selected from n. This sum can also be written as a weighted sum of terms R(0,n), R(1,n), ..., R(s-1,n):

$$(6) \qquad R(s,n) = \sum_{j=1}^{s} a_j \, R(s-j,n) \; / \; s \; , \qquad \text{with } R(0,n)=1.$$

In order to compute R(s,n), one first computes R(1,n) using R(0,n), then R(2,n) using R(0,n) and R(1,n), etc. until R(s,n) is computed. (Note: for the conditional logit analysis, involving iterative convergence to the maximum likelihood estimates of B' using the Newton-Raphson procedure, all the intermediate terms, R(i,n), are also required for each iteration.) The coefficient of R(i-j,n) in the equation for R(i,n) is just:

$$a_j = (-1)^{(j+1)} \sum_{k=1}^{n} (w_k)^j \; .$$

Note that these coefficients do not depend on i, so there are only s coefficients to compute, each consisting of n terms. For s=10 and n=50, the number of terms involved in the computation of R(s,n) is only on the order of s(n+s) = 600, compared to 10,000,000,000.

The proof that this works goes as follows. First, we note that equation (4) defines a probability space, with probabilities summing to unity over all possible combinations of s successes in a pool of size n. We create the random variable, S, which is the number of successes in each sample. S is a constant (s), of course, but we can compute its expected value according to the definition of expected value, yielding the equation:

$$(7)\quad s = (1/R(s,n)) \sum_{r \in C(s,n)} \left(\sum_{k=1}^{n} Y_{r(k)}\right) \prod_{k=1}^{n} (w_{r(k)})^{Y_{r(k)}} ,$$

The inner sum is equal to s, of course, for every combination r. The trick is to rearrange the terms in (7), yielding:

$$(8)\quad s = (1/R(s,n)) \sum_{k=1}^{n} \left(\sum_{r \in C(k,s,n)} \prod_{j=1}^{n} (w_{r(j)})^{Y_{r(j)}} \right) ,$$

where C(k,s,n) is the subset of combinations of s out of n that include item k as a success. For a particular k, the number of terms in the inner sum is the number of ways of selecting the remaining s-1 successes from the remaining pool of n-1; and w_k is a common factor of all of those terms.

After factoring out w_k, the inner sum is almost R(s-1,n), missing only those combinations of s-1 out of n that include item k as a success. We can rewrite the inner sum, adding and subtracting the terms that would have been included had we been computing R(s-1,n), yielding:

$$(9)\quad s = (1/R(s,n)) \sum_{k=1}^{n} (w_k)\left(R(s-1,n) - \sum_{r \in C(k,s-1,n)} \prod_{j=1}^{n} (w_{r(j)})^{Y_{r(j)}} \right),$$

where the product at the right now has only s-1 factors, $w_{r(j)}$.

The operation that transformed (8) into (9) can be repeated s times, yielding an equation for s as a weighted sum of terms R(s-j,n). The sum for s can then be simply switched to a sum for R(s,n), with coefficients a_j.

Comparison of the accuracy of approximations

The unconditional logit model and the approximations for the conditional logit can be compared against the exact model, using the now feasible computation of R(s,n). One hundred replications of a pool in which s=15 and n=60, with two predictive factors, Q, a normally distributed variable, and G, a dichotomous factor, were generated. Some results of a Monte Carlo study using these data are shown in Figure 1. The statistic graphed in Figure 1 is $-2\log(\lambda)$ for the contribution of G, where the true values of B' were -.50 for b_G and 1.00 for b_Q and the correlation between Q and G was -.30. The relations among the estimated beta weights reflect the same pattern as shown in Figure 1.

First, the results from the exact unconditional logit model and the exact conditional logit model are very similar. Evidently, when as few as 25% of a pool of 60 are successful, the dependence introduced by the prior constraint on the number successful has little effect. Second, the approximation proposed by Efron appears to be reasonably good. Finally, however, the approximation proposed by Breslow underestimates the effects. Thus, use of this approximation should be avoided.

Although it appears from Figure 1 that Efron's approximation and the unconditional logit model both provide good approximations to the exact conditional logit model for pools with 15 successes out of 60, examples with small sample sizes can be generated in which the approximations are not good. Therefore, the exact conditional logit model should be used when it is the model that fits the process being analyzed.

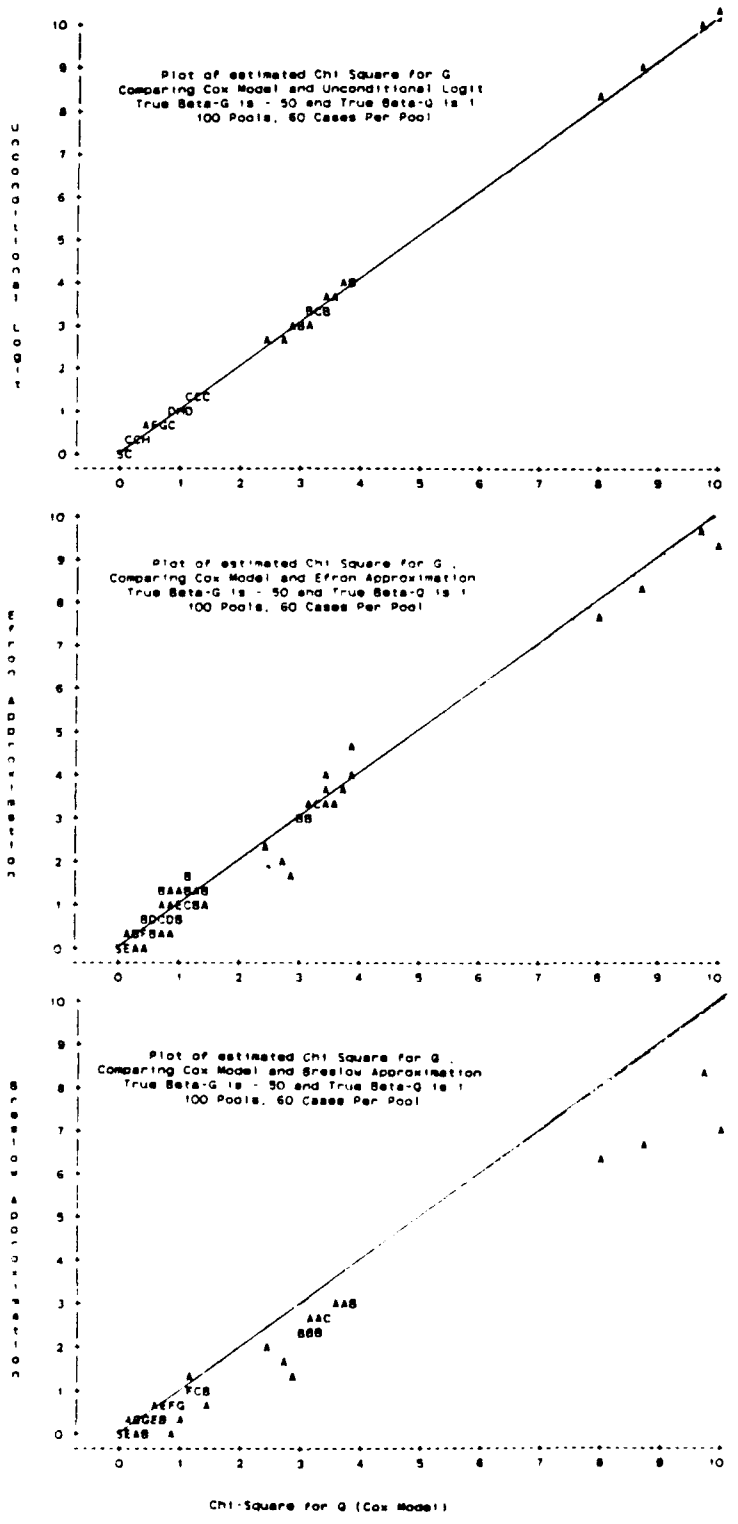


Figure 1. Comparison of chi square statistics for three approximations to the exact conditional logit model.

Infinite beta weights

Finally, comparison of the exact conditional logit model with the approximations has highlighted an important difference. The maximum likelihood estimates of B' for the conditional logit model do not have finite means, because with some probability greater than zero, they diverge to infinity. This occurs whenever the data are "consistent" in the sense that the predictive factors account perfectly for the successes and failures, except possibly for a set of ties on the predictive factors, as in the following example.

Case:	#1	#2	#3	#4	#5	#6
Success:	1	1	1	0	0	0
Predictor:	2	1	1	1	1	0

In this example, the predictor accounts perfectly for Cases #1 (the highest score is a success) and #6 (the lowest score is a failure), and the other cases are all "tied" on the predictor. The exact conditional logit analysis fits this result by setting the beta weight for the predictor to infinity.

The approximations, by contrast, yield finite estimates of B' in these cases. This difference limits the value of the approximations, but it also creates a problem for the exact model computations: how to identify data sets for which some combination of components of B' can be infinite. Surprisingly, one cannot take divergence of the Newton-Raphson algorithm as a criterion for the existence of an infinite solution because the divergence is so slow that it is frequently "caught" by the stopping rule of a sufficiently small change in the log likelihood function. It seems that one must screen for infinite solutions by identifying "consistent" data sets as a part of the analysis. While for a single predictor (as in the example above) this is straightforward, it is not simple to identify whether there exists a linear combination of two or more factors that would render the data consistent. A good place to search, however, is at the B' the Newton-Raphson procedure identifies as a potential maximum. The algorithm I have programmed checks each B' tried in the Newton-Raphson iteration process as a candidate for rendering the data consistent.

Summary

It is now feasible to perform conditional logit analyses for pools with multiple successes; and this method should be considered for analyses aiming to estimate the effects of quantitative factors on dichotomous outcomes.

References

Cox, D. R. (1972). Regression models and life tables (with discussion). J. R. Stat. Soc. B, 34, 187-202.

Efron, B. (1977). Efficiency of Cox's likelihood function for censored data. J. Am. Stat. Assoc., 72, 557-565.

Kalbfleisch, J. D., and Prentice, R. L. (1980). The statistical analysis of failure time data. Wiley, New York.

Smith, P. G., Pike, M. C., Hill, A. P., Breslow, N. E., and Day, N. E. (1981). Multivariate conditional logistic analysis of stratum matched case-control studies. Applied Statistics, 30, Algorithm AS 162.

REV. 11/7/86

A METHODOLOGY TO IDENTIFY THE RELATIVE IMPORTANCE OF TRAINING REQUIREMENTS

F.J. Hawrysh and Major J.P. McMenemy
Directorate of Military Occupational Structures
National Defence Headquarters
Ottawa, Canada

In 1984/85 the Directorate of Military Occupational Structures at National Defence Headquarters, conducted an occupational analysis of the Canadian Forces Maritime Surface and Sub-surface (MARS) Officers Military Occupation (MOC). Two of the aims were to:

a. provide the occupational data required to develop or validate selection standards; and

b. provide the data needed for the development of Course Training Standards (CTS) and Course Training Plans (CTP).

The MARS MOC is multi-dimensional, being responsible for virtually all Naval operations at sea as well as command and staff jobs ashore. The young Naval Officer is expected to perform in such diverse duty areas as seamanship, navigation, administration and leadership. Occupational training is long. The development of the new MARS officer, prior to the completion of the OA was conducted in five phases as follows:

a. Phase 1 - 11 weeks of general military and leadership training, common to virtually all Canadian Forces commissioned officers;

b. Phase 2 - 9 weeks - basic naval knowledge, introduction to ships' small boat coxswain course, damage control, basic ship handling and general seamanship duties and evolutions;

c. Phase 3 - 8 weeks - basic theories and practices of safe navigation;

d. Phase 4C - 4 weeks - duties and responsibilities of the Officer-of-the-Day and damage control theories and practices;

e. Phase 4 - 9 weeks - advanced navigational theories and practices, an introduction to radio telephone procedures, basic fleet manoeuvring and relative velocity; and

f. Phase 5 - 23 weeks - advanced navigation techniques, radar and blind pilotage, astronomic navigation, ship handling and seamanship evolutions.

This five phase development required 64 weeks of training before the new officer was available for his first operational job. In addition to the long training period time, this training was very costly in both capital and operating expense. The program demands not only the time (and salary) of students and instructors but also the dedication of equipment (eg, ships) and other resources (eg, fuel, support personnel) to provide an adequate training environment. In this context, both training design and personnel selection become very important issues.

The Canadian Forces training system aims to provide training which is efficiently designed and performance oriented. Each trainee should be taught only those things which are needed to do the job. In initial MARS training, the junior officer is prepared for his first tour at sea, normally on a destroyer. Subsequent experience and training prepare him for more advanced employment. Preparing the MARS officer to do his job necessitates a careful and comprehensive analysis of training requirements to ensure that the development of Course Training Standards and Plans are maximally efficient. The pursuit of efficiency however, must not detract from his achieving operational effectiveness.

By the same token, inefficient or ineffective selection can increase overall costs and detract from operational effectiveness. Since training is so costly, it is vital that only personnel with high probability of success be selected. Otherwise training becomes a very expensive selection device.

A common method of identifying ability constructs or domains to be considered in selection and in designing training programs is job analysis, followed by "expert judgement". (Guion, 1976, Wexley & Yukl, 1977). In the case of MARS officers, the judgement would have to be extremely astute to validly discriminate among 900 tasks and over 1,000 knowledge items which were identified in the OA.

It was clearly inappropriate to proceed directly from the tremendous volume of job performance and knowledge requirement data to CTSs or selection standards. What was needed was a rational condensation of this data into a manageable criterion. The criterion, historically has been among the most difficult and least successfully addressed issues in industrial psychology. The requirements are that the criterion be relevant, valid, reliable and practical (Smith 1976).

Identification of the critical elements for training in such a broad occupation is difficult. Attempts at isolating and measuring personal characteristics or abilities considered essential for successful performance in the MARS occupation proved frustrating, largely because the criterion has not been defined adequately. The

criterion, in the statistical sense is success in training, but the underlying structure or constituent characteristics which contribute to success have been elusive. Since a global assessment was not likely to be fruitful, a method which started from basic elements or tasks was developed.

Method

When the OA large and complete task list was assembled, a panel of officers, including occupational analysts, experienced MARS officers, and personnel research officers who were working on MARS officer selection procedures was convened. This panel reviewed the task list to eliminate those tasks which were not performed by first term MARS officers (eg, prepare staff papers at HQ). The panel selected a list of 293 tasks likely to be performed by junior officers. A separate questionnaire containing these tasks was developed.

The selection of raters to perform these ratings was considered a critical element. The raters had to be close enough to the operation to be familiar with the tasks but broad enough in their responsibilities and span of control to avoid narrow, departmentally oriented responses. After consideration of supervisors, department heads, and instructors, it was concluded that Commanding Officers and Executive Officers were the appropriate raters.

The inventory was administered to Commanding Officers and Executive Officers of Destroyers (N=47). They were instructed to rate, on a seven point scale, how important it was that first term MARS officers be able to perform each of these 293 tasks without further training upon reporting to their first operational ship. Responses were recorded on machine readable response sheets and entered into a computer file for analysis.

Results

The data were analyzed using the REXALL program from the Comprehensive Occupational Data Analysis Programs (CODAP) developed by the US AFHRL. This program calculates individual rater correlation with the mean rating of all raters, calculates inter-rater reliability (Spearman-Brown prophecy formula), and prints the distribution of means and standard deviations of those ratings. It was found that the raters were in substantial agreement on the relative importance of tasks (Rxx=.977). The most encouraging aspect of the data is that the raters achieved this very high level of agreement while using the whole scale. The range of task importance means runs from 1.10 to 6.60.

The sorted data present a very reliable estimate of the importance of the tasks which make up the job of the junior MARS officer. Samples of the most important and least important tasks are displayed at tables 1 and 2 respectively.

TASK	MEAN RATING	SD
Execute Rescue Station Procedures	6.64	.67
Execute Man Overboard Procedures	6.49	1.09
Execute Steering Gear Breakdown Procedures	6.57	.68
Execute Cryro Failure Procedures	6.28	.96
Manoeuvre Ship Singly and in Company	6.13	1.02
Direct Bridge Activities	6.38	.96
Direct Ship's Routine Alongside	6.11	1.06

Table 1
Sample of Highly Important Tasks

TASK	MEAN RATING	SD
Prepare Ship's Historical Report	1.64	.84
Prepare Ship's Activity Report	1.76	.94
Prepare/Review Annual Physical Fitness Report	1.92	.90
Advise Operations Room Officer on Communications	1.83	1.03
Select Equipment Settings for Ship Degaussing	1.66	.91

Table 2
Sample of Low Importance Tasks

Applications

Following the acceptance of the MARS OA report, work proceeded to the preparation of Occupational Specifications, Course Training Standards, and Course Training Plans. The Specification is a summary document which presents the minimum acceptable standards of task involvement, and knowledge and skill requirements for an occupation at a given level. Course Training Standards are based upon the Specification and stipulate the performance objectives which must be met in training. Course training plans are designed by Canadian Forces School Standards Personnel. The CTP essentially prescribes the

sequence of training, the method media and aids applied in the conduct of instructional courses. The availability of the training importance data greatly streamlined the process. It is relatively easy to differentiate between very important and trivial tasks but the boards found our importance data to be extremely valuable when considering the large number of tasks which fall in the range from less than average importance to more than average (point 3 to 5 on the scale employed). Availability of the data resulted in more accurate specifications, in reduction in board time and expense and significant saving in terms of efficiency of training.

The Canadian Forces Personnel Applied Research Unit (CFPARU) was working on a parallel study of MARS Officer selection standards. CFPARU was able to use the task importance ratings to apply a derivation of the ability analysis procedure developed by Levine, Mallamad, and Fleishman (1978). This trial was very successful, providing estimates of the importance of 42 abilities (Fleishman and Quaintance, 1986).

With the important abilities identified, CFPARU can now proceed to identify or develop effective measures likely to be related to success. This method was particularly useful in this case because of the complexity of the MARS MOC and the large number of tasks performed.

Conclusion

After assessing the utility of this method in the two applications of training design and selection, the Director of Military Occupational Structures has concluded that the training importance method yielded entirely satisfactory reliability and excellent task discrimination. The unconditional and enthusiastic acceptance of the information by Naval operational and training authorities has been taken as evidence of the validity of the results. Substantial changes to training provided to junior MARS officers have been made as a result of the project with considerable savings achieved. Of equal importance, the Navy has been reassured that Naval officers, even at a junior level, perform highly complex jobs requiring lengthy training. Training importance methodology is not required in every OA but will continue to be used by occupational analysts when the complexity of occupation the under study warrants its application.

References

Fleishman, E.A., and Quaintance, M.K. (1984). Taxonomies of Human Performance. Toronto, Ontario: Academic Press.

Guoin, R.M. (1976). Recruiting, Selection and Job Placement. In M.D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally.

Levine, N.G., Mallamad, M.S. and Fleishman E.A. (1978). Decision Aids in Estimating Personnel Requirements. Washington DC: Advanced Research Resources Organization.

Rodgers, M.N. (1984). An Overview of the Naval Officer Production Systems. Technical Note 12/84. Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Wexley, K.N. and Yukl, G.A. (1977). Organizational Behaviour and Personnel Psychology. Georgetown, Ontario: Irwin-Dorsey Limited.

RETENTION OF COMMON MILITARY SKILLS

by

Major Grahame Brown MA RAEC
Systems Consultant (Training Development)
Army School of Training Support (UK)

Introduction

Background. In March of this year, the Director of (British) Army Training (DAT) expressed his concern about soldiers' retention of basic common military skills and the in-unit continuation training problem that it created. The Army School of Training Support (ASTS), as the Systems Approach to Training (SAT) proponent, was tasked to advise DAT on what assistance could be given to unit commanders to ensure that individual resources are used in the most cost effective and efficient manner to maintain performance standards in the requisite skills. It so happened that, at that time, the US Exchange Officer serving at ASTS had been involved in the research, conducted by the US Army Training Board (ATB) and the Army Research Institute (ARI), into skills retention and the development of computer prediction models based on task difficulty. Thus, based on that acquired US knowledge and other experience gained using the British Army three factor, or DIF, task analysis model, ASTS was able to advise the DAT that the state-of-the-art was such that investigation of the skills retention problem was feasible, and provided the US Army retention model was culturally transferable to the British Army context, a management of training decision aid could be developed for unit commanders.

The Tasking. It was apparent that a two stage study would be required to achieve the desired outcome: the first involving a field validation of the US retention model as applied to a limited number of representative tasks, and the second, a comprehensive survey of all the common military tasks for assessment of those factors identified as affecting retention, and estimation of the minimum continuation training required to ensure sustainment of the requisite skills. This advice was accepted by the DAT, who subsequently formally tasked ASTS to conduct such a study.

US Army Experience

ARI Research. The empirical evidence on which the UK study design depends is that obtained by the USATB and ARI respectively in the development of a skills retention model for predicting the average performance of a task, and a users' decision aid for predicting the percentage of soldiers who will perform a task correctly. (It should be noted that both models were based on 'de post facto' rather than 'a priori' reasoning). These became known within ASTS as the 'by step' and 'whole task' models. The most important feature revealed by the ARI 'whole task' research was that task difficulty is the major factor affecting skills retention, and that it consists of a number of identifiable components. ARI subsequently developed a nine-item checklist of task characteristics which could be used to derive a numerical value, the 'magic number', for task difficulty. In essence, the 'magic number' is the sum of loadings placed on specific responses to questions posed for each of the nine items: the loadings having been obtained from factor and regression analyses of field data.

The checklist was subsequently modified by the ATB to consist of ten items as follows:

Are job/memory aids used?
How good are the job/memory aids?
How many steps in the task?
Is there a fixed step sequence?
Is there an inter-step logic?
Does the task have a time limit?
What are the mental requirments?
How many items of information have to be memorised?
How hard are the items to remember?
What are the physical/motor skills demands?

Based on this evidence, predictions could be made about retention of skills and knowledge implicit to a task.

McFan and Gray Associates. Concurrent with the ARI research, ATB had addressed the problem of skills retention by collecting field data on the number of steps performed correctly in the execution of a task. This 'by step' approach was an extension of some earlier, unsolicited research conducted by McFan and Gray Associates (MGA), for the ARI, which focussed on the number of errors committed in performing individual task steps. Retention was measured from a known start point of 'competence', which MGA had defined as 'N' correct successive performances of the task during initial training. Despite the fact that the MGA predictions in skills retention were better than anything else available at that time, the prediction model was dependent upon first trial error data as the sole parameter of task difficulty, and this placed the model on a somewhat doubtful foundation.

ATB Research. In the development of the MGA model, the ATB considered task categorisation (instead of errors made as the indicator of task difficulty) and method of training as the factors affecting the retention of skills and knowledge. With regard to the former, the ATB designed a five-item checklist similar in some respects to that developed by the ARI. Ultimately, however, it failed to produce as good a fit between the predicted and actual performance data as that obtained using the ARI ten-item check list. In the case of training, four methods were identified which offered increasing levels of practice within the period of training to increase the probability of retention for tasks as they increased in difficulty. (The validity of this approach had been established by the research of Annette et al into the effectiveness of spaced versus block training.) The four methods identified were:

Single session: Train to proficiency then test

Double session: As for 'single session', but provide a second 'revision' session prior to testing.

Progressive: Train the task gradually, using a series of three to five sessions, reaching correct performance towards the end of the sequence.

Comprehensive: As for 'progressive' but giving at least two more complete sessions of continuation training.

Subsequent field trials revealed that the ATB predictions, based on post trial data, task characteristics and training method, correlated better with the obtained data than did the MGA predictions. Other interesting data that emerged from the US programme of research was that for revision purposes only half to one third (depending on the training method utilised) of the initial training time was required to ensure retention.

British Army Experience

The Sandhurst Review. Although prior to 1984 the British Army had not investigated skills retention 'per se', it had, over a period of time, refined a task analysis model that utilised subjective assessment of the three factors difficulty, importance and frequency of performance of tasks, to enable training decisions to be made. The obvious drawback of such a model is that it is dependent upon qualitative, which can be notoriously unreliable, rather than quantitative assessments. It was not until the fall of 1984 that three events conspired to require the ASTS to address skills retention, albeit indirectly. First, ASTS was tasked to provide a consultancy team to assist the Royal Military Academy, at Sandhurst, in a review of its courses as part of the overall Army Review of Officer Training and Education (ROTE) Study. Secondly, the newly arrived US Exchange Officer at ASTS had been involved in the USATB/ARI skills retention research programme, and he was deployed on the Sandhurst phase of the ROTE Study. The final event was the realisation by the ASTS consultancy team, during the initial planning of an analysis of the tasks performed by a young officer in his first appointment, that it was without a quantitive tool for assessing what training treatment would be required to enable the officer cadets at Sandhurst to successfully perform the tasks identified by the job survey. The immediate solution was to adapt the US Army retention models, without prior cross validation, by incorporating the UK DIF (three factor) model, and create a training decision aid. This aid offered nine possible training treatments for a task based on the 27 (ie 3^3 matrix) combinations of DIF rating: each factor being rated on a three point scale. Using this aid, worst case estimates were derived for the training required to address all young officer tasks, and this data was used in the planning of the new Sandhurst course. Despite the obvious criticism that the consultancy team applied the US model without testing its ability to transfer to the UK situation, the course of action adopted had utility: training decisions were made on a more objective footing than could otherwise have been achieved, and these decisions made sense to the training staff.

The Formal Retention Study

By the time that ASTS was formally tasked to conduct a study of common military skills retention, there had been significant development of the training decision aid derived from the integrated US retention and UK DIF models. This development work had been effected by the same military training consultants who had been engaged on the Sandhurst review and who were now deployed, with the addition of a senior psychologist, on the formal retention study.

DIF Ratings. These developments included refinement of the original UK three point scale definitions of difficulty and frequency by providing them with an empirical foundation. This was achieved by relating the definitions to various characteristics manifest in the US Army trials performance curves. Thus, 'very', 'moderately' and 'not difficult' were associated with

three distinct bands of 'whole-task' retention curves. Frequency definitions were obtained by the application of the 90% rule to the 'by step' and 'by whole task' performance curves (90% was the minimum percentage performance considered acceptable by the consensus of battalion and company commanders). Thus, examinaton of the 90% 'whole task' curves revealed significant dispersion at the two week point, whereas for the 90% 'by step' curves such dispersion occurred after eight weeks: these two points became the critical cut-offs between 'very', 'moderate' and 'not' frequent.

Training Methods. Although the training decision aid developed during the Sandhurst consultancy initially depended, amongst other things, upon the four training methods listed by the US ATB, five additional methods were identified and incorporated into the aid. However, during subsequent developmental work, this nine-method list was found to be not only too cumbersome, but closer scrutiny revealed that three of the additions were in fact only variants of the original ATB methods. The list was therefore reduced to six methods, consisting of the original four, plus 'explain and demonstrate only' and 'leave to field unit'.

Training Decision Aid. The major advance in developing tools for the retention study was completion of a microcomputer program, with accompanying users' manual, for the training decision aid. The program was designed to run on an Apple II variant utilising the DOS 3.3 operating system. Based on an input of difficulty rating (obtained using the US 10-item check list) and the training method used initially to train a task, the program will deliver the following output:

a. Predicted percentage of soldiers that can be expected to perform a task correctly after a given interval of time from last correct performance eg. after 12 weeks, 72% of the soldiers would still be able to perform the task correctly.

b. Predicted average 'by step' performance on a given task over time eg. the soldiers will still be able to perform 92% of the performance steps correctly after 17 weeks.

c. Predicted length of time soldiers can go without training on a task and yet still maintain a given proficiency level eg. the soldiers will be able to go 20 weeks and still perform the task 80% correctly.

Field Trials

Sample Size. The aim of the field trials is to collect data on the task performance of randomly selected soldiers (from an opportunity sample) which can then be used to validate the US retention model. A random sample of 300 soldiers will be drawn from nominated units located in the UK and Germany ie. 150 soldiers from each Command. Currently, sampling frames are being developed to include an additional 100 soldiers to cover trial attrition due to sickness, leave or other priority duty. The sample of 300, drawn from a total population of 132,000 male soldiers in the active army, will satisfy the conditions for a confidence level of at least 90% in the results. In addition, it provides three subgroups of 100 subjects each, a nicely rounded and convenient size.

Personnel Criteria. As the US Army research did not present any evidence on the effect of personal variables on skills retention, it is intended to investigate such variables during the course of the UK study. Five personal variables have been selected on the basis that they relate to job experience, and because that data is readily available from a central computer facility. The variables are:

Arm/Corps
Summed Selection Group (SSG) ie. 'intelligence'
Military Service
Age
Rank

A personal variables matrix has been devised to ensure that the full British Army range of each variable is represented in the sampling frame. An upper limit of eight years service and rank of staff sergeant has been imposed, because the bulk of the total population falls within those limits.

Tasks for Testing. A list of 187 basic military tasks has been composed from which approximately 30 representative tasks will be selected for testing during the field trials. The list includes all the tasks for which there is mandatory annual testing, as laid down in Army Training Directives, and a number of common military subjects listed in the recruits' syllabus for which there is no mandatory annual testing requirement. The inclusion of the latter tasks is to ensure that the UK task list encompasses the full range of the US Army skills retention model. Currently a panel of subject matter experts and training development personnel is being convened to assess the 187 tasks on difficulty. This assessment will be a two phase process during which the tasks will be assessed initially on the basis of a yes/no response to a five-item checklist, thus creating a 2^5 matrix covering all difficulty values. The second phase will require the panel to select 32 tasks, one from each difficulty level, to be the representative tasks for trials purposes. These representatives tasks will then be reassessed for difficulty, using the US Army 9 and 10-item checklists, to obtain the 'magic numbers' for each task. Predictions will then be made, using the US retention model, on 'by step' and 'whole task' performances.

Field Testing. The test population, provided by the UKLF and the BAOR, will be divided into three subgroups and be tested on the representative tasks according to the following schedule:

	Elapsed time in months			
Group	0	1	3	6
A	*	*	*	*
B	*		*	*
C	*			*

At the start of the trials all the soldiers will be trained to a known competency level in each task. Thereafter, those soldiers failing to demonstrate task mastery during subsequent testing will be retrained on the failed task. It is intended to collect data on both 'by step' and 'whole task' performance. There are two reasons for collecting data on only four occasions over a six month period, and they are; first, it is manifestly

apparent from the US Army data that the most significant effect of skills decay occurs within six months of training; and secondly, the field army cannot cope with the resource demands of testing on more than four occasions during the trials period.

Trials Programme. Originally it had been planned to complete the field trials (validation) stage of the study by April 1987. Unfortunately, higher priority field commitments and other administrative constraints have seriously delayed the trials programme. The trials have had to be reorganised in the UK for the period June to December, 1987, and in Germany for the period January till July, 1988. This will have a knock-on effect on the Stage 2 survey.

Trials Data. The trials data will be used to plot performance 'curves' for each task on a 'by step' and 'whole task' basis. These actual performance graphs will then be tested for best fit. If such statistical tests prove unsuccessful, the gross data will be grouped according to personal variables and be resubmitted for test-of-fit with the predicted data. An unsucessful result from this exercise will require the prediction model to be restructured.

Stage 2 of the Study

Survey. The intended survey will not be implemented until either the US retention model has been proved valid in the British Army context, or the model has been restructured to accommodate any UK variation. The aim of the survey is to disseminate a questionnaire, to British Army units dispersed world wide, and solicit DIF ratings, using a given three point rating scale, for all the common military tasks. The returned data will be used to further refine the training decision aid.

Training Decision Aid. The ultimate aim of the retention study is to produce a tool that will enable unit commanders to plan effective and efficient continuation training programmes. The intention, therefore, is to modify the current computer version of the training decision aid to respond to DIF inputs rated on a three point scale: the desired output will be predictions of task decay rate, and statements on the required training strategy to repair the decay, or to sustain the task skills. The modified aid will be adapted to paper format for distribution to unit commanders.

Concluding Remarks

Close scrutiny of the US and British Armies' skills retention studies will reveal a number of differences in design, the most important of which are:

a. The UK study includes cognitive tasks: The US study did not.

b. The UK test population includes soldiers of up to eight years service: the US test population consisted of first term soldiers only.

c. The focus of the UK study is on continuation training: the US study was based on initial training.

Thus, on methodological grounds, the UK study is open to criticism for trying to generalise the retention model from the particular US Army situation. Notwithstanding such criticism, if it works it will be to the benefit of all.

BEST COPY AVAILABLE

Review of Air Force Task Identification Methods and Data Bases

Sharon K. Garcia
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

INTRODUCTION

The Air Force uses several methods to collect and analyze data on the tasks and task performance requirements of Air Force jobs. These methods and their associated task data bases, support varied Manpower, Personnel, and Training (MPT) research applications and decision making. Perhaps the most commonly utilized are the a) Logistics Composite Model (LCOM); b) Maintenance Data Collection System (MDCS); c) Logistic Support Analysis (LSA); d) Occupational Survey Methodology (OSM); and e) Instructional Systems Development (ISD).

Each of these methods rely on task data; yet each uses different task data. Task data are collected and organized independently and for the most part are diverse in their applications. As such, there are a number of disconnects in our human resources technology base which weaken the MPT decision process and preclude a fully coordinated MPT analysis methodology.

One disconnect is represented by the disparity between methods of task identification based on equipment maintained versus methods based on occupations and personnel. LSA and ISD are representative of the equipment orientation. Task identification is organized around specific hardware; manpower, training , and related analyses depend on hardware characteristics (e.g., maintenance concept, and predicted failure rates). On the other hand, the Occupational Survey Methodology is organized around an occupational group. Thus, while equipment maintained is identified, the level of detail may not be adequate to differentiate specific equipment.

Another dimension of disconnect involves the three domains to which task identification methods and data bases apply: system acquisition, peacetime force management, and training. Within system acquisition, the task identification procedure is based on detailed hardware work unit code structure. It is an equipment-oriented task analysis, and LSA is the method used. After the system is fielded, the emphasis shifts from hardware to personnel, and the Occupational Survey Methodology becomes the method for conducting occupational-personnel oriented analysis. For identification of training needs and development of training standards, ISD and OSM provide the necessary data sources for making training decisions.

While these disconnects appear to represent a fatal flaw in the MPT process, it must be kept in mind that when each method was originally developed, there was a specific goal in mind; namely to respond to an M, a P, or a T need. Little thought was given to how task data collected for one application could be related to task data used for another application to

****

Author's Note

The author gratefully acknowledges the contributions to this paper provided by published reports and working papers by Dr Walter E. Driskill (MAXIMA Corp.) and Mr Edward S. Boyle (AFHRL/LR).

provide a coordinated MPT data base. Thus, no single present system fully serves all MPT uses. Across the various task data methods, however, there is a wide variety of information that could be highly useful for integrating MPT analyses if a method of aggregation were developed.

The Air Force Human Resources Laboratory has begun research aimed at providing a sound basis for task identification and evaluation. This effort, called the Task Identification and Evaluation System (TIES), will seek to develop specifications and criteria for measuring, collecting, analyzing, and managing task data in the Air Force. Implementing these specifications and criteria through modifications and extensions of current methods will help streamline data collection efforts and help in establishing a common frame of reference for the diverse applications of task data. The purpose of this paper is to review each of the five methods currently in use by the Air Force for task identification, description, and analysis that have been selected for initial development of a TIES.

TASK IDENTIFICATION METHODS

Logistics Composite Model

The Logistics Composite Model (LCOM) is a multi-functional computer simulation model designed to determine the resource requirements of emerging weapon systems. Resources may include maintenance personnel, facilities, support equipment, and supply items. "The necessary inputs to LCOM include: daily mission schedules (defining when aircraft are to fly and for how long); aircraft servicing networks (defining the tasks, times, and resources to prepare and launch an aircraft at its scheduled time and service it upon return); corrective maintenance networks (defining the tasks, times, and resources to fix each subsystem when it breaks); failure rates (defining how frequently each subsystem is likely to require corrective maintenance); and quantities of each resource (e.g., aircraft by type, personnel by AFSC and shift, LRU spares and support equipment)" (Richards, 1983). The LCOM simulation uses these inputs to simulate a sequence of maintenance activities that would take place in an operational unit flying a specified schedule. Aircraft are preflighted, loaded with munitions, taxied, flown, recovered, and maintained. The simulation tracks the number of personnel and physical resources used to run the operation as each aircraft is flown.

Output of the LCOM simulation include statistics describing the simulated operations that can be used to answer "what if" questions. For example, a manager may ask, "If specific logistics or manpower resources were limited, how would sortie generation be degraded?" or, "If a designated sortie rate were flown, what would be the manpower requirements for specific AFSCs?" Using LCOM, the manager can thus see the results of a new policy or restriction on the system that was simulated and then make the best decision on a course of action.

Maintenance Data Collection System (MDCS)

Much of the maintenance data used as input to the LCOM come from the Maintenance Data Collection System (MDCS). This system contains detailed maintenance data for on-equipment, off-equipment, and depot aircraft maintenance work. Data for input to the MDCS is extracted from an Air Force

Technical Order (AFTO) Form 349. For every maintenance task performed, a maintainer must complete an AFTO Form 349. Relevant data contained on this form include: the workcenter performing the work; type of maintenance performed (e.g., preflight inspection, unscheduled maintenance, and servicing); work unit codes identifying the system, subsystem, and component on which the work was performed; specific action taken (e.g., bench-check, troubleshoot); time taken to perform the task; crew size needed; and employee identification number.

Data collected on the Form 349s are input to a centralized data bank providing maintenance data by base and major command From these data a variety of analyses can be performed to provide data on the reliability and maintainability of equipment and weapon systems, manhours, weapon systems readiness, and supply requirements. The primary uses of the data appear to be in the areas of reliability and maintainability as well as product improvement. The data is useful for identifying components whose reliability is low or for which maintenance is expensive in terms of maintenance manhours and resource replacement. Especially at base-level, the data are used for maintaining benchstock.

Logistic Support Analysis (LSA)

Logistic Support Analysis (LSA) is the application of scientific and engineering efforts undertaken during the weapon system aquisition process to identify, define, analyze, and process logistics support requirements. The general requirements for the conduct of an LSA are contained in MIL-STD 1388-1A, Logistic Support Analysis. The MIL-STD specifies 72 different analyses that can be procured from the contractor which describe various aspects of the logistical support required for the new system.

Data compiled in the conduct of LSA are provided via Logistic Support Analysis Records (LSAR), which yield detailed engineering oriented maintenance task requirements for every hardware item going into a new weapon system. Much of the information regarding the maintenance tasks are contained on data sheets entitled the Task Analysis Summary, Maintenance and Operator Task Analysis, and Skill Evaluation and Justification. The Task Analysis Summary provides a detailed listing of tasks that may be performed on an equipment item. Tasks performed on existing comparable systems serve as the basis for generating the task lists. The Maintenance and Operator Task Analysis sheet consolidates the operations and maintenance tasks identified for each repairable unit and indicates the necessary support requirements (e.g., facilties, tools, training equipment). Tasks are identified as a composite of the equipment item name, estimated task time, frequency, and estimated personnel by skill level and specialty. The Skill Evaluation and Justification data sheet is used to describe and justify any new or modified personnel skills required to support an emerging weapon system/equipment.

LSA data provide an array of valuable information for developing MPT requirements for new weapon systems. The data are useful for determining the impact of design features on logistic support, for determining the impacts of proposed logistics support systems on system/equipment availability and maintainability, and for providing data for life cycle costing and logistic support modeling.

Occupational Survey Methodology

Occupational Survey Analysis is conducted by the USAF Occupational Measurement Center (USAFOMC) under the provisions of AFR 35-2, Occupational Analysis, and ATCR 52-22, Occupational Analysis Program. Data generated through the analysis of Air Force occupations or specialties have been used for a variety of MPT applications. The key uses of the data, however, are in the areas of personnel and training.

The basis for analysis is a job/task inventory, which is an exhaustive list of tasks that may be performed in an Air Force occupation. The inventory is developed using subject matter experts in the occupation and is then administered to a large sample of occupational incumbents. Completed inventories yield six basic types of information: a) tasks comprising the Air Force specialty; b) percentage of incumbents performing the tasks; c) relative percentage of time spent on each task; d) relative difficulty (time needed to learn to perform) of each task; e) relative training emphasis recommended for each task; and f) summaries of background information (e.g., job satisfaction indicators, and equipment used and/or maintained). These data are generated and analyzed using the Comprehensive Occupational Data Analysis Programs (CODAP), a very powerful set of computer programs. Key products of the analysis are statistical descriptions of jobs and analyses of what people do by pay, grade, skill level, and major commands.

Over the past 19 years, occupational analysis data have been used for a variety of purposes. For example, occupational classification structures described in AFR 39-1, Airman Classification Manual are based to a large extent on survey data. The data can be used to identify dissimilarity of work performed across a specialty to serve as a basis for occupational restructuring and development of new Air Force specialties. A further use for classification has been to use task difficulty data to estimate aptitude requirements for Air Force enlisted occupations. Occupational data are also used by the training community to identify tasks for training and to develop appropriate training plans and standards.

The Occupational Survey Methodology and CODAP exist as one of the most powerful and versatile task identification and analysis approaches currently in use by the Air Force. Since its development, CODAP has been universally adopted for use throughout DOD, in other government agencies, in industry, and in academia.

Instructional Systems Development

In the early 1960s, the Air Force began the development of a systematic approach to the development of training known as Instructional Systems Development (ISD). This approach is currently being used by the Air Force to provide a systematic, flexible decision-making process for instructional programs, and is regulated by AFR 50-8, Policy and Guidance for Instructional Systems Development. ISD is used for planning, developing, and managing training programs to assure that personnel learn the necessary skills and knowledge needed to perform their Air Force jobs. It is applied to virtually all new and modified training programs.

In practice, ISD is a model, consisting of five broad steps. These steps

are a) analyze system requirements; b) define education and training requirements; c) develop objectives and tests; d) plan, develop, and validate instruction; and e) conduct and evaluate instruction. Descriptions of these steps and how they are applied can be found in AFP 50-58, Handbook for Designers of Instructional Systems. ISD can be applied to the design of training for an entire career ladder or more narrowly focused around particular systems within the career ladder.

Many products are produced by the ISD approach, however, the two that are of most interest to TIES are the Job Performance Requirements (JPRs) and Training Requirements (TRs). JPRs provide a listing of tasks that must be performed to accomplish a job, along with the standards for adequate performance of those tasks. For the most part, the Occupational Survey Report (OSR) serves as a starting point for the ISD practioner to use in identifying the tasks provided in the JPR for a given specialty. If an OSR is not available, this information must be obtained from subject matter experts who assist the ISD practioner to break down a job or jobs into duties, tasks, and subtasks. The task identification process is iterative and continues until an appropriate level of detail exists to support the training need. Once tasks have been identified in a JPR as requiring training, TRs are then established for each task. TRs reflect the analyst's judgements regarding the necessary skills, knowledge, and aptitudes required to satisfy the JPRs (Joyce, Garcia, and Collins, in preparation).

FUTURE DIRECTIONS

A major goal of TIES is to integrate the five methods discussed in this paper. While only a brief description of each method is provided here, a more thorough review has been accomplished in an attempt to identify similarities to aid in linking the five methods. Common to all is the identification of tasks. Hence, the next objective in TIES will be to develop computerized procedures to cross-match or map task statements across each of the five methods. Once developed the techniques will permit the interfacing of various tasks/task data across methods so that relevant data could be readily accessible, collated, analyzed, and reported in a useable format. A TIES presents a formidable task. But, if successful TIES could make a significant contribution to the Air Force manpower, personnel, and training decision and policy-making process.

REFERENCES

Joyce, R. P., Garcia, S. K., & Collins, D. L. (in preparation). Review of Air Force task identification methods and data bases. Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Richards, Jr., E. R. (1983). Building and Operating the Logistics Composite Model (LCOM) for New Weapon Systems, Part A. Wright-Patterson AFB, OH: Aeronautical Systems Division, Air Force Systems Command.

Relationship of SQT Scores to Project A Measures

Jane M. Arabian and Jeanne K. Mason
U. S. Army Research Institute
Alexandria, Virginia

The Army develops and administers Skill Qualification Tests (SQT) to soldiers in many of the Military Occupational Specialties (MOS). The testing program was originally intended to diagnose needs for training. However, SQT scores are also used for personnel management decisions (e.g., promotion policy decisions, distribution goals for soldier quality, etc.)

Although SQT are not developed for all MOS, particularly the smaller MOS, the MOS that do have SQT represent a variety of occupational specialties and a large proportion of Army accessions. Further, the test administration and score reporting program is well-established, rendering the SQT scores readily accessible to the Army research community. Since these skill tests are administered to soldiers after school training (AIT), when soldiers have had experience performing in their specialty, the SQT scores have been employed as proxy measures of job performance to support personnel policy decisions. However, the assumption that SQT can be validly used as a measure of job performance has not been tested directly.

Converging evidence does suggest that SQT are viable measures of job performance. For example, the distribution of SQT scores by ASVAB (Armed Services Vocational Aptitude Battery) scores, more specifically Aptitude Area (AA) composite scores from ASVAB, were employed by proponent schools to support particular MOS AA entry score requirements. Along with the proponents' input, the Army's submission to Congress on Army manpower quality goals also included data on the relationship of written and hands-on performance scores, obtained from TRASANA, with ASVAB scores (Office of the Assistant Secretary of Defense, 1985). While both sets of data (SQT/ASVAB and TRASANA data/ASVAB) produced similar results, namely a positive relationship between ASVAB and the performance measures, direct examination of the relationship between SQT and TRASANA hands-on and written test scores was precluded by the small number of cases available with both sets of scores. Consequently, it was not possible to determine the validity of SQT scores as measures of job performance at that time.

With the collection of job performance data from the 1985 (concurrent validation) testing phase of the Army's Project A, "Improving the Selection, Classification and Utilization of Army Enlisted Personnel", and the merging of SQT data into the Project's research database, it has become possible to validate SQT scores against independently developed criteria of job performance. The Project A measures selected for this SQT validation research include paper and pencil measures of school knowledge and job knowledge as well as a work sample (hands-on) measure of job proficiency. If the results of this research demonstrate a strong positive relationship between SQT scores and the Project A measures, then it could be confidently asserted that the SQT are valid measures of job performance. Use of SQT data would then be empirically justified as a measure of job performance for personnel management decisions.

Method

Subjects

The subjects in the present research are a sub-sample of the Project A concurrent validation sample. The data for Project A were collected from June to November 1985. The soldiers were all at Skill Level 1 with 18 to 24 months

experience in the Army at the time of testing. The sub-sample had 3,117 soldiers with test scores for each measure of interest (SQT and three Project A measures as well as ASVAB). Soldiers from the following eight MOS were represented in the sub-sample: 11B--Infantryman; 13B--Cannon Crewman; 19E--Tank Crewman; 31C--Radio Teletype Operator; 63B--Light Wheeled Vehicle/Power Generation Mechanic; 64C--Motor Transport Operator; 71L--Administrative Specialist; 95B--Military Police.

Measures

The SQT is a multiple choice, written test of overall MOS knowledge designed for a 2-hour administration period. Soldiers are tested by MOS and Skill Level. Tasks included in the SQT are randomly selected from the Soldier's Manual for a given MOS. Approximately 20-35 tasks (maximum of 161 items) are included in an SQT. The notice announcing the test includes a list of 150% of the tasks that will appear on the test. The overall SQT score is a percentage computed by adding all scores from each task and dividing the sum by the total number of tasks on the test. Further information about SQT development and administration is available in the SQT Test Development Manual (TRADOC, 1983). SQT scores used in the present research were from the 1985 administration with the exception of MOS 31C, whose scores were from 1986.

The Project A School Knowledge tests (K3), also labelled Job-Relevant Knowledge tests, were developed to measure the cognitive component of training (school) success. Test items were based on, e.g., the Army Occupational Survey Program and Program of Instruction (course curriculum) information for each MOS. All items were reviewed by job incumbents, school trainers and appropriate MOS training proponents for content, accuracy, etc. The K3 test for each MOS contained approximately 150 multiple choice items and was administered in a 2-hour period. A detailed description of the test development procedure and psychometric properties of the tests can be found in R. Davis, G. Davis, Joyner, and de Vera (1985).

The development process and psychometric properties of the Task-Based MOS-Specific, Job Knowledge (K5) and Hands-On (HO), measures are described in C. Campbell, R. Campbell, Rumsey and Edwards (1985). Briefly, the job performance domain for each MOS was determined from several sources, including: the Army Occupational Survey Program results, Soldier's Manual of Common Tasks, MOS-specific Soldier's Manuals, and input from the MOS proponent agency. Subject matter experts provided judgments of task criticality, difficulty and similarity. Separate panels of subject matter experts in each MOS used the judgments to select MOS tasks for K5 measure development. The written K5 measures cover some 30 tasks and have approximately 150-200 multiple choice items which require about 2 hours for administration. The HO measures are a sub-set of 15 of the 30 tasks covered in the K5 measure for each MOS. Aside from logistical constraints (e.g., tasks too hazardous to test), tasks selected for testing in the HO mode entailed physical strength or skilled psychomotor performance, performance within a time limit, many procedural steps, and/or steps that are uncued in their normal sequence.

Data and Analyses

A workfile was created from the Project A longitudinal research database. The workfile contained Skill Level 1 SQT score, average percent correct K3, K5 and HO scores and ASVAB AA composite score for each case (subject). The AA score is used in the Army enlistment process as the primary classification eligibility measure for each MOS. Univariate descriptive statistics and correlation analyses were performed using the SAS statistical package.

Results and Discussion

The univariate descriptive statistics for each performance variable (SQT, K3, K5 and HO) by MOS are presented below. There is satisfactory variance and range in the data to permit further analyses.

Table 1

Descriptive Statistics for Each Variable By MOS

MOS		SQT	K3	K5	HO
11B	M	79.27	59.89	60.16	71.99
	SD	10.75	12.72	11.46	7.73
	N	614	599	594	598
	Min	44	17	25	45
	Max	100	84	86	93
13B	M	73.18	54.03	60.64	63.37
	SD	11.50	11.37	10.69	11.07
	N	547	529	528	500
	Min	28	17	24	34
	Max	100	79	84	91
19E	M	74.30	66.08	62.08	76.80
	SD	8.78	12.73	9.81	8.01
	N	433	418	396	40?
	Min	8	20	34	50
	Max	94	86	85	93
31C	M	74.79	59.80	59.68	70.51
	SD	8.72	11.55	10.23	8.68
	N	313	298	280	295
	Min	38	21	27	37
	Max	91	84	82	90
63B	M	62.21	59.54	63.52	84.85
	SD	9.18	12.43	10.76	5.44
	N	525	505	488	472
	Min	23	20	27	62
	Max	86	84	86	96
64C	M	82.17	61.13	58.25	71.49
	SD	7.10	12.27	9.83	8.17
	N	561	547	548	521
	Min	52	20	30	43
	Max	98	84	81	89
71L	M	71.44	59.35	57.06	63.34
	SD	13.58	11.18	10.18	10.16
	N	431	416	421	415
	Min	21	24	30	29
	Max	99	86	84	90
95B	M	78.66	58.79	62.00	70.71
	SD	5.89	10.12	9.56	6.87
	N	628	610	606	603
	Min	49	19	26	48
	Max	95	81	86	85

Correlations were obtained between the appropriate AA composite score, SQT, and each Project A performance measure by MOS for cases with complete data. The correlations, in the table below, are generally consistent with data from other studies. The SQT are positively correlated with the ASVAB AA composite scores as well as with the Project A performance measures. Since the focus of this report is on the relationship between SQT and other measures (i.e., K3, K5 and HO) of job performance, weighted averages using the Fisher z transformation were computed across MOS only for the SQT and Project A correlations and the intercorrelations among the Project A measures. As would be expected, the correlations between same-mode measures (paper and pencil, e.g., SQT:K5, K3:K5) are somewhat higher than the cross-mode (paper and pencil vs hands-on, e.g., K3:HO, SQT:HO) correlations.

Table 2

Correlation Coefficients: Cases With All Variables

MOS	AA COMPOSITE	N	AA:SQT	AA:K3	AA:K5	AA:H0	SQT:K3	SQT:K5	SQT:H0	K3:K5	K3:H0	K5:H0
11B	CO	502	.432	.439	.522	.343	.525	.566	.381	.660	.387	.616
13B	FA	411	.293	.340	.410	.142	.488	.503	.433	.694	.442	.443
19E	CO	315	.590	.494	.577	.309	.565	.616	.395	.726	.341	.491
31C	SC	220	.524	.392	.402	.322	.572	.537	.410	.686	.460	.482
63B	MM	390	.501	.641	.542	.299	.588	.597	.367	.735	.412	.356
64C	OF	467	.490	.435	.473	.323	.391	.465	.374	.634	.368	.444
71L	CL	349	.474	.508	.544	.378	.570	.536	.497	.720	.611	.602
95B	ST	463	.405	.403	.342	.331	.387	.355	.335	.583	.278	.364
	N = 3117											
	$\bar{r}$						.503	.517	.395	.676	.409	.479

The correlations between SQT and the three Project A measures were corrected for attenuation and range restriction. The reliability estimates for the Project measures, used for the attenuation correction, are presented below. SQT reliability estimates were not available; therefore, the corrections were based on only the Project A measures.

Table 3

Internal Consistency Reliability Estimates

	Test		
MOS	Hands-on	Job Knowledge	School Knowledge
11B	.54 (682)	.89 (678)	.93 (684)
13B	.75 (612)	.85 (639)	.89 (640)
19E	.63 (474)	.89 (459)	.93 (485)
31C	.79 (341)	.86 (326)	.93 (349)
63B	.52 (569)	.87 (596)	.94 (612)
64C	.64 (640)	.85 (668)	.90 (669)
71L	.73 (494)	.82 (501)	.88 (493)
95B	.58 (665)	.84 (665)	.88 (674)

Note: The second entry (in parentheses) is the sample size.

With respect to the correction for range restriction, a formula was employed which is appropriate for the correlation of a new measure, such as the Project measures, with an existing criterion, the SQT, when selection has been made on a third variable, in this case AA composite score (Guilford, 1965). The correlations between SQT and the Project A measures, corrected for attenuation and range restriction are presented below. Again, weighted averages of the validity coefficients across MOS were computed. It can be seen in the table below that SQT is strongly correlated with each of the independent measures of job performance. The somewhat lower average correlation between SQT and HO scores may be attributable at least in part to measurement mode differences (written vs hands-on).

Table 4

Cases With All Variables: Corrected For Attenuation and Range Restriction

MOS	SQT: K3	SQT: K5	SQT: HO
11B	.646	.674	.631
13B	.593	.625	.534
19E	.703	.765	.606
31C	.679	.675	.563
63B	.756	.768	.646
64C	.661	.729	.674
71L	.705	.698	.676
95B	.689	.653	.665
$\bar{r}$	.679	.699	.652

In order to compare scores across the four measures, equi-percentile equating was performed; the results are presented below. Since 60 is used as the passing score for SQT, the percentile for a score of 60 on SQT was used to determine comparable (in terms of percentile) scores for the K3, K5 and HO measures. Thus, 11B soldiers with an SQT score of 60 are in the 6.03 percentile. For the K3 measure, an 11B soldier in the 6.03 percentile would have a score of 37. The percentile for SQT scores of 60, 70 and 80 were determined along with the comparable scores on the Project A measures. Scores for SQT, K3, K5 and HO tests at the 50th and 85th percentile were also calculated.

The lower scores on the Project A measures, compared to the SQT scores, suggest that the Project tests may have been somewhat more difficult. Whether or not the apparent differences in difficulty can be attributed to test content versus the opportunity to study for the test cannot be ascertained. However, it should be noted that SQT test dates with 150% of the tasks to be covered are published before testing; this is not the case with the Project A testing.

Table 5

Equi-Percentile Equating

		SCORE						SCORE			
MOS	%ILE	SQT	K3	K5	HO	MOS	%ILE	SQT	K3	K5	HO
11B	6.03	60	37	40	59	63B	39.43	60	57	61	84
	18.89	70	49	49	65		8..91	70	70	74	89
	48.86	80	62	61	72		98.86	80	81	83	94
	50	80	63	61	73		50	62	60	64	85
	85	90	72	72	80		85	71	72	75	90
13B	13.35	60	41	48	50	64C	0.71	60	24	32	47
	40.40	70	52	59	61		5.88	70	39	42	58
	72.58	80	61	68	69		35.12	80	58	55	69
	50	73	55	62	63		50	83	63	59	72
	85	85	66	72	76		85	69	73	69	79
19E	5.54	60	40	45	62	71L	21.11	60	50	48	55
	27.71	70	61	57	71		43.39	70	58	55	62
	74.13	80	75	69	84		71.23	80	66	63	69
	50	75	69	64	78		50	73	60	57	63
	85	82	78	72	85		85	85	70	68	73
31C	5.43	60	40	43	55	95B	0.48	60	25	31	48
	32.27	70	56	55	67		8.92	70	43	48	60
	69.65	80	66	65	75		61.94	80	63	65	73
	50	75	61	61	71		50	78	60	63	71
	85	83	71	70	79		85	84	68	71	77

Note. N.B. equating is approximate values rounded to nearest whole numbers.

The equi-percentile equating performed on this data set should not be taken to suggest cut off scores for the Project measures. (Nor would it be reasonable to alter the SQT cut off given only the data presented here.) While it would be possible to apply standard setting procedures to the Project A data, it would not be advisable to use the SQT score of 60 to set standards on the other measures. The primary reason for this position is that the SQT cut off score of 60 was not necessarily derived empirically or validated against a definition of minimally acceptable performance. In order to evaluate the SQT cut off, and perhaps determine cut offs on the Project A tests, additional information would be needed about satisfactory and unsatisfactory performance levels.

Conclusions

Project A research has provided a unique opportunity to validate SQT against independently derived measures of job performance. The research presented in this paper strongly supports the validity of SQT as a measure of job performance. Although only a limited number of MOS were in the sample, the variety of occupations and the consistency of the results suggest that SQT in general (i.e., including MOS not in the sample) may serve as a valid measure of job performance for personnel management decisions. Further research is particularly needed, however, to validate the SQT cut off score.

References

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1985). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). Alexandria, VA: Army Research Institute.

Davis, R. H., Davis, G. A., Joyner, J. N., & de Vera, M. V. (1985). Development and field test of job-relevant knowledge tests for selected MOS (ARI Technical Report __). Alexandria, VA: Army Research Institute.

Guilford, J. P. (1965). Fundamental statistics in psychology and education (Fourth Edition). New York: McGraw-Hill Book Company.

Office of the Assistant Secretary of Defense (Manpower, Installations and Logistics). (1985). Report to the House and Senate Committee on Armed Services, Quality of Military Enlisted.

SAS Institute, Inc. (1986). Statistical Analysis System, Version 82.4. Cary, NC: SAS Institute, Inc.

TRADOC. (1983). Skill qualification tests (SQT): Policy and Procedures (TRADOC Reg 351-2). Fort Monroe, VA: Department of the Army.

ACKNOWLEDGMENTS

The opinions, views and conclusions contained in this document are those of the authors 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. Project A is an Army Research Institute contractual effort, #2Q263731A792, performed by Human Resources Research Organization, American Institutes of Research, and Personnel Decisions Research Institute. The authors wish to express their appreciation to Winnie Young for preparing the data file used in this research.

Test Validity in RAAF Air Traffic Controller Selection

Stephen J. Elliott
RAAF Psychology Service,
Department of Defence (Air Force Office),
Canberra ACT 2600 Australia.

The Royal Australian Air Force (RAAF) employs approximately 234 Air Traffic Control (ATC) officers, located at most RAAF bases in Australia. To ensure an adequate supply of trained ATCs, about 24 candidates must be selected each year for basic training in ATC procedures. The basic ATC course, which is conducted at RAAF Base East Sale, lasts 19 weeks, involving a 10 week theory phase followed by a 9 week practical phase of training. Training is given in the areas of Flight Planning Office (FPO), aerodrome control and procedural approach control. New graduates of the basic course are employed in FPO and ATC tower duties, whilst an advanced course qualifies ATC officers with 18 to 24 months job experience in the field for radar and approach control duties.

The selection process for RAAF ATC Officers currently involves ability testing and interviewing. As well as passing the RAAF Commissioning battery of verbal and intelligence tests, ATC applicants must attain set levels on tests of spatial ability and, from 1987, speeded simple addition and multiplication. About 35% of applicants are presently 'screened out' on aptitude, whilst those passing the testing must undergo interview by both a psychologist and a recruiting officer. Applicants who are recommended by the recruiting officer are presented to an Officer Interview Board, comprising an ATC specialist, two other RAAF officers and a psychologist, for final selection. Interview selection criteria, covering required personal qualities, abilities and aptitudes, educational attainments, experiences, interests and circumstances, are utilised by selection interviewers.

To date, selection validation has only been conducted against the criterion of performance on ATC basic training courses, from which about 10% of ATC Officer Cadets fail to graduate. Following is a list of the variables thus examined:

VM. A spatial ability test, requiring the visualisation of an aircraft's manoeuvres, and developed to select pilots.
DHT. A speeded information processing test requiring the identification of conflicting or agreeing directional symbols.
MCAT. A speeded test of ability to make predictions from graphically displayed flight data.
UQMR and INSTR. The University of Queensland Map Reading and Instructions tests. UQMR requires the measurement of distances from a map, whilst being periodically interrupted by complex verbal instructions for INSTR, which requires the applicant to complete the test paper according to the conditional logic of the instruction.

TNRA. ACER Test of Number, measuring speed of basic multiplication and addition.
AVAR. Speeded aviation arithmetic reasoning problems.
CPAB-NA. A subtest of the SRA Computer Programmer Aptitude Battery requiring speeded arithmetic estimations.
NSQ. IPAT Neuroticism Scale Questionnaire. A personality test comprising Cattell's 16PF I,F,E and Anxiety scales.
LOQ. Fleishman's Leadership Opinion Questionnaire, with scales of Consideration (C) and Structure (S) as preferred leadership styles.
EPI. Eysenck Personality Inventory, comprising Neuroticism, Extraversion and Lie scales.

Of the variables above, only VM, DHT, MCAT, and UQMR and INSTR were administered in the selection battery. The arithmetic tests, TNRA, AVAR and CPAB-NA, were all given at the commencement of ATC basic training, whilst the remaining personality and attitude tests were administered at the Officer Initial Training Course, prior to commencement of ATC training.

Two validation studies have been conducted with these variables. The first study (Elliott, 1984) compared all test variables (barring the arithmetic tests and tests MCAT and DHT) with overall pass versus fail, overall weighted score in theory subjects and overall weighted score in the practical phase, for 16 ATC basic courses between 1977 and 1981, comprising 106 RAAF trainees. The second study (as yet unpublished) examined the predictiveness of all aptitude, but not personality, variables, against overall pass versus fail and final ratings of performance in practical work, for 6 courses between 1984 and 1986, comprising 58 RAAF and 7 RAN ATC cadets.

Tables 1 and 2, below, show the correlations (corrected, where appropriate, for restriction of range due to selection), and their significance, of the test variables with the training criteria.

So, what can be concluded from the above studies? Study 1 showed disappointing predictiveness for aptitude tests, particularly for practical work. In part, this seems to have been due to some inconsistencies in assessment between courses. Improved assessment techniques employed in Study 2, obtained through use of better standardised examinations and performance ratings, seem to yield generally higher validity coefficients all round, although the sample is small. Clearly, basic arithmetic skills, measured particularly by TNRA-Mult, are the most critical aptitude currently assessed for ATC cadets. Table 3 shows a contingency table of course outcomes by scores on TNRA-Mult, where it can be seen that low multiplication scores are associated with half of all failures, but few pass or higher grades.

Table 1 Validity coefficients of tests against ATC basic course performance for study 1.

Test	Pass/fail	Theory	Practical
VM	0.16	0.13	0.08
UQMR	0.04	0.22*	-.00
INSTR	0.04	-.13	-.03
NSQ-I	-.06	-.17	-.30*
-F	-.15	-.30*	-.12
-E	-.21*	-.01	0.01
-Anxiety	-.15	-.27*	-.25*
-TOTAL	-.26*	-.34**	-.30*
EPI-N	-.13	-.26*	-.23
-E	0.13	-.08	-.01
-L	-.01	0.37**	0.16
LOQ-C	-.05	0.02	-.04
-S	0.32**	-.15	0.27*

* p<0.05, ** p<0.01.

Table 2 Validity coefficients of tests against ATC basic course performance for study 2.

Test	Pass/fail	Practical
VM	0.27	0.25
DHT	-.03	0.09
MCAT	0.17	0.18
UQMR-DIST	0.23	0.34*
INSTR	0.13	0.22
TNRA-Add	0.17	0.43**
-Mult	0.29*	0.50**
CPAB-NA	0.15	0.33*
AVAR	0.16	0.35*

* p<0.05, ** p<0.01.

Table 3 Contingency table of scores on TNRA-Mult against overall grade on RAAF ATC courses in Study 2

TNRA-Mult Score	Distinction/ Credit	Pass	Fail
60 or more	4 (=40%)	5 (=50%)	1 (=10%)
35-59	8 (=24%)	23 (=68%)	3 (=9%)
34 or less	0 (=0%)	3 (=38%)	5 (=63%)
Total	12 (=24%)	29 (=59%)	9 (=20%)

The correlations found for personality and attitude tests, in table 2, are more problematic in interpretation, for two reasons. Firstly, the tests were administered to cadets already selected for ATC, and therefore perhaps less likely than job applicants to distort their responses to cast themselves in a favorable light. Secondly, there is the possibility that the predictiveness observed simply reflects a degree of instructor bias against ATC officers with certain personality features. Unfortunately, administration of personality tests, pre-course, was discontinued in 1980 and has only just been recommenced, so data from the more reliably assessed recent courses are not available.

The correlation of the LOQ Structure scale with performance on course is perhaps best explained in terms of having a favorable disposition to working in a highly structured, procedural work environment.

The training failures from Study 1 scored higher on NSQ than the graduates, but lower than the IPAT general population norms, suggesting that 'neuroticism' per se is not the issue. Rather, it may be that the NSQ is tapping a tendency to repress anxiety, depression (F), emotional sensitivity (I) and low confidence (E), with non-repressors more likely to be assessed as poor performers. Evidence that possibly supports this hypothesis comes from the finding that higher 'Lie' scale scorers on the EPI are likely to do better in the theory phase, perhaps through a common process.

The RAAF Psychology Service is currently conducting a research program that will hopefully throw additional light on these interesting findings. NSQ will be introduced into the ATC selection battery, to see if there is continued predictiveness when social desirability responding could be expected to increase. Perhaps, however, 'revealers' of neurotic features that are undesirable in the ATC world will continue to reveal, even in a selection situation. As Dr A.G.P. Elliott (1981) concludes, in an interesting article on test distortion in real-life selection, distortion on personality scales may still provide useful information. As a check of the validity of selection NSQ scores, the 16PF and EPI will be administered to selected ATC cadets, along with the psychodynamic, projective Defence Mechanisms Test currently being trialled by the RAAF. In two or three years time, the RAAF may have a useful data base on personality characteristics and ATC performance.

Additional research needs to be conducted in two areas. Firstly, the validity of tests against basic training course performance needs to be replicated with advanced training course and actual on-the-job performance as criteria, as perhaps 5% of graduates of the basic course fail to make satisfactory controllers. Secondly, new aptitude measures need to be developed to better measure an applicants potential to learn to develop an air picture, assess a situation for possible conflicts and devise a plan to most expeditiously separate conflicting aircraft. In

this latter regard, ATC cadets are being tested with the computerized USAF PORTABAT testing devices currently on loan to the RAAF, and it is hoped that some useful leads will be gleaned from the range of aptitudes measured by the PORTABAT.

Technological change seems unlikely to change the basic nature of ATC duties until computers can learn to separate aircraft with the same 'creativity' as human controllers. However, those technological innovations likely to influence ATC have the potential to result in the better assessment of both performance and aptitude. One development in Australia is the Tower Simulator being developed at the Aeronautical Research Laboratory in Melbourne, using state-of-the art graphics and instructional software. This seems likely to bring increasing objectivity into the assessment of ATC performance, which in turn will allow more accurate validation data to be gathered for existing and new aptitude measures.

The 'bottom line' of the above research, of course, is the goal of developing greater efficiency and effectiveness in the ATC system, by selecting and training those individuals whose aptitudes and dispositions best fit them for the ATC role. There would seem to be plenty of scope for moving towards this goal.

References

Elliott, A.P.G. Some implications of lie scale scores in real-life selection. Journal of Occupational Psychology, 1981, 54, 9-16.

Elliott, S.J. RAAF Air Traffic Controller Selection: The relationship of psychological test scores and other variables with performance on Air Traffic Control Courses Nos 68-83. Department of Defence (Air Force Office), Canberra, 1984, Psychology Service Research Note 2/84.

TWO FOR THE PRICE OF ONE: PROCEDURAL TUTORIAL AND TESTING IN AIRCREW TRAINING

Josephine M. Randel
ManTech Mathetics Corporation

PROGRAM GOAL

A computer- assisted instruction (CAI) program was designed and developed at Miramar Naval Air Station as part of the training program for aircrew learning to fly the F-14A aircraft. The goal of the CAI program was to prepare the student for the simulator exercises so that valuable time would not be spent familiarizing the student with the cockpit layout and some basic procedures.

Before the development of the CAI lessons, training consisted of the NATOPS Flight Manual readings, lectures, slide/tapes and simulator exercises. While NATOPS readings are necessary, this book was written as a manual rather than an instructional vehicle, so other learning materials were needed. The slide/tapes used in the program are informative, but they do not provide hands-on practice and have a tendency to put active aircrew members to sleep. Thus, there was a need for an interactive medium which would be more congruent with the actual job of flying an airplane.

Flight simulators are excellent for aircrew training, but they are very expensive and there never seems to be enough time available on their schedule. In addition, they usually require one-on-one interaction with an instructor. CAI is a very good preparation for the simulator. It is one step further up on the hierarchy of interactive media above lecture/discussions and provides feedback for every student response.

Although much cheaper than a simulator, CAI can be quite expensive. Estimates of 50 to 700 hours of development time per hour of instruction have been given (String and Orlansky, 1979). The question arises of how to produce a good training program while keeping the development time within bounds.

COST CONSIDERATIONS

Costs of CAI are driven by the authoring system, graphics and the design used. A good authoring system will allow an instructional designer to develop a training program without having to spend time to master a programing language. An authoring system should allow the designer to write text incorporate graphics, ask questions, deliver feedback for correct and incorrect answers and automatically score performance, all through the use of a menu or simple English language commands. In addition, the authoring system should have

enough flexibility to allow for deviations from the prepackaged authoring model. This is the ideal; most authoring systems use either stringent models or require excessive programing.

Any learning program which is to serve as a pre-simulator must have a good graphics capability to produce the desired product, and the less labor intensity involved the better. To be able to enlarge and decrease, rotate and scale a drawing are extremely desirable in terms of economy of production time. The one element that is often forgotten in graphics input is the artist. Good CAI cannot get along with an instructional designer acting as an artist.

The greater the resolution of the CRT on which the graphics are delivered the more realistic the presentation. However, for most purposes where the drawings are mainly horizontal and vertical lines, the requirements are not as stringent.

For all CAI programs, the most important ingredient is the design of the lessons. Consideration must be given to the mix of graphics and text, screen design, the type of response required of the student, feedback, testing and record keeping.

Careful consideration was given to the factors involved in the CAI aircrew training program so that the final product would be effective and efficient. The MicroTICCIT authoring system was a given and a large quantity of well executed graphics were required. Program design would have to be the area where good planning could keep costs of the program within bounds.

PROGRAM DESIGN

A large number of the lessons chosen for CAI presentation involved a procedure of some type, for example, turning on a system or performing cockpit checks. The first step in performing the procedure was to choose the appropriate panel from an outline of the cockpit (Figure 1).

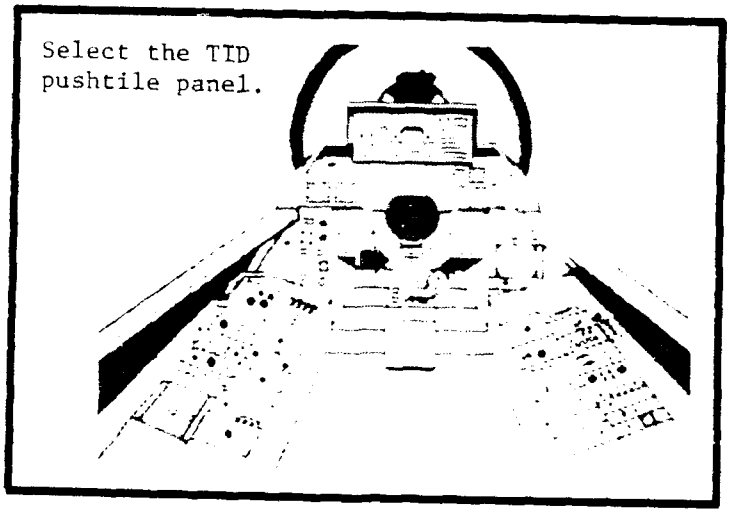


Figure 1. CAI screen of the rear cockpit.

The procedures were to be taught to the student through a tutorial and the student would be prompted to make the appropriate response, typically by choosing a switch setting with a light pen. Feedback would then occur with the switch moving to the correct response graphically or further instruction being given in the case of an incorrect response. A logical test of the tutorial instruction would be to have the student go through the same procedure without the prompts. The tutorial is the prompted version while the test is the unprompted version. Since the prompt is the only difference between the two versions, the author needs to provide only one script, with the programing controlling which version is being presented: prompted for tutorial, unprompted for testing. The need for only one script reduced authoring time as well as programing time.

This design was achieved by organizing the CAI screen presentations into the following items, which can be seen in Figure 2.

1. Direction or Cue - The Direction is a general instruction to the trainee regarding a particular procedure to be performed, e.g., "Set up the TID pushtile panel for preflight. The pushtiles can be selected and deselected as desired." When more than one response is required on a page, the student will be told what to do to indicate he is finished: "When finished select □." Sometimes a change in a graphic will serve as a cue to perform the next step in the procedure and the only direction will be, "Continue with the procedure."

2. Prompt - This is a specific instruction detailing the exact steps that would accomplish the Direction, presented one step at a time where appropriate. For example, "Select all pushtiles except RID DSBL, LAUNCH ZONE and VEL VECTOR." Occasionally the prompt in the form of an arrow will appear in the graphic.

3. Feedback and Observation - The pushtile selected will light up to indicate it is on and when deselected will be unlit. On other panels, switches will go from

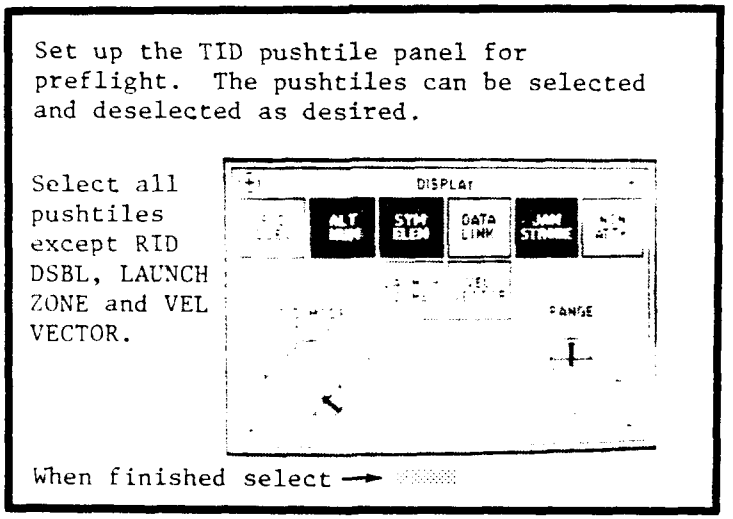


Figure 2. CAI screen of a panel in the rear cockpit.

the off to the on position when selected. For incorrect responses the student is told to "Try again" and is given further instruction.

4. Graphic - Usually the front or rear cockpit (Figure 1) or one of the panels in the cockpit (Figure 2) is displayed on the right two-thirds of the screen. Switch positions change as selected and dial readings change when manipulated.

For the procedural part of the training all four of the above items are seen by the student. The testing aspect is achieved by withholding the prompt until the student makes an error, in which case, the prompt is displayed as feedback.

Each lesson is composed of several segment or scenarios. A multiple-choice pretest based on required NATOPS readings has to be completed before attempting any of the scenarios. This pretest is given to motivate the student to read the NATOPS Manual, which is often ignored in favor of other training media. To successfully complete a scenario, the student is required to complete the unprompted version without making more than the specified number of errors. Following completion of the scenarios a multiple-choice posttest is given to test concepts learned in the lesson. However, the unprompted scenario is considered the best test of the procedure being tested.

While it is recommended that the prompted version be taken before the unprompted, this is not required. Beginning students are encouraged to complete both versions, but more advanced students may choose to see only the unprompted version, if they are familiar with the material. This provides for economy of instructional time.

SCREEN DESIGN

Each of the items presented on the screen was placed in the same location on each page so the learner would come to expect the information in a specific location. This allows the learner to concentrate on the material being taught rather than be distracted by the mechanics of the training vehicle. The screen design is shown in Figure 3.

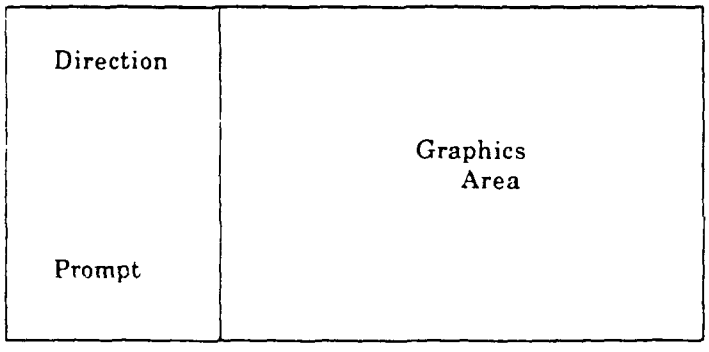


Figure 3. Screen design for CAI scenarios.

For incorrect responses, the words, "Try again," appear above the prompt. If there is no graphic change as a result of a correct response, the word, "Correct," appears under the prompt before moving on to the next page.

PROGRAM DEVELOPMENT

Following the Instructional Systems Development model, Mathetics training personnel updated the original program objectives and performed a Training Requirements Analysis and Media Selection. Upon approval of these reports, lesson specifications were completed for 35 lessons. These lessons are mainly procedural and included the following:

1. IFF (Identification Friend or Foe)
2. INS (Inertial Navigation System)
3. Aircraft Self-tests
4. Data Link
5. Navigation Controls and Displays

Other lessons considered appropriate for CAI involved graphic depictions of the internal workings of aircraft systems such as the following:

1. Hydraulic System
2. Electrical System
3. Fuel System
4. Power Plants

Besides showing how these systems worked, malfunctions and procedures for the corrections were included. This second group of lessons did not always lend themselves to the prompted/unprompted model and were sometimes shown as demonstrations only.

The lessons were developed by a team of instructional psychologists, aviation subject matter experts (SMEs), graphic artists and a programmer. Lesson specifications and storyboards were written by the instructional psychologist and SME working together as a team, the former contributing expertise in CAI and learning principles and the latter providing information on the aircraft systems. Upon completion of the storyboards, the graphic artist designed and executed the computer graphics for the lesson from the information supplied on the storyboard. Then the lessons were ready to be programmed using the ADAPT authoring language.

At first the instructional psychologist programmed the lessons on the computer. However, as it became clear that certain procedures such as rotating thumbwheels and changing several switches in any order were necessary, it was decided to add a professional programmer. The programmer along with one instructional psychologist became the programing team.

PRELIMINARY RESULTS AND CONCLUSION

At the present time approximately half of the lessons are on the computer and students have viewed about half of these. All of the students have been positive in their remarks. They have particularly enjoyed the interactive nature of the lessons. When they select a correct switch setting, the switch actually moves before their eyes. While this is not the airplane or the simulator, it is a fairly good facsimile which has promise of achieving the program goals.

By designing the program so that one script could be used for both the procedural tutorial and testing, time required to author and program the CAI lessons was reduced. Instructional time for some students will also be reduced by allowing more advanced students to bypass the prompted version. For the same amount of development time, two groups of students and two versions were produced for the price of one.

REFERENCE

Orlansky, J. & String, J. Cost-effectiveness of Computer-Based Instruction in Military Training. (IDA Paper P-1375) Washington, D. C. Institute for Defense Analysis, April, 1979.

ACKNOWLEDGEMENTS

Bruce Cordes and Ellen Le Vita along with the author were responsible for the program design. I wish to thank Bruce Cordes for his comments on this paper.

Development of a New System of Measurement

Ronna F. Dillon

Southern Illinois University at Carbondale

and

Richard K. Reznick

School of Medicine
University of Houston

The importance of sound prediction of both the success of medical school training and the quality of job performance is widely acknowledged. Traditional psychometric procedures center on a narow range of predictors (i.e., aptitudes), and only the products of performance are measured. The shortcomings of such product (i.e., test-score) information are receiving increasing attention (e.g., Dillon, 1986; Hunt, 1983). These traditional psychometric procedures have not accounted for adequate amounts of variance in criterion-task performance (Dillon, 1985; Sternberg).

With specific reference to prediction of success in the medical sphere, the Medical College Admissions Test (MCAT) is known to be of less-than satisfactory predictive power in accounting for medical school performance (Hobfol, 1981) and for predicting the outcomes of training during advanced work, such as residency (Kegel-Flom, 1975; Veloski, Herman, & Gonella, 1979). The National Board of Medical Examiners (NBME) assessment clearly is less than adequate in its predictive power and prescriptive utility with respect to residency programs (Gardner, 1973). Moreover, neither undergraduate grades (Hobfol, Anson, & Antonovsky, 1982) nor medical school grades (Wingard & Williamson, 1973) provide a potent system for predicting success in residency programs. Thus, the shortcomings of existing tests and testing procedures, for the entire medical education sector, underscore the importance of developing and operationalizing new measurement technologies for use in medical education.

What is needed is a robust conceptualization of aptitude that specifices the domain of aptitudes to be used as predictors, provides prescriptions for the use of new information-processing measurement technologies, and provides an expanded system of criterion measurement. Dilon (1986a) has provided a framework for considering new aptitude dimensions and new measurement techniques. New aptitudes include the range of procedural and declarative knowledge requisite to successful performance on the measure or measures of intelligent behavior comprising the criterion. Dimensions include knowledge acquisition, planning, and performance componential skills, coping with novelty and developing expert (e.g., automatized) performance, and enviornmental adaptation/contextual skills (Sternberg, 1985). New measurement techniques include recording ongoing information processing through measuring eye movement patterns (see Dillon, 1981, 1985, 1986). In

addition, testing may involve test administration manipulations, such as the use of dynamic testing procedures.

With respect to the nature of the criterion, there is no universal agreement as to what constitutes the ideal criterion measure for medical school or residency training, as well as for training in other academic/technical or occupational areas. In the medical school domain, the move toward a passs/fail system has severly limited the potential potency of predictor variables, and hence the need exists for a more elaborate and systematized criterion. Several options are available, and existing strategies include (a) class rankings, (b) overall performance grades, and (c) licensure examination scores.

In previous work, Dillon has found that criterion-related validity can be increased markedly when eye movement measures of information processing, recorded during solution of inductive reasoning items, are used in place of the examinee's score on the same test from which the information-processing measures are derived for predicting overall academic achievement for college undergraduates and graduate students (Dillon, 1986; Dillon, Radtke, Wood, & Koepping, 1984). In the present work, we advocate a system of measurement that expands both the predictor space and the system of criterion measurement. With respect to the predictor space, this study constitutes the first reported attempt to couple information-processing measures with declarative knowledge. Regarding the use of a robust criterion measure, the study is the first work of its kind to report the use of multiple raters evaluating examinees along multiple dimensions.

Method

Subjects

The sample was comprised of 15 third-year medical students. All students were enrolled in a surgical clerkship at the time of testing. The sample included 8 females and 17 males, between 23-34 years old.

Instruments

A 20-item test of analogical reasoning was given to all subjects in an untimed format. Analogies were solved by generating a transformational rule and applying the rule to medical knowledge. MCAT data also were compiled for each student. Criterion data were comprised of ratings given on a 10-point scale by 21 faculty members based on three tests of cognitive knowledge, two performance-based examinations and faculty reports.

Procedure

A Whittaker 1992 eye tracking system was used to track eye fixations. The system operates at 60 Hz, computing the subjects point-of-regard from horizontal and vertical locations in the stimuli. The subject viewed stimulus items, displayed on 35mm slidies, at a viewing distance of 57cm. A fixation was defined as the absence of change in eye position for > 125msec.

Results

Data were subjected to stepwise regression analysis (SAS; Helwig & Council, 1985). Analysis of all variables selected one MCAT score (i.e., science problems) and two information-processing measures as contributing signficantly to success on the criterion. The two information-processing measures were (a) the percentage of the total number of scans that were large, nonredundant scans (positively weighted), and (b) the total number of times the subject broke main array processing to scan the response set. The R^2 for the science problem subtest of the MCAT was .40. This value was augmented to .77 when the two information-processing variables were included. The increment in R^2 was signficant, $F(2,11) = 9.25$, $p < .01$.

Discussion

The absence of a sound system of measurement has posed difficulty for the selection process in medical, military, and other educational and occupational environments. Two reasons for this limited criterion-related validity are use of a predictive system that is limited conceptually and methodologically and use of a system of criterion measurement that is extremely narrow. An alternative program of measurement is proposed herein, and pilot data are provided. The program includes use of a series of new aptitude dimensions, new methods for measuring those aptitudes, and a robust system of criterion measurement. Measures of information processing during solution of complex reasoning items, tapping both procedural and declarative knowledge, are used to predict school success, defined by cognitive and affective dimensions of intelligent performance.

The information-processing approach described herein has been used to predict successful performance in college (Dillon, 1985), medical school, and military environments (Dillon & Wisher, 1981). Therefore, we feel its applicability is quite broad.

References

Dillon, R. F. (1981). Individual differences in eye fixations as descriptors of cognitive processes for figural analogies. (Tech. Rep. No. 2). Carbondale: Southern Illinois University.

Dillon, R. F. (1985). Predicting academic achievement with models based on eye movement data. Journal of Psychoeducational Assessment, 3, 157-165.

Dillon, R. F. (1986). Information processing and testing. Educational Psychologist, 1986, 20(3).

Dillon, R. F., & Wisher, R. A. (1981). The use of scanning indices to predict performance on technical school qualifying tests. Applied Psychological Measurment, 5(1), 43-49.

Hobfol, S. E., Anson, O., & Antonovsky, A. (1982). Personality factors as predictors of medical school student performance. Medical Education, 16, 251-258.

Hunt, E. B. (1983). On the nature of intelligence. Science, 218(4581), 141-146.

Kegel-Flom, P. (1975). Predicting supervisor, peer, and self ratings of intern performance. Journal of Medical Education, 50, 812-815.

SAS Institute Inc. (1985). SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: SAS Institute Inc.

Sternberg, R. J. (1984). What should intelligence tests test? Implications of a triarchic theory of intelligence for intelligence testing. Educational Researcher, 13(1), 5-15.

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. Cambridge: Cambridge University Press.

Veloski, J., Herman, M. W., Gonella, J. S., Zeleznik, C., & Kellow, W. F. (1979). Relationship between performance in medical school and first postgraduate year. Journal of Medical Education, 54, 909-916.

Wingard, J. R., & Williamson, J. W. (1973). Grades as predictors of physicians' career performance: An evaluative literature review. Journal of Medical Education, 48, 311-320.

Effectiveness of the Linking Format in a Technical Training Pamphlet

Karen Jones and Cheryl Bothwell
U. S. Coast Guard Institute

The linking format makes the structure of a publication's text highly visible to the reader through the use of side headings. In the application evaluated here, the side headings "link" each training objective to the related text to assist the student in meeting the objectives. Side headings also identify key ideas within the text to help the reader organize the information during reading and access information during review or scanning. The linking format compared favorably with the dual-column format when it was evaluated using written questionnaires from students and the students' performance on a multiple-choice knowledge test.

The linking format was developed during an editorial revision of a solid state pamphlet in an electronics correspondence training course. The course developers initiated the revision because student performance and comments indicated that the solid state material was more difficult than the other material in the course.

The course developers reviewed the pamphlet to determine if there was a way to make the material less difficult and, therefore, more useful as a training publication. They determined that the material covered in the pamphlet and the technical writing were appropriate for the target audience. However, the material covered -- theory of solid state components and troubleshooting techniques -- was more intimidating (difficult) than the rest of the course (i.e., operation of an oscilloscope, safety, and administrative paperwork). The course developers were not the first to identify writing about electronic theory as a problem. For example, Sawyer (1979) stated that it was more difficult to write in an easy to understand way about electronics engineering than other types of engineering.

The course developers wanted a way to make the presentation of the material less intimidating and hypothesized that visual organization was a key factor in the students' use and understanding of written material. They developed the linking format to make the presentation less intimidating. The format provides visual organization to help the student understand the material. It uses side headings (Waller, 1982) to emphasize the material's structure. The side headings "link" each training objective to the text for that objective and identify the key ideas within the text. This, in effect, provides the reader with a running outline of the material.

In revising the solid state pamphlet, the developers totally revised the format but left the text practically unchanged. To accommodate the side headings, the page layout was changed from the dual column format (refer to Figure 1) to a single column format with 5-inch lines and 2.5-inch margins (refer to Figure 2). This change also provided space for the student to make calculations and write study notes. In arranging the information, the course developers organized the material on a unit basis. For example, a topic was started at the top of a page and that idea was continued through the succeeding pages as necessary. They did not start a subject at the bottom of one page and finish it on the next. The presentation is slightly different from information mapping (Horn, 1982) which treats each page as a totally

As mentioned previously, there are two general types of gate structures. First, there is the junction FET (JFET) which has the gate "junctioned" into the channel similar to a PN junction of a diode or bipolar transistor. This type of FET has the advantage of simplicity, ease of testing with an ohmmeter, and little chance of static-charge damage. (To avoid such damage, you must short the leads of some FET's together until the FET is installed in the circuit.) In the JFET, the gate is reverse-biased by the circuit so that only a very small amount of current flows between the gate and channel.

The other general type of gate structure has
r of insulation b n the gate annel.
is kn SFET al-

HOW FET'S
ARE USED

The junction field-effect transistor (JFET) operates with a high impedance input only as long as the gate-to-channel diode is reverse-biased. Circuits using JFET's use either fixed or self bias (figure 73) to make sure the signal input does not forward bias the gate. For an N-channel JFET, the bias from gate to source is negative, and on a P-channel JFET, the bias must be positive, as shown in figures 73A and B.

FIGURE 1

OTHER SEMICONDUCTOR DEVICES

HOW FET'S ARE USED

Uses

The JFET operates with a high impedance input only as long as the gate-to-channel junction is reverse-biased. Circuits using JFET's use either fixed or self bias (figure S-2) to make sure the signal input does not forward bias the gate. For an N-channel JFET, the bias from gate to source is negative, and on a P-channel JFET, the bias must be positive (figure S-2A and B).

objective 1...

JFET

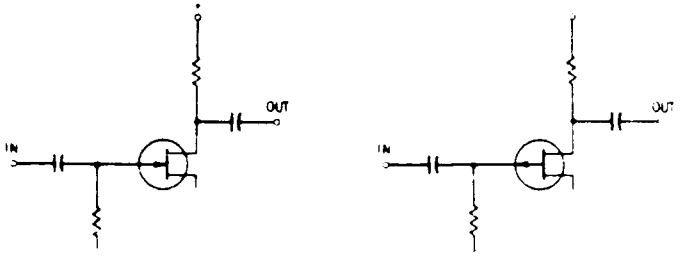


FIGURE 2

separate entity. Other design techniques were used in the page and pamphlet layout (e.g., running headings were added, the information on each page was balanced, and assignments were separated by colored dividers). As a comparison of Figures 1 and 2 shows, these changes produced a less intimidating presentation of the material.

Method

The revised pamphlet was developed for immediate incorporation in a correspondence course which the students were required to complete for career advancement. Therefore, the linking format was evaluated within the existing course administration procedures rather than by a controlled experiment.

Student Questionnaires

The format's effectiveness was immediately evaluated using student questionnaires so that the course developers could decide whether or not to use the format in developing/revising other pamphlets. Two questionnaires were developed -- one for the revised pamphlet only; the other for both the original and revised solid state pamphlets. Each had questions about the students, their study methods for the pamphlet, their opinion of the way the information was presented in the pamphlet, and their opinion of the relevance and usefulness of the pamphlet's content. The questions were multiple-choice, Likert-type, free response, and yes/no (with free response follow-ups).

For a few weeks after release of the revised pamphlet, all students who enrolled in the electronics course automatically received the revised pamphlet and a questionnaire covering it. The students who were already enrolled in the course were notified that a revised pamphlet was available upon request. The questionnaire on the original and revised pamphlets was included in the pamphlets distributed to these students.

Student performance

To determine if the results from the questionnaire were supported by a change in student performance, the format was evaluated two and one-half years later using student test scores. These test scores were from a proctored multiple-choice test which the student must pass to complete the course. The linking format and the dual column format were compared using percent of questions answered correctly on the solid state section of this test.

Results and Conclusions

The survey results indicated that the linking format was effective from the student's point of view -- the presentation was less intimidating and the material was easier to understand. However, as has been found with other studies using test scores, e.g., Duffy and Kabance's (1982) research with readable writing techniques, the change in format was not accompanied by a significant improvement in test performance.

Student questionnaire

Twenty-one students from the target audience responded to the questionnaire. All 21 students were high school graduates stationed at operational units.

Many of the students had completed some college or vocational-technical training; 19 had completed the Coast Guard's basic technical training in electronics; and 2 had completed only recruit training. In responding to the survey, 8 students completed the questionnaire on both the original and the revised pamphlet, 2 on the original pamphlet only, and 11 on the revised pamphlet only. To compare the students' opinions of the two formats, the responses were divided into two groups -- responses on the original pamphlet and responses on the revised pamphlet. The responses to the Likert-type questions were dichotomized into the equivalent of yes/no responses and combined with the responses for the yes/no questions. The responses were analyzed using Pearson chi-square tests for association (Hays, 1973).

Although the sample sizes were small (10 respondents for the original pamphlet; 19 for the revised), the results were quite consistent. As shown in Table 1, most students (90% and 95%) stated that the pamphlets helped them learn troubleshooting techniques. They preferred the revised pamphlet for future reference use as indicated by an increase (70% to 100%) in the number of students planning to keep the revised pamphlet for a reference. However, this increase was not statistically significant.

Table 1
Number and Percent of Respondents Answering "Yes" to Questions

Question about Pamphlet	Original Pamphlet		Revised Pamphlet	
	Number	Percent	Number	Percent
Help learn troubleshooting?	9	90%	18	95%
Keep pamphlet as reference?	7	70%	19	100%
Appropriate reading level?*	3	30%	15	79%
Easy to study?***	0	0%	16	84%
Logical stopping places?**	5	50%	19	100%
Well organized?***	0	0%	18	95%
Hold your interest?*	3	30%	16	84%
Assignment length appropriate?**	3	30%	17	90%
Illustrations interfere with reading?***	9	90%	2	16%
Parts confusing?	5	50%	6	32%
Headings help you study?	--	--	17	89%
Expand use of format?	--	--	18	95%

Note: "*" indicates difference between pamphlets was significant at the .05 level; "**", at the .01 level; and "***", at the .001 level.

The comparison of the original and revised pamphlets indicated that the linking format improved the usefulness of the pamphlet for training and made the presentation less intimidating. As shown in Table 1, a larger percentage of students answered that the revised pamphlet (1) had an appropriate reading level, (2) was easy to study, (3) had logical stopping places, (4) was well organized, (5) held their interest, and (6) had assignments of appropriate length; a smaller percentage responded that the illustrations interfered with reading. All of these comparisons were statistically significant. However, the finding that a larger percentage rated the reading level in the revised pamphlet as appropriate is particularly noteworthy, since the actual text was practically unchanged. Although the students' evaluations were extremely favorable, the survey responses indicated that the revised format did not solve everything. In the comparison of the original and revised pamphlet, fewer respondents rated parts of the revised pamphlet as confusing but this decrease was not statistically significant. This finding is understandable in view of the highly technical nature of the material.

On questions which applied only to the revised pamphlet, the students indicated that they liked the linking format. Most (89%) stated that the side headings helped them study and 95% stated that they wanted more pamphlets developed using the new presentation method. The students' comments supported the hypothesis that making the pamphlet's structure visible makes the material appear better organized and easier to understand -- even with only the minimal changes to the text which were made in this revision. For example, one student commented that in the revised pamphlet the "organization flowed together from topic to topic and [the information] was presented in an easier to understand manner than the old course".

Student performance

Test scores were available for 251 students who had used only the original pamphlet and 621 students who had used only the revised pamphlet. To equalize group sample sizes, 251 students were randomly sampled from the 621 students who had used only the revised pamphlet. The large difference in sample size between the questionnaire results and the test results occurred because the students were not required to complete the questionnaire.

Although the students preferred the linking format, the test scores indicated only a trend toward the format's improving test performance. The percent of questions answered correctly on the solid state section of the test was slightly higher for the students who studied the revised pamphlet ($\bar{x}$ = 54.15%, σ = 19.12) than for those who studied the original pamphlet ($\bar{x}$ = 52.25%, σ = 15.77). However, when the test scores for the two groups were compared using an independent sample t-test (Hays, 1973), the improvement was not statistically significant ($t = 1.21$, $df = 500$, $p > .05$).

The finding that the students' improvement in performance on the solid state section of the test was not statistically significant should not be taken as a negative indicator for the new format. One factor which should be considered in evaluating this finding is the reliability of this performance measure. The solid state section has only 12 questions which, even with a large sample size, does not result in high reliability. In addition, in related studies other researchers have found no significant improvement in test performance. For example, in five experiments Duffy and Kabance (1982) found little effect

on reading comprehension when they revised written material using readable writing techniques. Our findings were in agreement with their research and the research they cited -- large improvements in test scores do not occur when the ability level of the students is high and the training objectives are presented to the student. Although the students' test scores did not improve significantly with the revised pamphlet, there has been a decrease in the time the students are taking to complete the course. However, at this time there are insufficient data to draw a firm conclusion from this decrease.

In summary, the format change caused the expected result -- the students were less intimidated by the material. As a result the material was easier to read and use. This outcome -- enabling the students to read and use the material more easily -- indicates that the format produced successful results. Based upon this finding, the course developers decided that it was worthwhile to use the new format in other courses.

Recommendations

This type of format should be effective for achieving one or more of the following:

* make the presentation less intimidating
* alert the reader to key points in the text
* help the reader search for information
* link related information

It could be used in publications other than technical training and reference manuals. For example, it would be useful in a reference book detailing the application of rules and regulations, a job aid linking the situation and the procedure or in a gardening "how-to" book linking the problem and solution.

References

Duffy, T. M. and Kabance, P. (1982). Testing a reading writing approach to text revision. Journal of Educational Psychology, 8, 733-748.

Hays, W. L. (1973). Statistics for the social sciences. New York: Holt, Rinehart and Winston, Inc.

Horn, R. E. (1982). Structured writing and test design. In D. H. Jonassen (Ed.), The technology of text: Principles for structuring, designing and displaying text (pp. 341-367), Englewood Cliffs, NJ: Educational Technology Publications.

Sawyer, R. (1979). It is easy to communicate electronically; it is hard to communicate electronics. Journal of Technical Writing and Communication, 8, 121-131.

Waller, R. (1982). Text as design: Using typography to improve access and understanding. In D. H. Jonassen (Ed.), The technology of text: Principles for structuring, designing and displaying text (pp. 137-166), Englewood Cliffs, NJ: Educational Technology Publications.

EXAMINATION OF ALTERNATE SCALES FOR JOB INCUMBENT OCCUPATIONAL SURVEYS

LAWRENCE A. GOLDMAN, Ph.D. — USA Soldier Support Center - NCR
DARRELL A. WORSTINE — USA Soldier Support Center - NCR

Background. Since 1973, the Army Occupational Survey Program (AOSP) has collected, on a routine basis, job incumbent information from enlisted soldiers using a 7 point Relative Time Spent (RTS) scale. The soldiers who participate in AOSP survey are first instructed to review all tasks in the questionnaire. Then, they are to rate each of those tasks which they perform on the RTS scale, according to the amount of time devoted to that task relative to all other tasks done in their current job. For those job incumbents whose work varies widely and/or who perform a substantially large number of tasks, this requirement represents a rather complex undertaking subject to a considerable degree of error.

While each of the values used for the RTS scale is defined, ranging from "1" (Very Much Below Average) to "7" (Very Much Above Average), the individual values alone are not relevant in the interpretation of the information collected. Theoretically, a common job description could be obtained from two different individuals performing the same tasks, although one rated the tasks consistently lower than what one might expect while the other soldier rated them consistently higher than what one might expect. In general, the data collected using the RTS scale is most useful for: 1) assisting in job "typing" (involving the grouping of individuals performing similar work); and 2) examining work performed by sub-samples of soldiers based on the time they spent on each duty (each duty representing a category of related tasks). Very little, if any, useful information is obtained from the estimates of percent time spent by individuals on individual tasks. Partially for this reason several U.S. Army schools asked that a frequency scale as opposed to the RTS scale be used in surveys of their enlisted Military Occupational Specialties (MOS). In using the latter type of scale, information could be obtained on each task, particularly with respect to how often the task was performed. Also, it would be much easier for job incumbents to use since each task would be evaluated independently, regardless of the number of tasks performed. Nonetheless, if a frequency scale were to be used for occupational surveys, it should have the same capabilities as the RTS scale, particularly with respect to assisting in job "typing" and being analogous to the RTS scale with respect to providing estimates of percent time at the duty level.

Therefore, it was desired to determine the extent to which a frequency scale was comparable to the RTS scale in the survey of an enlisted MOS. In other words, could a frequency scale be used interchangeably with the RTS scale? In this study, two types of frequency scales were considered, a relative frequency scale and an absolute frequency scale, to ascertain if both or just one of these types could be deemed comparable to the RTS scale.

Methodology. The results reported in this study were based on an Army-wide sample survey of job incumbents in MOS 93P (Flight Operations Coordinator),

with authorized skill levels (SL) of SL1 through SL5 corresponding to paygrades E-3 through E-9. In July 1983, questionnaires were distributed using a Relative Frequency (RF) scale and an Absolute Frequency (AF) scale to each location where soldiers in this MOS were assigned. Roughly the same number of questionnaire booklets using the RF scale as those using the AF scale were distributed to each location. In the fall of 1984, approximately one-half of the total number of MOS 93P questionnaire booklets distributed in 1983 were sent to the same duty locations using the RTS scale. Analysis of the data was based on the following number of soldiers using these scales: (a) RF scale-176 soldiers (of which 116 were in SL1 corresponding to paygrades E-3 and E-4); (b) AF scale - 183 soldiers (of which 112 were in SL1; and (c) RTS scale - 230 soldiers (of which 147 were in SL1). The sample sizes of SL2 (E-5 soldiers) ranged from 15 to 30; those for SL3 (E-6) only ranged from 12 to 14; those for SL4 (E-7) ranged from 20 to 30; while those for SL5 (E-8 and E-9) ranged from 6 to 12. Twenty-four soldiers who had filled out a 93P questionnaire with the RTS scale later filled out one using the RF scale. Thirty other individuals who had filled out answer booklets for questionnaires with the RTS scale subsequently responded to this survey using a questionnaire with the AF scale. In terms of the alternate scales used in this study, the values for the 7-point RF scale ranged from "1" (Very Seldom) to "7" (Very Frequently). Similarly, the values for the 7-point AF scale ranged from "1" (Less Often Than Once A Month) to "7" (More Often Than Once A Day).

The data collected from MOS 93P soldiers using these three scales were first integrated into a common data file. Then, using the Comprehensive Occupational Data Analysis Programs (CODAP), combined with the Statistical Package for the Social Sciences (SPSS), the following analyses were conducted to examine the extent of inter-changeability between these three scales:

(1) Identification of the major types of jobs performed by MOS 93P soldiers. The process used within CODAP is based on "clustering" or "hierarchical" grouping whereby individuals are grouped according to the similarity of work performed based on commonality of time spent values (or estimates of time spent values) based on task performance ratings. After identification of all job types, the percentage of soldiers using each scale was determined for each job type. It was hypothesized that if these three scales were inter-changeable, then the percentages noted for each job type should be closely comparable to the overall percentages of 93P soldiers using each scale.

(2) Examination of the average percent time spent (or estimates of average percent time spent) with respect to the RF and AF scales by SL for each of 16 duties relating to 493 tasks in the task section of the 93P questionnaire. These duties included MOS-specific areas (e.g., Flight Planning and Dispatching, Flight Records, Aviation Safety, etc.) as well as duty areas performed by Army soldiers in general (e.g., General Military Training, Vehicle Operation and Operator Maintenance, Personnel Management and Supervision, etc.).

(3) Examination of the inter-correlations of average percent time values (average percent estimate time values pertaining to the RF and AF scales) in

terms of tasks. This review was done for each skill level. The closer that the PEARSON correlation coefficients approached 1.00, the greater the degree of inter-changeability between the three scales.

Findings

A. Job Structure Analysis. The job typing done for MOS 93P yielded nine (9) distinct jobs, including 558 (93 percent) of the 589 soldiers in this study. The primary work performed by soldiers in seven of these nine job types could be thought of as MOS-specific while the work performed by two other job types (Platoon Sergeants and General Military Personnel) is often encountered in many other Army MOS. Table 1 displays, for each job type and all soldiers whose work was identified, the percentage of 93P soldiers responding to each of these three answer scales.

TABLE 1 - CROSS-TABULATION REPORTS FOR MOS 93P JOB TYPES BY ANSWER SCALE USED

JOB TYPE \ SCALE USED	RTS (%)	RF (%)	AF (%)	NUMBER OF SOLDIERS IN JOB TYPE
MOS-SPECIFIC				
Flight Operations Specialist	37	31	32	258
Flight Operations Clerk	57	14	29	14
Flight Records Clerk	39	39	22	41
Senior Flight Operations Specialist	35	31	34	88
Assistant Flight Operations Sergeant	30	25	45	20
Flight Operations Sergeant	35	20	45	42
Operations Sergeant	40	30	30	10
NON-MOS SPECIFIC				
Platoon Sergeant	38	43	19	42
General Military	63	14	23	43
TOTALS	39	30	31	558

Examination of Table 1 indicates that for those job types where the number of soldiers is substantially large, the percentage of soldiers responding to each of the three answer scales is closely comparable to the overall percentages of soldiers responding to these scales. For example, the percentages of Flight Operations Specialists (the predominant job type) responding to the RTS, RF, and AF scales, in comparison to the overall percentages for each of these scales were 37 (vs 39) percent; 31 (vs 30) percent); and 32 (vs 31) percent, respectively. With respect to the MOS-specific job types, percentage differences greater than ten percent were noted just for Flight Operations Clerk (where the sample size was only 14) and for an appreciably higher

percentage of Flight Operations Sergeants and Assistant Flight Operations Sergeants using the AF scale. Concerning the non-MOS specific job types, the only major difference from the overall averages related to a substantially higher percentage of General Military personnel responding to the RTS scale (63 percent) than all 93P soldiers using this scale (39 percent).

B. Group Summary by SL. Review of the average percent time values for each of the 16 duties included within the MOS 93P questionnaire (or estimates of average percent time in the case of the RF and AF answer scales) showed a remarkable degree of consistency between the RTS, RF and AF scales. This was especially true for SL1 where the sample size for soldiers responding to each of these scales exceeded 100. Table 2 identifies those duties where differences between any two of these three answer scales, considering SL1 and SL2 separately, exceeded five (5) percent. Table 3 is analogous to Table 2, highlighting differences for SL3, SL4, and SL5.

TABLE 2 - MOS 93P DUTIES IDENTIFYING DIFFERENCES OF AT LEAST FIVE PERCENT BASED ON AVERAGE PERCENT (ESTIMATED) TIME VALUES COMPARING ALTERNATE SCALE RESPONDENTS - SL1 AND SL2

DUTY TITLE	SL1 RTS	SL1 RF	SL1 AF	SL2 RTS	SL2 RF	SL2 AF
FLIGHT PLANNING & DISPATCHING				16.2	27.0	18.1
FLIGHT RECORDS				6.2	5.8	11.5
AVIATION SAFETY						
TACTICAL OPERATIONS				15.6	5.4	9.5
PERSONNEL MGT & SUPERVISION						
GENERAL MILITARY TRAINING	24.7	20.1	18.0	24.3	17.1	18.4

TABLE 3 - MOS 93P DUTIES IDENTIFYING DIFFERENCES OF AT LEAST FIVE PERCENT BASED ON AVERAGE PERCENT (ESTIMATED) TIME VALUES COMPARING ALTERNATE SCALE RESPONDENTS - SL3, SL4, SL5

DUTY TITLE	SL3 RTS	SL3 RF	SL3 AF	SL4 RTS	SL4 RF	SL4 AF	SL5 RTS	SL5 RF	SL5 AF
FLIGHT PLANNING & DISPATCHING							1.5	8.7	12.1
FLIGHT RECORDS									
AVIATION SAFETY	8.0	2.1	9.9				18.4	10.0	11.5
TACTICAL OPERATIONS									
PERSONNEL MGT & SUPERVISION							18.1	13.1	11.4
GENERAL MILITARY TRAINING	14.0	18.7	11.8				13.3	7.5	13.5

As indicated in Tables 2 and 3, there were relatively few prominent differences between respondents using alternate scales, regardless of SL. Moreover, there was no single instance where the average (estimated) percent time values differed more than five percent among all three sub-groups of scale respondents for any SL. The only duty for which "prominent" differences appeared for more than two SLs was GENERAL MILITARY TRAINING which is non-MOS specific. From another point of view, out of 80 possible comparisons (based on 16 duties and five SL), there were only 11 instances where differences of at least five percent were noted - four of these associated with the non-MOS specific duty of GENERAL MILITARY TRAINING, another with the non-MOS specific duty of (unit level) PERSONNEL MANAGEMENT AND SUPERVISION.

c. Inter-Correlations. The Pearson correlation coefficients among the respondents to each of the three answer scales for SL1 through SL5 associated with this MOS are shown in Table 4 below.

TABLE 4 - INTER-CORRELATION MATRIX OF MOS 93P RESPONDENTS TO ALTERNATE ANSWER SCALES BY SL BASED ON AVERAGE PERCENT (ESTIMATED) TIME VALUES

		SL1			SL2			SL3			SL4			SL5		
		RTS	RF	AF	RTS	RF	AF	RTS	RF	AF	RTS	RF	AF	RTS	RF	AF
SL1	RTS	1.0	.95	.92												
	RF		1.0	.95												
	AF			1.0												
SL2	RTS				1.0	.70	.76									
	RF					1.0	.74									
	AF						1.0									
SL3	RTS							1.0	.68	.77						
	RF								1.0	.66						
	AF									1.0						
SL4	RTS										1.0	.71	.73			
	RF											1.0	.82			
	AF												1.0			
SL5	RTS													1.0	.68	.59
	RF														1.0	.64
	AF															1.0

As shown above in Table 4, the correlations for SL1 are all above .9, accounting for over 82 percent of the common variance. While the correlations noted for the other SL are much lower, falling within the .6 to .8 range, they are nonetheless statistically significant to a very high degree, accounting for about 40 to 60 percent of the common variance. In interpreting these results, it was believed that sample size played an important role; specifically, the SL1 sample sizes all exceeded 100 while the sample sizes for the other four SL only ranged from 6 to 30.

Conclusions and Implications for Future Studies. It was evident that there was a remarkable degree of consistency between these three answer scales for each type of analysis performed in this study. This was especially true whenever the sample sizes involved were fairly large, particularly those for SL1 soldiers which, for each of these scales, was in excess of 100. The impact of large sample sizes was also evident in review of the job structure analysis for this MOS. With respect to the latter, the percentages of soldiers responding to each of these scales in the two largest job types (Flight Operations Specialists consisting of 258 soldiers and Senior Flight Operations Specialists comprised of 88 MOS 93Ps) were closely comparable to the overall percentages of respondents to each answer scale. Regardless of the analysis done, there was no clear-cut evidence that either the RF or the AF scale was more comparable to the RTS scale. To validate the findings of this study of MOS 93P personnel, the AOSP, in Fiscal Year 1987, will survey soldiers in three other MOS. These three MOS (12B-Combat Engineer; 31K-Combat Signaler; and 94B-Food Service Specialist) were selected primarily because they represent widely different types of work and because each is comprised of substantially large numbers of soldiers. Since there are no essential differences noted between the RF and AF scales, and because the AF scale can yield more useful information at the task level, only the RTS and the AF scales will be used in these follow-on studies. If the results of this validation study confirm the findings noted for MOS 93P, then it is likely that future AOSP surveys will make more use of the AF scale.

In addition to providing more information for training school course developers, the soldier filling out the AOSP questionnaire should find it much easier to accomplish (especially if he/she performs a large number of tasks). In addition, the time required in questionnaire administration should be shortened with a concomitant increase in data accuracy and reliability.

Preliminary Holland Code Classification of Navy Entry-Level Occupations

John L. Holland, PhD
Johns Hopkins University

Herbert George Baker, PhD
Navy Personnel Research and Development Center

Abstract

A system by which to classify military occupations and individual occupational preferences according to a common principle has not been available heretofore. To fill that need, Navy entry-level occupations were classified according to the widely used Holland coding system. This preliminary classification has great potential for use by both the military recruiter and the school guidance counselor. Further research is indicated to verify its usefulness in improving selection, assignment, productivity, and job satisfaction in the Navy.

Introduction

Although there are significant occupational differences (e.g., applicant characteristics, legal commitment, constraints on freedom), factors affecting vocational choice and occupational placement within and outside the Armed Forces are similar (Clark, 1955). That is, personal abilities, interests, and preferences must be juxtaposed with institutional factors such as job openings, minimum standards, and employment incentives.

Vocational interests, values, and preferences have been investigated in light of their potential contributions to the selection, classification, and assignment of military job applicants, with the idea of increasing job satisfaction and individual productivity and reducing attrition. However, no instrumentation has been developed within the military research community that permits easy linkage of interests to occupations.

A number of occupational classification systems are used by the Armed Forces. Unfortunately, these classification systems group occupations according to logistical or administrative convenience. In addition, occupations and occupational preferences are not dealt with in common terms, making the task of relating individual preferences and occupational information highly problematical.

In military recruiting, where job applicants tend to be career naive and no professional guidance is available, there is a particular need for preenlistment vocational guidance that links applicant characteristics with military work (Baker, 1985). However, to accomplish person-job matching systematically, occupational preferences and occupations must be classified

according to the same scheme. The objective of this effort was to classify entry-level Navy occupations according to the Holland (1985) coding system, for later use in developing a computerized military vocational guidance system.

The Holland coding system is intended to be applied to both occupations and vocational interests. Based on the premise that vocational interests are expressions of personality, the Holland coding system (1985) classifies people and work environments according to six main types (see Fig. 1). (Refer to Holland, (1985) for complete descriptions of the six basic types.)

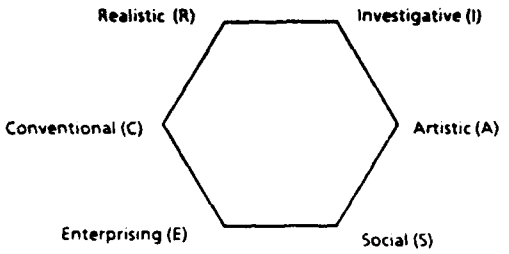


Figure 1. Holland hexagonal classification.

Because no person or occupation is characterized as a pure type, the three-letter code was developed. The first letter of the code is the principal personality or occupational descriptor, with the second and third letters providing supplementary information. A variety of types of people are found successfully working within any single occupation, but some types are more common than others (see Holland, 1979; Holland & Holland, 1977). In short, all occupations tolerate a range of types, but some types appear to cope more successfully with an occupation's demands than do others.

Occupational classification according to the hexagonal model provides a method for estimating the psychological distance between occupations and between preferences. Psychological distance refers to the dissimilarity of types and is represented by their relative positions within the hexagon (see Fig. 1), i.e., proximal occupations are more alike than are distant occupations.

Because the classification system is integral to Holland's theory of vocational choice, the act of classification makes it possible to use the theory to interpret or predict the behavior and activities of persons and the influence of occupations or environments assigned to a particular category. The Holland classification schema has undergone a number of revisions and tests of its usefulness from 1959 to 1982 (Holland, 1985) and is now the most widely used organizing principle in the field of vocational choices.

Method

Information on Navy occupations was collected from a variety of sources that contain an abundance of information on occupations and occupational groups (Dept. of Defense, 1985; Dept. of Labor, 1977; Dept. of Navy, 1986; Dept. of Navy, unpublished; Gottfredson, Holland, & Ogawa, 1982). Entry-level occupations were defined as the 96 ratings used by the Navy's automated assignmnent system for recruit classification.

Each entry-level occupation was classified with a three-letter code, which was determined by comparison with related occupations found in the Dictionary of Holland Occupational Codes (DHOC) (Gottfredson et al., 1982). The DHOC empirically assigns a Holland code to all of the civilian occupations and many of the military occupations that are listed in Dictionary of Occupational Titles (DOT) (Dept. of Labor, 1977). Because the DHOC is based on the ratings of experienced occupational observers, the classification is tied closely to occupational data. In addition, the DHOC illustrates that the codes based on the older interest inventory data closely approximate codes based on data from job analyses.

Although a few military occupational specialties were equivalent to DOT codes and therefore explicitly related to a Holland code via the DHOC, the assignment of most codes required expert judgment. This process was accomplished under contract to the Navy, and was guided by a review of related DOT occupations military occupational information (Dept. of Navy, 1986).

Results and Discussion

By this review procedure, 96 entry-level Navy occupations were assigned three-letter Holland codes (Holland & Baker, in press). Results show that recruits with realistic interests are most likely to find the military a compatible environment.

The application of Holland's classification to military occupations provided a stringent test of the classification scheme because the majority of military occupations fall in a single major type: Realistic. These are occupations that require technical, electronic, or mechanical skills and physical agility and strength.

Previous findings indicate that over 75% of job transitions fall within the same major category (Holland, Sorenson, Clark, Nafziger, & Blum, 1973). Nevertheless, there is some evidence to support the expectation that the subtypes in the classification can distinguish occupations within the Realistic category and between the Realistic and the other main categories. Specifically, among the Realistic type, two and three-letter codes were found to predict occupation 5 and 10 years following classification (Holland et al., 1973).

Based on this preliminary classification, the subtypes that would be most likely to find military occupations to match their interests are RE, RI, IR, RS, CS, and RC. The groupings in Table 1 resemble many occupational groups created by Clark (1961), Clark and Campbell (1965), and Norman (1960) for large samples of skilled tradesmen in Navy and civilian occupations. For example, seven of the nine Area scales (Mechanical, Health Service, Office Work, Electronics, Food Service, Carpentry, Outdoors) in the Minnesota Vocational Interest Inventory (Clark & Campbell, 1965) appear analogous to occupational groups in Table 1.

A major caveat is in order. The Holland coding of Navy occupations as accomplished herein is predicated on the supposed similarity of military and civilian occupations. That is, a welder is a welder, and so forth. To the degree that same-titled military and civilian occupations are dissimilar, any occupational classification system, other than one unique to the military, should be used with caution.

Nevertheless, the most feasible approach to the problem of developing military occupational exploration and vocational guidance systems is to consider the bulk of military and civilian jobs (qua job) as identical, while addressing the military-unique factors of the job or occupation separately (Baker, in press). In that light, the occupational classification system developed here should prove sound.

Conclusion

Assigning Holland codes to Navy entry-level occupations is the first step toward developing a prototype vocational guidance system for military recruiting. Future research will be needed to determine the validity of this preliminary classification scheme and its usefulness in improving selection, classification, and assignment of military job applicants and individual productivity and job satisfaction. Certainly, Navy job experts should rigorously assess the validity of these preliminary classifications.

Identifying specific military occupations or occupational groups that are congruent with a prospective recruit's vocational preferences is potentially beneficial both to the Armed Forces and to those persons who wish to investigate a military vocation. This classification system should facilitate and focus occupational exploration not only for potential recruits but also for those who may not have considered military enlistment. Recruiters and school guidance counselors should find this coding of Navy entry-level occupations very useful.

References

Baker, H. G. (1985). Antecareer crisis: Military recruiting and the youthful job applicant. Armed Forces and Society, 11, 565-580.

Table 1
Distributions of Navy Entry-Level Enlisted Occupations

Code	Navy
RIS	3
RIE	21
RIC	1
RSE	4
RSC	0
RSI	0
REI	11
RES	10
REC	1
RCI	0
RCS	0
RCE	3
	54 (56%)
ISR	0
ISA	0
IER	0
IRS	2
IRE	18
IRC	0
	20 (20%)
ASE	0
ASI	0
AES	0
AER	1
AEI	1
	2 (2%)

Code	Navy
SEC	1
SER	1
SEI	0
SEA	0
SCR	0
SCI	0
SCE	0
SRE	0
SAE	0
SAI	1
SIA	2
SIR	1
	6 (6%)
ECS	0
ERI	0
ERS	1
ESC	1
EAS	1
ESR	1
ESI	0
ESA	0
	4 (4%)
CRS	1
CRE	0
CSE	2
CSR	1
CES	2
CER	1
CEI	1
CIE	1
	9 (9%)

Baker, H. G. (in press). Designing a vocational guidance system for military recruiting: Problems and prospects. II. Theoretical and methodological considerations. San Diego: Navy Personnel Research and Development Center.

Clark, K. E. (1955). The use of interest measures with naval enlisted personnel. Minneapolis: Department of Psychology, University of Minnesota.

Clark, K. E. (1961). Vocational interests of nonprofessional men. Minneapolis, University of Minnesota Press.

Clark, K. E., & Campbell, D. P. (1965). Manual for Minnesota Vocational Interest Inventory. New York: Psychological Corporation.

Department of Defense. (1985). Military occupation and training data. Arlington, VA: Defense Manpower Data Center.

Department of Labor. (1977). Dictionary of occupational titles. Washington, DC: Government Printing Office.

Department of the Navy. (undated). Classifier rating fact sheets. Washington, DC: Chief of Naval Operations. OP-01.

Department of the Navy. (1986, January). Manual of Navy enlisted manpower and personnel classifications and occupational standards. II. Navy enlisted classifications. Washington DC: Author.

Gottfredson, G. D., Holland, J. L., & Ogawa, D. K. (1982). Dictionary of Holland occupational codes. Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L. (1979). Professional manual for the Self-Directed Search. Palo Alto, CA.: Consulting Psychologists Press.

Holland, J. L. (1985). Making vocational choices: A theory of vocational personalities and work environments. Englewood Cliffs, NJ: Prentice-Hall.

Holland, J. L. & Baker, H. G. (in press). Preliminary classification of Army and Navy entry-level occupations by the Holland coding system. San Diego: Navy Personnel Research and Development Center.

Holland, J. L., & Holland, J. E. (1977). Vocational indecision: More evidence and speculation. Journal of Counseling Psychology, 24, 404-414.

Holland, J. L., Sorensen, A. B., Clark, J. P., Nafziger, D. H., & Blum, Z. D. (1973). Applying an occupational classification to a representative sample of work histories. Journal of Applied Psychology, 58, 34-41.

Norman, W. T. (1960). A spatial analysis of an interest domain. Educational and Psychological Measurement, 20, 347-361.

THE JOB DIFFICULTY INDEX - REVALIDATING THE EQUATION FOR SUPERVISORY JOBS

Squadron Leader Kenneth C. Given
Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks AFB, Texas

Introduction

In the early 1970s, the Air Force Human Resources Laboratory (AFHRL) developed the Job Difficulty Index (JDI), a measure of the relative difficulty of jobs within an Air Force Specialty (AFS). Historically, job difficulty had been associated with the subjective and qualitative evaluation of a job. The major systems in use in industry at that time did not lend themselves to the military situation with its vast number of jobs and wide dispersal throughout the world. As a recurring need had been identified for a technique to derive a quantifiable index of the relative difficulty of Air Force apprentice, journeyman and technician jobs, Donald F. Mead at AFHRL addressed the problem of developing such an index.

Mead published three Technical Reports in 1970 detailing his research and also reported his findings to the MTA in 1971. Since that time little work has been done to validate his research although Wiley (1972) used Mead's index to help predict job difficulty. The purpose of this paper is to summarize an ongoing effort to revalidate Mead's equation and to determine whether or not the existing JDI is valid when applied to enlisted supervisory positions.

Background

The problem, as perceived by Mead (1970), was to develop a job evaluation approach which would satisfy four criteria. The system would be quantitative, it would be easy to administer, it would be objective, and it would provide maximum interface with existing military occupation analysis data processing programs (or CODAP).

Mead's approach was to have experienced supervisors rank 25 randomly selected enlisted Position Descriptions (PDs) for an AFS according to their relative difficulty, and then to apply Christal's (1967) judgement analysis model to capture the judgement policy of these supervisor rankers. The resulting prediction equation could then be applied to all enlisted positions in the AFS to determine their difficulty levels. In the three AFSs studied by Mead, the same three predictor variables combined in the multiple regression equation to capture the supervisors' evaluation policy. These predictor variables were the number of tasks appearing in the position description, the square of this variable, and the average difficulty per unit time spent (ADPUTS). This last variable, ADPUTS, is calculated using all the tasks performed in any PD. It is the sum of the cross products of the percent time spent and the mean task difficulty for each task divided by 100.

The correlations found by Mead between the predicted job difficulty values, based on these three predictor variables, and the criterion values for the three AFSs he studied are shown in Table 1. These correlations, measures of the predictive efficiency of the captured policy, indicate that there is very substantial predictive power.

Table 1
Correlations Between Predicted & Criterion Job Difficulty Values [a]

AFS	r	r^2
Vehicle Maintenance	.93	.86
Accounting and Finance	.95	.90
Medical Materiel	.95	.90

a Data from Mead,1970

Having found that three quite independent and diverse AFSs used the same three predictor variables to capture the supervisors' job difficulty judgements, Mead proceeded to develop a constant standard weight equation to predict the difficulty level of jobs across AFSs. For each AFS, standard score beta weights were developed for the three variables to maximize the correlation between the predicted values and the supervisors' evaluations. The similarity in the standard score weights suggested that the most suitable constant standard weight would be derived by computing the mean weights for each predictor. The results of Mead's work are shown at Table Two.

Table 2
Standard Score Weights for Selected Predictor Variables

AFS	Standard Score Weight			Criterion Standard Deviation
	Variable 1	Variable 2	Variable 3	
Vehicle Maintenance	1.29125838	.51612430	-.61529753	4.9992
Medical Materiel	1.12582776	.45263499	-.58673340	5.7705
Account & Finance	1.58510913	.39230372	-.95835786	5.4198
Mean (3 AFSs - Mead)	1.33406509	.45368767	-.72012963	5.3965
Mean (12 AFSs - Wiley)	1.42366	.38343	-.81392	5.4816

Note: Variable 1 is the number of tasks performed.
Variable 2 is ADPUTS.
Variable 3 is number of tasks performed squared.

This average-weight equation was applied to all cases in the three career ladders to produce new values of job difficulty. These newly computed values, using the single equation, were shown to still correlate highly with the supervisors judgement. Wiley (1972) developed new standard score weights for 12 AFSs and found little change in weights (see Table 2) or predictive efficiency.

Emerging Issues

The results of Mead's research were highly encouraging and the JDI was introduced into the reporting of occupational data by the USAF's Occupational Measurement Center (USAFOMC). Operationally, the JDI has been used to compare the difficulty level of work assigned to various individuals, to assist in establishing minimum aptitude requirements for enlisted positions, and to ensure that individuals are given increasingly difficult jobs and responsibilities as their careers progress. However, a number of problems have arisen in the use of the JDI and concerns have been expressed about the validity of the equation in general and the applicability of the equation to the enlisted supervisory jobs in particular. A brief summary of the concerns leading to the present study is given below.

The current equation reliably estimates the difficulty of apprentice and journeymen enlisted jobs but is ineffective when applied to supervisory jobs. The reason for this lack of sensitivity towards supervisory jobs may be found in the initial research design. In order to establish the criterion value, 250 individual position descriptions were randomly selected from an AFS. These 250 PDs were divided into subsets containing 25 PDs. However, as only some 3% of the Vehicle Maintenance personnel, for example, are either 9-skill level or Chief Enlisted Managers (CEMs) only 7 of the 250 PDs randomly selected would represent these higher skill level positions. Many of the subsets which were ranked would therefore, have no higher skill levels PDs, making the ranking judgement for such jobs impossible to capture.

The insensitivity of the JDI to the higher skill level jobs may also be explained by the emphasis given in the equation to the number of tasks performed in any position. Lower skill level positions tend to consist of many tasks while the higher skill level positions consist of relatively few tasks. The underlying assumption in occupational analysis is that there is, in some undefined sense, some degree of equality in each task listed in a task inventory. This, of course, is not the case. The task of removing a fuel pump is not equal in any sense to the task of counselling a subordinate. The JDI makes no allowance for the difference in the type and number of tasks which supervisors perform as opposed to the type and number of tasks performed by journeymen. This is largely because the current JDI captured a policy which involved using primarily non-supervisory tasks.

Part of the concern with the present JDI is that the expected increase in job difficulty with skill level begins to plateau at the 7-skill level. On this basis, some 9-skill level supervisors are doing jobs which appear to be no more difficult, perhaps even easier, than 7-skill level people. The fact that a graph of the number of tasks performed against skill level mirrors a graph of job difficulty against skill level has led to the suspicion that the JDI is too closely tied to the number of tasks performed. Again, this may reflect the failure of the initial study to adequately capture supervisory jobs which generally consist of a low number of tasks.

An Alternative Design

The alternative design reported in this paper seeks to be as

faithful as possible to Mead's original design while recognizing the problems which have come to light in the application of the JDI. The criterion value is still the mean rank order value based on supervisor rankings of 25 PDs, but the selected PDs include more supervisory jobs.

The career field chosen for this study was AFSC 472XX (Vehicle Maintenance), one of the AFSs studied by Mead. The total population was divided into four separate groups representing each skill level but combining the 9-skill level and CEMs in one group. Sixty individual PDs were randomly selected and printed from each of these four groups giving a total of 240 PDs. These 240 PDs were randomly ordered into 16 lists and each of these lists was subsequently divided into 10 subsets each consisting of 24 PDs. The end result was 160 packages each containing 24 PDs.

One additional 5-skill level PD was selected at random from the PDs not already selected. This PD serves as a benchmark and was added to each of the 160 packages previously prepared. Therefore,the final packages mailed to the supervisors for ranking contained 25 PDs. Each supervisor was asked to rank order the 25 PDs in his package from the least difficult (Rank=1) to the most difficulty (Rank=25). The printed task listing was restricted to those tasks which comprise approximately 60% of time spent. However, if the total task listing for any position was 20 tasks or less, the full task listing was given. In all cases, the total number of tasks performed by the job incumbent was given at the end of the task list.

Although this design is not identical to Mead's original design, it does maintain the essential features while, at the same time, addressing the need to make the JDI sensitive to supervisory jobs. As 25% of the PDs to be ranked represent supervisory jobs, the ability to capture the rankers policy is greatly enhanced. However, the design retains a good representation of lower skill level responses which will provide sufficient data to validate Mead's findings for journeyman level jobs.

The design also emphasizes the nature of the tasks performed rather than the number of tasks performed. In addition to the task listing, supervisors were given information on the percent time spent on each task and on the task difficulty of each task. These factors could be used, in addition to the number of tasks performed, to make the ranking judgement.

Results

Of 160 ranking packages mailed to Vehicle Maintenance (472XX) personnel, 107 surveys were returned. Seven of these had to be excluded as they did not have complete ranking information. Therefore 100 (62.5%) useable surveys were available to establish the criterion value. The responses adequately represented the population. For example, 31% of the population serves with USAFE and some 30% of the returns were from USAFE.

While entering the data, two facts became very clear. First, some respondents had obviously reversed their responses making 1 the most difficult position and 25 the least difficult. These respondents were identified by GRPREL, one of the ASCII CODAP suite of programs, and their input was corrected accordingly. Secondly, the additional benchmark 5-skill level PD

enclosed with all packages had a few rankings at the extremes of the scale. Generally however, the rankings were in the 9 to 16 range as expected of a 5-skill level position. In fact, the mean of this PD was 13.56 with a Standard Deviation of 5.54 and a range of 1 - 25. It was ranked 104th of the 241 PDs used in the survey.

Some other descriptive statistics give an indication of the reliability of the rankings provided. The top 40 ranked PDs were all 7- or 9-skill level positions with generally low standard deviations (average SD of this group was 3.37) The 30 PDs ranked lowest were all 3- or 5- skill level positions (average SD of this group was 2.70). The highest ranked 5-level PD was 58th on the list. The lowest ranked 9-skill level PD was ranked 167th.

These results indicated that there was a consistent ranking policy for the supervisory jobs on the one hand and the journeyman jobs on the other. The question remained as to whether or not the ranking policy was the same for the two different job types and if the policy was the same as that reported by Mead.

A comparison of the intercorrelations between the predictor variables found in this study and those reported by Mead (see Table 3) reveal some large differences. However, a comparison of the regression weights, and the beta weights (see Tables 4 and 5 respectively) reveal minor differences in the results. The figures in brackets are Mead's results.

Table 3
Correlations Among Selected Predictor Variables

Variable	Criterion	V1	V2	V3
Criterion	1.00 (1.00)[a]	-.029 (.75)	.802 (.54)	-.003 (.63)
V1	-.029 (.75)	1.00 (1.00)	-.459 (.06)	.919 (.93)
V2	.802 (.54)	-.459 (.06)	1.00 (1.00)	-.324 (.09)
V3	-.003 (.63)	.919 (.93)	-.324 (.09)	1.00 (1.00)

a Data in parentheses from Mead, 1970

V1 Number of Tasks Performed
V2 Average Difficulty per Unit Time Spent
V3 Number of Tasks Performed Squared

Table 4
Regression Weights for Predicting Job Difficulty Values

	Predictor Variables	Present Study	Mead's Study
V1	Number of Tasks Performed	.05144	(.084324)
V2	Average Difficulty per Unit Time Spent	8.46217	(7.302877)
V3	Number of Tasks Performed Squared	-.00006285	(.000121)
	Regression Constant	-32.48000	(-22.969625)

Table 5
Beta Weights for Predicting Job Difficulty Values

	Predictor Variables	Present Study	Mead's Study
V1	Number of Tasks Performed	0.93782	(1.29126)
V2	Average Difficulty per Unit Time Spent	1.06449	(.51612)
V3	Number of Tasks Performed Squared	-.52007	(-.61530)

In addition to the differences illustrated in these tables, in this study the correlation between the predicted and criterion job difficulty values was lower; R was .91 (.93) and R Squared was .83 (.86). Several other regression models were examined. These models differentiated between supervisory and journeyman jobs in an attempt to highlight any changes in policy which rankers may have used to distinguish these different types of jobs. However, the use of a number of alternative predictor variables resulted in only very minor changes to the regression weights and overall predictive efficiency.

Conclusion

Bearing in mind the high positive correlation found by Mead between the criterion value and the number of tasks performed (.75), the very low negative correlation of this same predictor variable found in this study (-.029) would indicate that concerns about the influence of this variable on the JDI equation are unfounded. In fact, when the number of tasks performed in supervisory and journeyman jobs are separated, their correlations with the criterion are .621 and -.218 respectively. However, the differences found in these correlations suggests that journeyman and supervisory jobs should be viewed separately until further analysis is done. At the least,the inclusion of supervisory jobs has slightly diluted the predictive efficiency of Mead's JDI. Further similar studies using additional AFSs are required to determine the validity of the results obtained in this study.

References

Christal, R.E. (1967). Selecting a Harem - and other Applications of the Policy-Capturing Model (PRL-TR-67-1, AD-658 025). Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division.

Mead, D.F. (1970). Development of an Equation for Evaluating Job Difficulty (AFHRL-TR-70-42, AD-720 253). Lackland AFB, TX: Personnel Division, Air Force Human Resources Laboratory.

Mead, D.F. (1970). Continuation Study on Development of a Method for Evaluating Job Difficulty (AFHRL-TR-70-43, AD-720 254). Lackland AFB, TX: Personnel Division, Air Force Human Resources Laboratory.

Mead, D.F., and Christal, R.E. (1970). Development of a Constant Standard Weight Equation for Job Difficulty (AFHRL-TR-70-44, AD-720-255). Lackland AFB, TX: Personnel Division, Air Force Human Resources Laboratory.

Wiley, L.N. (1972). Analysis of the Difficulty of Jobs Performed by First-Term Airmen in 11 Career Ladders. (AFHRL-TR-72-60, AD-757 876). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

THE DEVELOPMENT AND VALIDATION OF THE PETERS PERSONNEL TEST

Glenn E. Peters
National Security Agency

INTRODUCTION

The National Security Agency is an excepted Agency and thus its Employment Division is responsible for conducting all tests relevant to the hiring and placement of applicants. One of the tests most frequently administered is the Career Qualification Battery II (CQB II). The CQB II is comprised of nine separate and distinct aptitude tests which are administered in combination. Each subtest is then scored; the raw score is converted to a staten (a score ranging zero to nine), and then combined in a unique grouping with different weightings to produce a score which is the best predictor of performance on the job as rated by their supervisors. Categories of jobs which contain the same loading factors are established by performing a stepwise regression analysis which establishes relationships between independent variables (subtest performance) and the dependent variable (job performance). For example; for the position of clerical assistant, the subtest Spelling (weighted by a factor of two) combines with the subtest Pattern Matching (weighted by a factor of one) to form the best predictor of success. This category may contain any number of jobs which will have the same test coverage.

Eight of the nine subtests used in the COB II are multiple choice, machine scored format. This makes scoring of the test fast, efficient, and cost effective. The one exception is the Wonderlic Personnel Test (WPT) which is a commercial test of general mental ability written in a constructed response format. The examinee responds to each question by writing the answer directly on the disposable test booklet. These in turn are hand scored by the test proctor who also converts the raw score to a staten before it is factored with other tests to complete the category score. This procedure holds a distinct possibility of human error being introduced into the scoring process. It also creates a delay in the scoring process, requires extra man-hours, and is not cost effective.

The WPT has been commercially available for many years and extensive tables have been produced which displays relationships of tests scores in different job areas. Presumably, scores that are either too high or too low indicates poor job compatibility and thus, low probability or success.

The WPT has been used as part of the Career Qualification Battery II for the past 20 years. It is factored into two categories which cover 16 different jobs. To eliminate the test or substitute one of unknown parameters would be unacceptable since the interaction effects would be unpredictable. Thus, a test which would parallel the WPT in terms of content and difficulty would have to be developed. If a parallel test could not be developed it would necessitate new normative tables being developed and significant software changes. In order to preserve the consistency of the tests over time, changes of this magnitude were avoided.

PROCEDURE

Multiple Choice questions of a similar content (either verbal, mathematic, or general knowledge) were developed. Form A consisted of 50 questions for which 15 minutes testing time was allowed. The experimental test was given concurrently with the WPT to samples of high school seniors living in the Baltimore/Washington area. They were not informed that the test was experimental so it is assumed that their motivation to perform was high. Sample testing of 100 or more subjects were collected and the results were item analyzed.

The item analysis determines the merit of a test question by providing three kinds of information: 1) the difficulty of the item, 2) the discrimination index of the item, and, 3) the effectiveness of the distractors. Questions and responses are repeatedly modified with regard to subsequent item analyses. A total of three separate analyses were performed before the final form was established. Once Form A was properly established the same process was repeated to develop alternate Forms B and C.

RESULTS

Data was analyzed using SAS statistical software package on a VAX 11/730 minicomputer. Descriptive statistics were calculated using the Mean procedure and a Pearson product-moment correlation between the experimental test and the WPT was performed using the Corr procedure. Although the table below shows differences between all three forms of the test, those differences are insignificant. Correlations between performance on the experimental test and the WPT are significant (P .01). As a result, the staten conversion tables used for WPT was retained for the experimental test now titled the Peters Personnel Test (PPT)

FORM A

VARIABLE	N	MEAN	STANDARD DEVIATION
PPT	145	22.17	7.91
WPT	145	21.94	5.99

Pearson Correlation Coefficient: R = .82*

FORM B

VARIABLE	N	MEAN	STANDARD DEVIATION
PPT	145	20.20	6.41
WPT	145	19.59	5.27

Pearson Correlation Coefficient: R = .65*

FORM C

VARIABLE	N	MEAN	STANDARD DEVIATION
PPT	146	20.69	5.54
WPT	146	19.54	5.94

Pearson Correlation Coefficient: R = .76*

* Probability less than .01 ($P < .01$)

DISCUSSION

The availability of a steady supply of willing subjects (applicants for entry level position) enabled numerous revisions of the PPT to be item analyzed until a satisfactory test was developed. It is clear that parallels forms of a test can be constructed regardless of the test format. In this case a multiple choice test was made to closely emulate the qualities of a constructed response test. The potential for other kinds of constructed response tests to be converted to computer scored, multiple choice formats is excellent. The reward is lower cost, increased accuracy, and faster results.

REFERENCE

Lyman, H. B. (1968) - Intelligence, Aptitude, and Achievement. Boston: Houghton Mifflin Co.

Warren, J. R. (1984)- The Validity of the National Security Agency Career Qualification Test Batteries. Berkeley: Educational Testing Service, unpublished report.

A Refined Item Digraph Analysis of a Cognitive Ability Test

William M. Bart and Ruth Williams-Morris
Educational Psychology
University of Minnesota

Refined item digraph analysis (RIDA) was first introduced in a paper presented by W. Bart at the 1985 Military Testing Association Conference held in San Diego. RIDA permits the assessment of the diagnostic and prescriptive value of test items and the evaluation of cognitive microtheories and instructional microtheories. This paper illustrates the utility of two of those indices in analyzing the diagnostic value of a prominent cognitive ability test and its items.

The Dense Item

Central to RIDA is the dense item concept. A dense item is any test item for which one can infer 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 item from the viewpoint of cognitive diagnosis and instructional prescription.

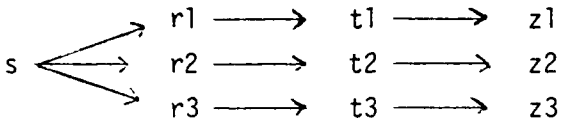
A dense item has ten properties. Each property engenders a quantitative index which indicates the extent to which a test item has the dense item property under consideration. This report highlights the first two properties: response interpretability and response discrimination.

The Refined Item Digraph

To use RIDA, one first constructs a refined item digraph for each item under consideration. A refined item digraph is a graphical representation of the inferences interrelating the item stem (e.g., "3 + 3 = _"), the responses to the item (e.g., "6"), the cognitive 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 cognitive 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, uses 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 implemented, 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 that the response was wrong and 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 (Bart, 1985).

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 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). Only between-set inferences will be considered for an item and that is why only refined item digraphs and not item digraphs are examined in this paper.

Two Properties of the Dense Item

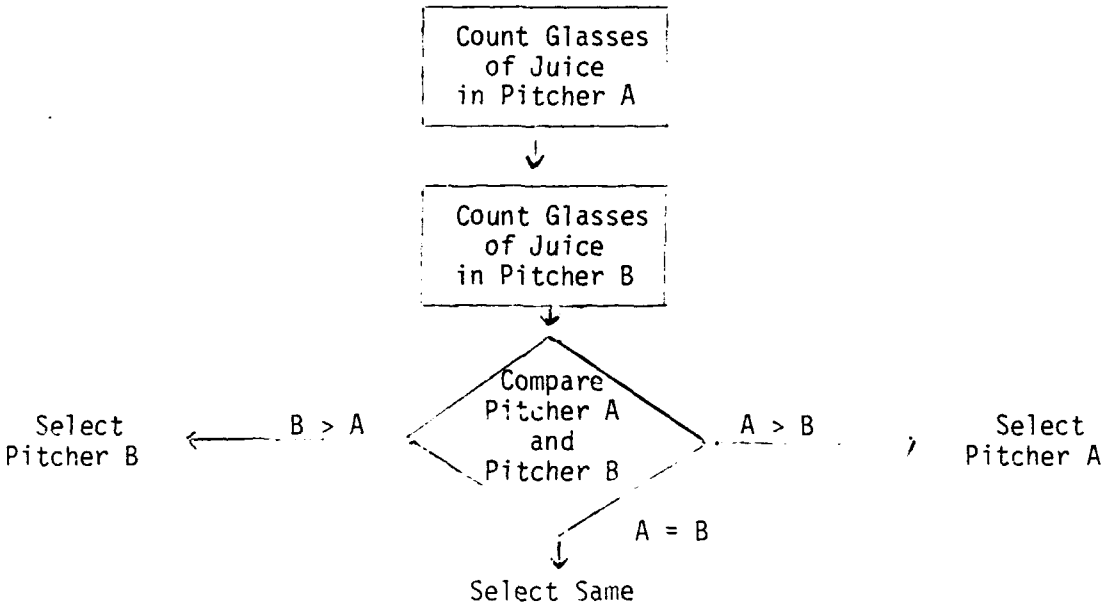
The first two properties of a dense item are response interpretability and response discrimination. A test item has response interpretability to the extent to which each response to the item is interpreted by at least one cognitive rule. The index of response interpretability for an item indicates the extent to which an item has response interpretability. This index can vary from 0 to 1.00, with 1.00 indicating the highest level of response interpretability.

A test item has response discrimination to the extent to which each response to the item is interpreted by only one cognitive rule. The index of response discrimination for an item indicates the extent to which an item has response discrimination. This index also varies from a value of 0 to 1.00, with 1.00 indicating the highest level of response discrimination. The exact definitions of these two indices were provided by Bart (1985).

The Cognitive Ability Test

The Orange Juice Test formulated by Noelting (1980a, b) was selected because that test is an important measure of proportional reasoning and because the rules used in responding to each item are specified. Figure 2 depicts one of those four rules.

Figure 2: A Rule Used in Solving Orange Juice Test Items



The Noelting Orange Juice Test (NOJT) has 25 items. Each item consists of the comparison of the relative orange juice tastes of two drinks, each composed of a certain number of glasses of orange juice and a certain number of glasses of water.

Results

The RIDA of the 25 NOJT items yielded four types of refined item digraphs and items. Each Type I item has a response interpretability index of .667 and a response discrimination index of .444. Three of the items were Type I. There were three Type II items and they had a response interpretability index of .333 and a response discrimination index of .083.

Each Type III item has a response interpretability index of 1.000 and a response discrimination index of .611. There were 15 Type III items.

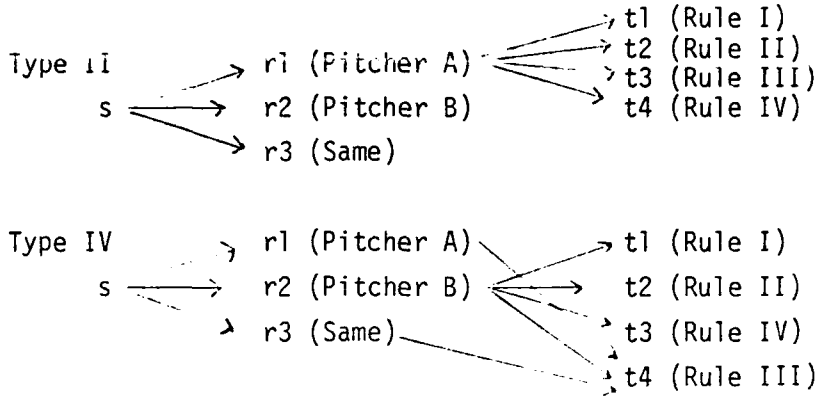
The two indices of response interpretability and response discrimination are also definable at the item response level. Use of those item response indices permitted the Type III items to be subdivided into Type IIIa and Type IIIb. There were ten Type IIIa items and five Type IIIb items.

Type IV consisted of four items with each having a response interpretability index of 1.000 and a response discrimination index of .750.

Thus RIDA resulted in a clustering of 25 items into five types with one type subdivided into two subtypes. Regarding response interpretability, all of the items except Type I and Type II items had each of their responses interpretable by at least one cognitive rule posited by Noelting. Regarding re-

sponse discrimination, Type I and Type II items had many responses that were not interpretable by only one rule. Type I and Type II items had, in general, poorer diagnostic value than Type III and Type IV items. Figure 3 depicts the refined item digraph for a Type I item and the refined item digraph for a Type IV item respectively.

Figure 3: Refined Item Digraphs for a Type II Item and a Type IV Item



One puzzle that emerges from this analysis is why the NOJT has a preponderance of Type III items marking its unequal distribution of items across the four types. This item analysis of the NOJT items highlighting only two of the ten dense item property indices demonstrates the need for further refinement of the NOJT, so that it has better diagnostic and prescriptive value.

Conclusion

This paper illustrates the use of only two of the ten dense item properties. More useful information could be gained from further analysis of the remaining eight dense item properties. Refined item digraph analysis provides a method by which the diagnostic and prescriptive value of tests and their items can be evaluated. RIDA with its emphasis on the dense item concept should facilitate research on and usage of cognitive diagnosis and instructional prescription and should predispose the concurrent consideration of test items, cognitive theories, and instructional theories.

References

Bart, W. (1985). How qualitatively informative are test items: A dense item analysis. Paper presented at the Military Testing Association Conference, San Diego, CA.

Noelting, G. (1980a). The development of proportional reasoning and the ratio concept. Part I - Differentiation of stages. Educational Studies in Mathematics, 11, 217-253.

Noelting, G. (1980b). The development of proportional reasoning and the ration concept. Part II - Problem-Structure at successive stages: Problem-solving strategies and the mechanism of adaptive restructuring. Educational Studies in Mathematics, 11, 331-363.

Siegler, R. (1976). Three aspects of cognitive development. Cognitive Psychology, 8, 481-520.

Acknowledgement

The first author wishes to express his appreciation to the Wilson/University of Minnesota College of Education Alliance.

Cigarette Smoking, Field-Dependence and Contrast Sensitivity

Bernard J. Fine and John L. Kobrick
U.S. Army Research Institute of Environmental Medicine, Natick, MA

This study examined the separate and combined effects of cigarette smoking and field-dependence on contrast sensitivity. Each of these variables is related to many aspects of human perception (e.g., 13,20,23). As far as can be determined, there has been no previous research on these relationships.

Contrast sensitivity is the ability to distinguish an object from its background under varying contrast conditions; contrast is the difference in brightness between object and background. Visual acuity, the traditional threshold index of visual resolution, represents only 1 extreme of contrast sensitivity, that of discriminating stimuli of high spatial frequency under conditions of high contrast, e.g., the Snellen Eye Chart. In recent research (7,12,13), discrimination of highway signs and detection of targets were found to be related to contrast sensitivity but not to acuity. These results reflect the importance of measuring visual resolution throughout the entire continua of both contrast and spatial frequency, and suggest that acuity may not be the optimum measure of visual resolution.

Contrast sensitivity is based on the determination of thresholds for the detection of differences in contrast at various spatial frequencies of alternation. Frequencies are stated in cycles/deg of visual angle subtended by the retina. Contrast is defined as Lmax-Lmin/Lmax+Lmin in which Lmax is the highest and Lmin the lowest luminance of the alternation pattern. The reciprocal of this threshold contrast value is known as contrast sensitivity.

The infinite variety of spatial detail contained in the world around us generates spatial frequencies and contrasts which vary from moment to moment. This variable array of stimuli triggers contrast sensitivity responses. The limiting parameters within which this continual process occurs are determined by the physical characteristics of the prevailing field of view and by perceptual characteristics of viewers. Our experience with the perceptually-based variable "field-dependence" has led us to infer some of the relevant characteristics of viewers, and to include that variable in the present study.

Field-dependence refers to the ability to perceive a relatively simple shape ("figure") e.g., a triangle, when it is hidden ("embedded") in a more complex background ("field"). Individuals better able to detect the figure are referred to as "field-independent;" those who have difficulty overcoming the embeddedness are called "field-dependent." Significant relationships have been found between field-dependence and many psychological variables, but little attention has been paid to underlying "causes." In general, differences in field-dependence are assumed to originate in early childhood experiences.

Fine (8) conceptualized individual differences in field-dependence as representing basic, probably genetic, differences in nervous system development. Development, manifested by factors such as number, size, location, responsiveness and/or organization of neurons, and modulated by the amount and quality of neural transmitter or other substances at the cellular or molecular levels, was conceived to proceed so as to enable increasingly better discrimination among, and organization and integration of neural responses to stimuli. This presumably higher level of development, or greater

"sensitivity," of the nervous systems of field-independent persons was felt to underly their superior performance on the Hidden Shapes Test, a measure used to define field-dependence. Fine predicted that this sensory "superiority" should result in better performance on other perceptual tasks and verified the prediction in several studies of the discrimination of colors (9,10,11).

Consistent with the foregoing, it was predicted that in the present study field-independent persons would be more sensitive to contrast (have higher contrast sensitivity scores) than would field-dependent persons.

The effects of tobacco smoking on health generally are well known and quite clearly defined; the effects on performance are much less clearcut (14,15,22). Smoking has been found to facilitate some types of perceptual performance, but to impair other types. Given the aforementioned relationships between contrast sensitivity and important aspects of military performance (12,13) and the widespread incidence of smoking in the military (52-53% smokers;3), investigation of the relationship between the two variables seemed advisable. We have been unable to locate any previous research on this topic.

Method

Subjects (Ss): Ss were 25 military and 3 civilian volunteers ranging from 19-40 years of age (mean= 23.8; median= 23) Twelve participants were cigarette smokers (10 or more cigarettes/day) and 16 were non-smokers.

Procedure: Ss took part on 3 different days. Day 1 involved instruction and practice on a number of tasks. Then, Ss were tested on the tasks on 2 successive mornings or afternoons. Time of testing and order of tasks were constant for each S. Half of the smokers smoked a cigarette during each of several 5-minute rest periods prior to task performances on the 1st day and abstained on the second day; the rest of the smokers followed the opposite regimen. Results of only the contrast sensitivity task are reported here.

On a deprivation day, smokers were not permitted to smoke during the testing session or for 90 minutes prior to it. Time of deprivation prior to performing the contrast sensitivity task varied from 90-180 minutes, depending upon the order of task assignments. Smoking was done in an area removed from non-smokers. Measures of personality and cognitive style and questionnaires about smoking habits and demographics also were administered.

Measures: (a) Field-dependence was measured by the Gottschaldt Hidden Shapes Test (5). Based on norms from 1000+ soldiers, Ss were classified as field-dependent (scores of 19 or below), field-central (20-26) or field-independent (27 or higher). (b) Contrast sensitivity was measured with the Nicolet CS 2000 Contrast Sensitivity Testing System (16), using a standard method with 8 trials, in each of which a sinusoidal grating was presented on a video screen to an S seated 3 meters away. The first 2 trials (gratings of 0.5 and 6 cycles/deg of visual angle) were for practice. The remaining 6 trials, which were scored, were with gratings of 0.5, 1, 3, 6, 11.4 and 22.8 cycles/deg.

Each trial began with a preview of the grating pattern for that trial, introduced by 2 tones. Contrast of the pattern was increased from zero to maximum in 1 second, held for 2 seconds, then decreased to zero, ending with 2

tones. A single tone then noted the start of the test. The previewed grating was first presented at zero contrast. Then, contrast increased until S pressed a button upon detection of the grating. Keeping the button pressed, S caused a decrease in contrast until he could no longer see the grating, whereupon he released the button, causing the grating to gradually reappear. By alternately raising and lowering contrast in response to S's signals, the instrument tracked the threshold for perception of the grating. A trial consisted of 4 ascending and 4 descending responses at a given frequency. The 6 test trials were the bases of the contrast sensitivity functions (CSF's) for each S.

Results

Smoking and CSF: The study design permitted two types of smoking comparisons: (a) smokers when smoking compared with when deprived of smoking and (b) smokers compared with non-smokers.

A comparison of the CSF's of smokers immediately after smoking with those obtained in deprivation showed no significant differences. Accordingly, the CSF's for the smokers were averaged across smoking and deprivation sessions to compare with the data of the non-smokers. The data for the non-smokers also were averaged across both test days, an analysis of variance (ANOVA) having shown no significant differences in CSF's between the two days. A comparison of the 2-day averages of smokers and non-smokers is shown in Figure 1. The non-smokers had significantly (t-test) higher mean contrast sensitivities at frequencies of 3 cycles/deg (p=.09, 2-tails) and 6 cycles/deg (p=.06, 2-tails) and near significant superiority at 22.8 cycles/deg (p=.12, 2-tails).

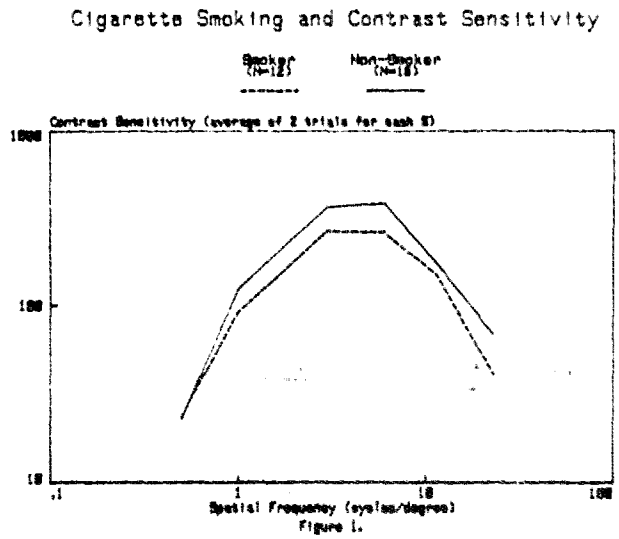


Figure 1.

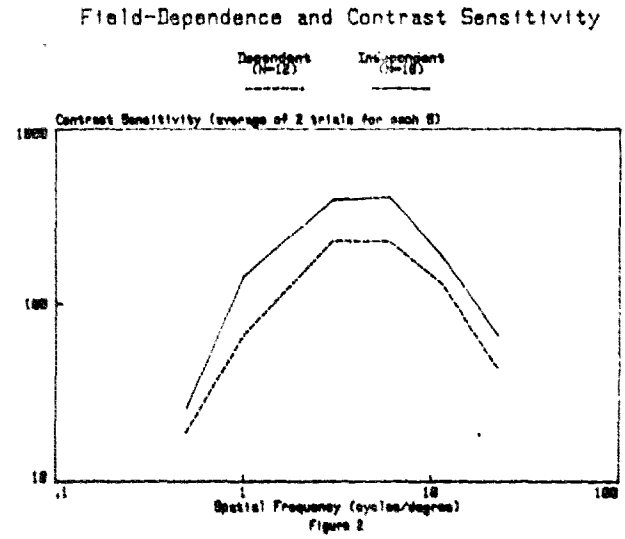


Figure 2

Field-dependence and CSF: Since field-independent and field-central groups did not differ in CSF's, their scores were combined to compare with the field-dependent group. (The combined groups were termed "field-independent.") The results are shown in Figure 2. The relationship predicted between contrast sensitivity and field-independence was substantiated. The "field-independent" group had significantly (t-test) higher mean contrast sensitivity levels at all spatial frequencies. The p-values (all 1-tail) for the .5, 1, 3, 6, 11.4 and 22.8 spatial frequencies were .07, .002, .002, .05 and .10 respectively.

Smoking, field-dependence and CSF: Analysis of data for the 4 sub-groups shown in Figure 3 indicated that at all frequencies except the lowest, the "field-independent" non-smokers had significantly higher mean contrast sensitivity scores than the field-dependent smokers.

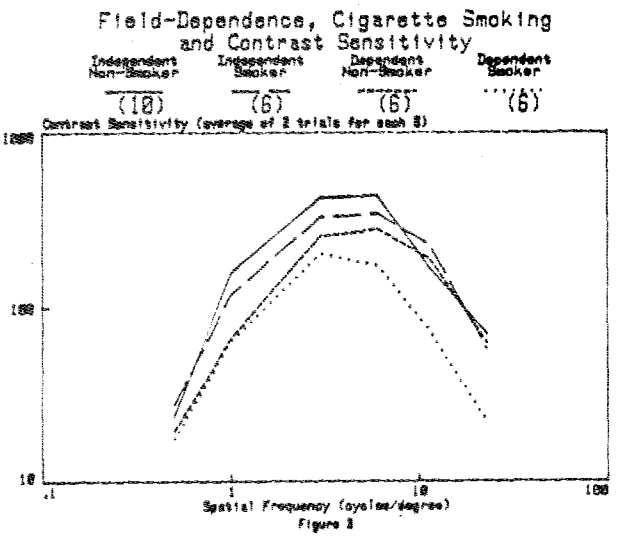


Figure 3

Discussion

Based on Ginsburg's research (12,13), we believe it reasonable to expect pilots with poor contrast sensitivity to be at greater risk for accidents when flying in mist, smoke, fog or whiteout or at dawn or dusk. Thus, our finding of an adverse relationship between smoking and contrast sensitivity raises a very serious question about the compatibility of being both a smoker and a pilot. Further research is needed to verify the effect of smoking on CSF and to relate both smoking and CSF to actual flight operations.

The strong interaction between field-dependence, smoking and contrast sensitivity illustrates the extent to which the CSF can vary from one type of person to another and has implications for studying motor vehicle accidents. We suspect that field-dependent smokers may be impaired perceptually whendriving under conditions of poor contrast and illumination. Various degrees of association already have been reported between field-dependence and automobile accidents (2), alcoholism (17), color discrimination (9,10,11), and age (18), between smoking and accidents (6), between age and CSF (19) and color discrimination (21) and between alcoholism and color discrimination (1,4).

While field-dependence was significantly related to both CSF and color discrimination (a color discrimination test was given in this study with results similar to those previously reported in 9,10,11), the latter two variables were not related at any spatial frequency. This suggests that our assumption of developmentally-related differences between individuals in "sensitivity" of the nervous system might be too general; the differences may be specific to each sensory modality. Differences between aspects of the same modality within a given individual also are likely. Thus, a person with excellent ocular color sensitivity may have poor cortical color integration

capability. Someone else may have an opposite configuration. A third person may have superior color sensitivity at all levels, but may be very poor at perceiving contrast, at the ocular level, the cortical level, or both, and this may vary with spatial frequency, and so on with all of the sensory modalities. Given these individual differences in development and function of aspects of the nervous system, concomittant differences between individuals in quantity and quality of information that has to be processed and the manner in which it is organized, integrated and responded to undoubtedly follow.

How these complex individual differences and their origins are perceived influences how science is applied to the solution of problems. If individual differences in behavior are attributed to environment, then it is implicitly assumed that people are biologically homogeneous and can be modified in similar ways by manipulations such as training or conditioning, e.g., training "good" driving "habits." However, if the differences are seen as predominately genetically based, then problem solving becomes more complicated; solutions must take account of differences in relatively unmodifiable innate abilities. With regard to motor vehicle operation, for example, we must ask what it means phenomenologically to have poor contrast sensitivity, poor ability to judge distances or discriminate colors, and we must develop data bases, principles and norms with which to modify vehicles, highways, road signs and lighting, etc. to accomodate much wider ranges of human abilities than at present.

There now appears to be a significant convergence of data denoting complex relationships between field-dependence, contrast sensitivity, color discrimination, alcoholism, smoking, aging and human performance. To integrate this knowledge and move forward effectively, unifying concepts are needed. We believe that the origin of such concepts lies in a consideration of the bases of differences between individuals. We express our concern here because we perceive that the focus of most researchers is not in that direction.

References

1. Adams, AJ, Brown B, Haegerstrom-Portnoy G, Flom MC, Jones RT. Evidence for acute effects of alcohol and marijuana on color discrimination. Perc. Psych. 1976;20:119-24.
2. Arbuthnott, DW. Effects of noise and perceptual style on reaction to simulated driving emergency. Canad. J. Behav. Sci. 1980; 12:78-86.
3. Bray, RM et al. 1982 worldwide survey of alcohol and nonmedical drug use among military personnel. Raleigh, NC: Research Triangle Institute, July 1983.
4. Burdick JA, Chebib FS. Heredity, color vision and alcoholism. Int. J. Addic. 1982; 17:815-22.
5. Cattell RB et al. The Objective-Analytic Personality Factor Batteries. Champaign , IL: Institute for Personality and Ability Testing, 1955.
6. DiFranza, JR et al. The relationship of smoking to motor vehicle accidents and traffic violations. N.Y.State J. Med. 1986; 86:464-67.
7. Evans DW, Ginsburg AP. Contrast sensitivity predicts age-related differences in highway sign discriminability. Human Fact. 1985; 27:637-42.
8. Fine BJ. Field-dependent introvert and neuroticism: Eysenck and Witkin united. Psychol. Rept. 1972; 31:939-56.
9. Fine BJ. Field-dependence-independence as "sensitivity" of the nervous system: supportive evidence with color and weight discrtimination. Perc. Mot. Sk. 1973; 37:287-95.

10. Fine BJ. Field-dependence and color discrimination ability in females. Perc. Mot. Sk. 1983; 57:983-86.
11. Fine BJ, Kobrick JL Field-dependence, practice and low illumination as related to the Farnsworth-Munsell 100 Hue Test. Perc. Mot. Sk. 1980; 51:1167-77.
12. Ginsburg AP, Easterly J, Evans DW. Contrast sensitivity predicts target detection field performance of pilots. Proc. Hum. Fact. Soc. Ann. Mtg. 1093; 269-73.
13. Ginsburg AP, Evans D, Sekuler R, Harp S. Contrast sensitivity predicts pilots' performance in aircraft simulators. Am. J. Optom. 1982; 59:105-09.
14. Heimstra NW, Bancroft NR, DeKock AR. Effects of smoking upon sustained performance in a simulated driving test. Ann. NY. Acad. Sci. 1967; 142:295-307.
15. Myrsten A-L, Andersson K, Frankenhaeuser M, Elgerat A. Immediate effects of cigarette smoking as related to different smoking habits. Perc. Mot. Sk. 1975; 40:515-23.
16. Nicolet CS 2000 Contrast Sensitivity System. Madison, WI: Nicolet Instrument Co., 1982.
17. O'Leary MR, Donovan DM, Chaney EF. The relationship of perceptual field orientation to measures of cognitive functioning and current adaptive abilities in alcoholics and nonalcoholics. J. Nerv. Ment. 1977; 165:275-82.
18. Panek PE. Age-differences in field dependence/independence. Exp. Aging R. 1985; 11:97-99.
19. Sekuler R, Owsley C. The spatial vision of older humans. In: Sekuler R, Kline D, Dismukes K, eds. Aging and human visual function. New York: Alan R. Liss, Inc., 1982: 185-202.
20. Tong JE, Knott VJ, McGraw DF, Leigh G. Smoking and human experimental psychology. B. Br. Psycho. 1974; 27:533-38.
21. Verriest G, Vandevyvere R, Vanderdonck R. Nouvelles recherches se rapportant a l'influences du sexe et de l'age sur la discrimination chromatique ainsi qu'a la signification pratique des resultats du test 100 Hue de Farnsworth-Munsell. Revue Opt. (Paris) 1962; 10:499-509.
22. Wesnes K, Warburton DM. The effects of cigarette smoking and nicotine tablets upon human attention. In: Thornton RE, ed. Smoking behaviour: physiological and psychological influences. Edinburgh: Churchill Livingstone, 1978.
23. Witkin HA, Goodenough DR. Cognitive styles: essence and origins. New York: International Universities Press, 1981.

Persons participating in this study did so only after giving their free and informed voluntary consent. Views opinions and findings herein do not reflect official Army position, policy or decision. We acknowledge the invaluable assistance of Douglas Dauphinee, Donna McMenemy, Shelley Strowman, William Tharion and Calvin Witt with data collection and of Edith Crohn with both data collection and the scoring of tests.

PERSONNEL VARIABLES AND ORGANIZATION/MISSION PERFORMANCE

Raymond O. Waldkoetter
U. S. Army Soldier Support Institute
Fort Benjamin Harrison, Indiana 46216-5060

Soldier performance is affected by any number of variables. Performance for an orgnization and mission must somehow relate to the cumulative effects of soldiers and units or personnel. Unless there is a straight forward progression between the results of personnel performance, known variable effects, and mission performance, inefficiency can readily occur. Many decision-makers and leader/managers do not recognize the potential effects of personnel variables on organizational results and vice versa. When there is stress in an organization and failure under heavy task demands, it can trace frequently to conflicting mission policies and personnel variables. The most obvious kind of organizational crisis occurs when minimally trained soldiers are assigned to perform complex tasks in hostile situations. On the other hand, when an organization constrains performance of highly trained personnel, conflict is apt to follow in many forms.

Personnel/soldier variables are intrinsic and extrinsic and can enhance or impair performance. Training and utilization of personnel must constantly inquire about levels of needed ability and achievement and whether given weapon systems and leadership will result in estimated proficiency and mission success. A soldier/system view has to be appreciated to begin to know how discrete mission objectives are to be defined and performed within allocated resources. Naturally, organizations will know to some degree which personnel variables are valued and can help or hinder success as defined. But the problem revolves about whether there are quality control procedures to insure such variables are ethically and scientifically used in decision making. Organizational tolerance may imperceptible move from allowing impaired performance for the sake of realism in operations to permissive disregard and inability to control and audit performance results. The extent to which personnel variables affect performance must have detailed analysis to define work and battle constraints and the respective corrective actions.

In the military setting a lack of proficient work is usually seen to lessen the probability of success in battle. However, there are almost infinite interpretations about skilled work, battle success, and their relationship. For one reason or another as an analyst looks at the meaning of personnel variable measures, and tries to relate them to mission results, organizational disconnects begin to defeat such auditing efforts. One has only to ask if any organization directly relates or dares relate individual performance to mission performance? I would say only when someone's performance is so incompetent, it threatens the evaluator. But then we know too there are occasional exceptions where the evaluator knows time can be used to escape an accurate report. Time and time again a mission is subverted by not dealing accurately with the effects of personnel variables. As missions become critical the nearly absolute need to assess and utilize personnel variables more precisely becomes equally critical.

The views and opinions expressed in this paper are those of the author and should not be taken as an official policy of the Department of the Army.

METHOD

In estimating mission performance levels, planners or strategists too often assume soldier and unit performance will be at a level of full potential. No matter what personnel variables are utilized they will seldom be applied at an optimal degree of efficiency. Then, the organization/mission will only approximate full success at given decision points and may ignore some needs altogether as personnel variables function erratically. There is a temptation to exaggerate limited success and excellence and devalue system failures until a crisis situation. It is apparent in view of missile and nuclear disasters that personnel variables and mission exigencies are so interconnected as to defy separate analysis. As soon as job consequences are treated more ethically and scientifically, the analyst or leader/manager can begin to estimate more accurately what personnel must do to lessen mission risk. The economic and political consequences for an organization are nearer predictable analysis, if there are clear pathways from personnel variables to measurable organizational objectives (Butcher, 1986).

Once it is agreed that soldier performance is readily subject to measurement, the tie in with critical mission needs can proceed to connect personnel to organization tasks requiring specific performance. (The reader can only note this is not a very original idea). Yet organizations continually fragment personnel analysis and lose their potential audit trail. A movement toward mission or strategic analysis is growing to resolve the unending crises for missions, which are changing frequently and demand predictable human resource and performance indexes. Although the need to refine personnel measurement is most defensible, it will not really pay off if personnel and decision-makers are unable to apply related data to improve the overall cohesion and efficiency. An organizational assessment survey (Short, Lowe, & Hightower, 1985) can show those variables which affect a mission and personnel, but may not clarify personnel or mission deficits toward a particular level of organizational performance. This paper is not protesting the lack of measurement capability. There is in this writer's opinion a lack of properly integrated measurement. By merging personnel variable, operational, and organizational measures increased efficiency will result with ethical decision-making to attack mission problems. The challenge goes to the leader/manager to work for progressive interpretations of data, and then test the cumulative information with ethical criteria as well as being satisfied technical objectives are secured. Much can be done using basic ordinal (ranking) measurement, if the integrity of the measurement situation is safeguarded.

It is rather well known that hardly any leader/manager selects a marginal performer to do a critical task, and rarely insists on having a high performer assigned to a low-skill task without obvious reason. But if personnel utilization is not strictly regulated, some organizations must deal with inequitable skill allocations to execute impossible mission tasks. Any political manipulation of personnel variables without ethical analysis will tend to reduce mission success. The simplistic method advocated here is to strive for a scientifically valid audit scheme where all measurement procedures have predictable interface points and results with constant review of any adverse impact on personnel and the organization. Too few missions are

assessed from the viewpoint of ethical impact, but whatever the ethical framework, strategies are successful if inherent value judgments are made according to a verified ethical system. Without relatively accurate information and even a pragmatic scheme of "right" and "wrong," decision makers will tend to rely on the bias of personal experience, their own "human equation" of perception. We need to perceive things as correctly as possible, and then obtain some detailed consensus before formulating a judgment or decision. Most often when harried decision-makers make ineffective choices their excuses resort to the anomalies or ambiguity of situations. The real flaw could be improper preparation before reviewing analytical methods, or having no conscious metamethodology for the given problem conficts.

RESULTS AND DISCUSSION

To develop and predict accurate results for battles, missions, and specific operations, analysts have to consider valid sets of soldier and unit performance variables. These variables are not unknown; however they do have to be ordered in some acceptable aggregate. The time sequence for analyzing personnel variables can make a difference in terms of variables selected and expected effects on mission. The long-and short-range use of variables and "life" of some task skills must be understood in planning and mission execution. Along with time there are considerations of degrees of personnel skill and versatility and conditions under which performance is attempted. Besides the soldier and unit performance variables there are the ultimate conditions of battle. There is a need to think of conserving human and materiel resources, and yet risking everything if the objective is valuable enough. When resources are applied the effect desired is not only to show that tasks and skills are proficient but that a comprehensive objective of battle success is possible with justified cost. Even though personnel variables and organizational performance are coordinated for positive effects, there is a continuous tradeoff to exercise quality control on the positive and adverse factors. No matter how resolute a decision-maker is in the leader/manager role, the hidden variables associated with adverse conditions in continuous combat will reduce the performance first of soldiers and units and then the mission. Positive factors of high skill and materiel superiority are vulnerable to the most basic of adverse factors. The decision maker who insists on operating in spite of fatigue and stress will reduce mission capability and unit and soldier performance.

Looking at performance reduction from a quality control or command viewpoint, an analyst projected the progressive decline in performance of mechanized infantry squads and platoons under sleep loss, stress, and fatigue during continuous operations. Figure 1 shows the progressive decline of various types of combat activity computed to examine the critical abilities required for combat tasks (FM 22-9, 1983). Different types of units vary according to types of combat activity. Task behavior may be taken for granted, but as the types of activity requiring complex reasoning begin to suffer all performances gradually lessens across time. Even the addition of new personnel may not stop such decline with heavy casualties.

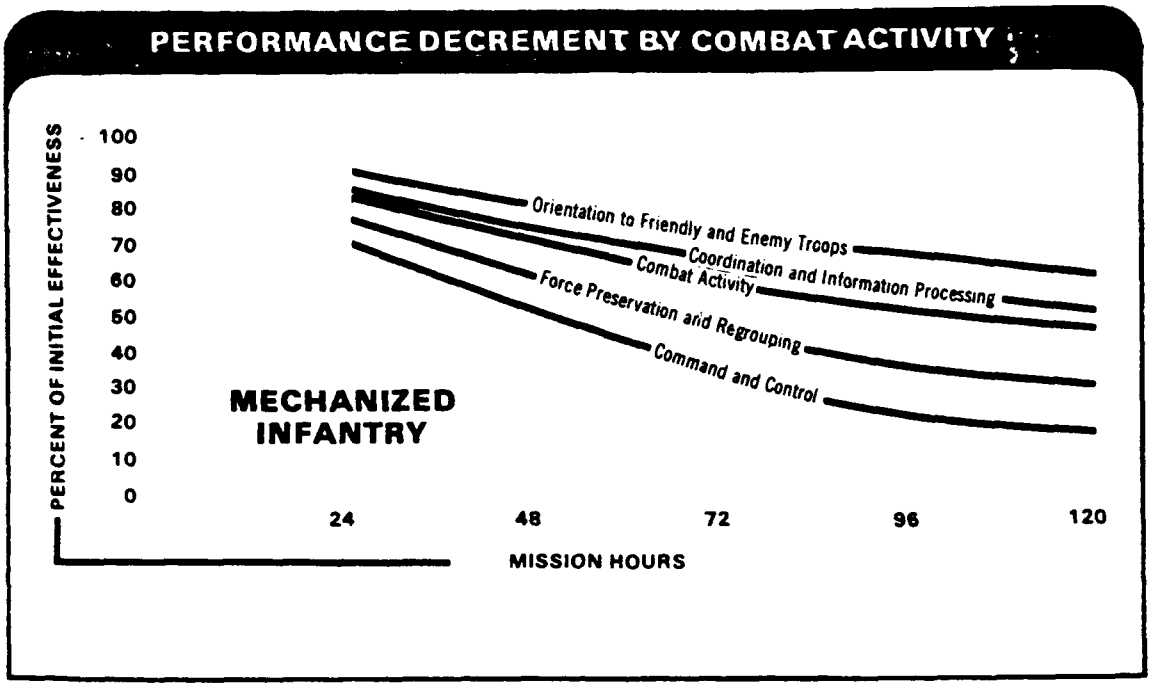


FIGURE 1

Another perspective of time effects was presented by showing a rating on the "probability of inadequate performance." In a future battle scenario, it was suggested that "frequency rate of task performance" was significantly correlated ($r = .66$, $p < .01$) and was more vulnerable to a battle context. Ordinarily "frequency rate" can assure a task is less apt to suffer a performance deficit by being routinely used. Under battle stress, though, greater demand and exposure increases a possibility of failure (Waldkoetter, Bridges, & Sterling, 1984), as noted by raters evaluating combat and communcation tasks. So as combat activity and tasks are examined, the changes over time are to be reckoned with as well as changes in value and priority of given objectives. The leader/manager capability to examine a vast array of data is not always as important as being able to have a series of situation samples that identify major personnel and mission variables. Performance decrements may act as warning signals, if there is a quicker loss of momentum and responsiveness than would be estimated under adverse circumstances.

Where computer models and games may partially explore soldier variables, it is necessary to study their effects much more closely to improve analysis. In one modeling study on soldier variables (Miller, 1985) the probability of a successful target sequence was significantly related to higher target engagement for soldiers with over two years in the MOS. Also the "time-to-fire" was significantly faster for soldiers with over six months in the MOS.

These findings in an air defense scenario showed that personnel variable modeling was sensitive to time experience and many standard personnel variables were of less practical value. Modeling of soldier performance though possible may have a difficult quality control structure, because numerous variables will require detailed analysis and the associated model or game may have limited application.

Whether in models or other analysis structures, the variable of performance time can only affect action when a level of task skill has been achieved. In a race or solving a problem, time will define the result or learning curve. Knowing at what point a measure will best describe the intended performance has always been the worst restriction on systematizing observations about personnel and mission variables. Not only does the "task life" length pose as an issue with its maintenance, the "time-location" decision for testing or using the task tends to affect is temporal validity. Both the soldier performance and relation to other action and combat phenomena must be viewed so that an evaluator is also included in the solution. This may cause observer bias or a great leap forward in ethical and scientific judgment, depending on the available data and observer's role (Novick & Cowley, 1986). If time is the velocity (V) or movement to a point where a variable is examined, and scope of task consequences relating to the variable is mass (M), the resulting action to use the variable data and respond to the consequences is the application and release of force or energy (E). Accordingly, there may exist a parallel reality dealing with personnel variables and mission performance that is in consonance with Einstein's theory of relativity.

Soldier variable effects (friendly/hostile) can be treated more realistically as computer models of missions/operations accept greater variations related to personnel impact. The obvious realization that personnel variables can determine the actual validity of models and games is causing many planners and players to devise techniques to more accurately enter the soldier in the simulated system. Being able to design models or games based on constrained assumptions and parameters does little to enhance the skill of leaders/managers, and may impose inhibitions toward creative ideas. In moving to apply and build effective decision-making systems and expand on our use of personnel variables to exert greater mission control, there will be a likely trend to indulge in extensive artificial intelligence and robotics solutions. A critical treatment of ethics or values must go simultaneously with the use of scientific development in these related areas (Chao & Kozlowski, 1986). This is not to advocate ethics and science as separate functions but rather that they be wholly combined in the analysis process.

Personnel variables must not be managed by policies which evade ethical analysis and do not comply with reliable operational practices. The organization and mission are fulfilled to the degree that personnel or soldiers are utilized on needed tasks to perform valid operations within conscious time limits.

REFERENCES

Butcher, J. N. (1986). The Minnesota report: The personnel selection system for the MMPI. Minneapolis, MN: National Computer Systems.

Chao, G. T., & Kozlowski, S. W. T. (1986). Employee perceptions on the implementation of robotic manufacturing technology. Journal of Applied Psychology, 71, 70-76.

Miller, C. R. (1985). Modeling soldier dimension variables in the Air Defense Artillery (TRASANA TEA-2-85). White Sands Missile Range, NM: U.S. Army TRADOC System Analysis Activity.

Novick, A., & Cowley, G. (1986). An imagined world. The Sciences, 26(6), 54-60.

Short, L. O., Lowe, J. K., & Hightower, J. M. (1985). Initial standardization of an Air Force organizational assessment survey instrument. Proceedings of the 27th Annual Conference of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

United States Department of the Army, Headquarters, Soldier performance in continuous operations. Field Manual No. 22-9. Washington, D.C., December 1983.

Waldkoetter, R. O., Bridges, L. T., & Sterling, B. S. (1984). Mission area analysis and expected soldier performance in battle. Proceedings of the 26th Annual Conference of the Military Testing Association. Munich, FRG: The Psychological Service of the Federal Armed Forces of Germany.

Building Cohesion the Old Way: From the Ground Up[1]

Robert F. Holz

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

In late February 1986 and continuing through the present, I began a study of cohesion among U.S. Army soldiers in the 2nd Brigade of the 10th (Mountain) Division-Light. This paper presents some perspectives derived from both interviews and non-participant observations and the "sensing" that comes when one has spent time talking with soldiers in the field and in garrison.

Units in the 2nd Brigade are not only light infantry, but are part of the Army's COHORT (Cohesion, Operational Readiness and Training) program. As such, all of the first term soldiers entered the Army together (in September 1985) and received one-station unit training (OSUT) at Fort Benning, GA. While the Division HQ of the 10th is located at Fort Drum, NY, the 2nd Brigade was activated at Fort Benning, GA, because Fort Drum is still in the process of building the troop billets, offices and facilities needed for a Division-sized unit.

The officers and NCOs in the 2nd Brigade were generally selected and assigned to the Brigade in late September 1985, and between then and the "arrival" of the first term soldiers, in January 1986 after OSUT, these officers and NCOs developed the necessary programs, procedures and endless other requirements associated with establishing a unit "for the first time." While most of the officers I met with appeared enthusiastic about their assignments to a light infantry division the same cannot be said for the NCOs. The latter, as it turns out, were frequently selected for assignment to the 2nd Brigade based on their recent experiences as first line trainers, supervisors of active duty troops. Moreover a number of these NCOs had just completed a three year tour as the Platoon Sergeant of another COHORT Company and tended to view their assignment to yet another similar unit as uncalled for "punishment" to them and their families.

The primary reason for this view is that an NCO in a combat unit will, of necessity, spend a significant amount of his time "out in the field" engaged in training soldiers to function as units. While this may well be seen by others as the "primary job" of the NCO, one needs to consider the potentially adverse impact on both the NCO and his family (the vast majority of NCOs are married and have two or more children). The more time one is required to spend engaged in field training the less time one has to devote to one's family and to restoring oneself. For the majority of officers and NCO's I talked with this issue was seen as a major sticking point. The NCO's felt put upon while they viewed the officers as having an "easier time" due to the fact that officers were rotated from position to position every 14-18 months. Additionally, some smaller number of NCOs reported

[1]The opinions expressed are those of the author and do not necessarily reflect the position of the U.S. Army Research Institute or of the Department of the Army.

that their previous assignments had been "behind the desk" at a staff, headquarters, or TDA organization. For these NCOs the rigors of field training were quite new and required that they become physically conditioned very rapidly.

In an effort to "train the trainers" all of the Officers and NCOs were exposed to a Light Leaders Training Program. This program, conducted prior to the arrival of the first term soldiers and lasting upwards of three weeks required the cadre (officers and NCOs) to conduct training in the field the way they would ultimately train their soldiers. What this amounted to was, then, role playing by the cadre. The Battalion Commander played the role of the Company Commander while the latter played at being a Platoon Leader. Senior NCO's trained as they would expect Mid-level and junior NCOs to train and so forth. Additionally, on the recommendation of the Brigade Commander the officers and NCOs throughout the Brigade were urged to undergo Ranger training as well as be jump qualified so as to ensure their physical and psychological fitness. These "rites of passage" for the officers and noncommissioned officers led, in the main, to a considerable degree of personal and unit pride which carry over to the present.

Lest the reader be lulled into complacency or be led to believe that everything is "just super" among the officers and NCOs of the brigade in question the following information (obtained from officers and NCOs alike) is offered: Company level and below officers and NCOs (E5-E7) report that they are unable to plan beyond a one or two week timeframe. These individuals are becomming increasingly concerned that the "pace of training", close to 70% of the unit's time is spent in the field on exercises of between four days and three weeks, is leading to burn-out for both themselves and for their troops. Mid-level NCOs are particularly vocal regarding the extent to which their assignment in a newly formed light infantry unit will preclude their "professional development" (their ability to attend Advanced NonCommissioned Officers Schooling or other specialized courses viewed by NCOs as necessary for their career development). Officers, on the other hand, appear less concerned with this issue (advanced schooling) probably because the majority of them know that their individual assignments will only last for between 14-18 months and that they will then be afforded the opportunity to take some other assignment within the Brigade or for some to attend specialized officer courses.

Another "sore point" raised by a number of Company grade officers and by many of the NCOs as well addresses the issues of "command sponsored family programs" for married personnel. With the Army's emphasis on families it is not surprising that family support programs are popping up within this brigade as well as elsewhere in the Army. What needs to be recognized, by senior leaders and policy makers alike, however, is that many officers and NCOs do not appreciate being told or directed to have their wives attend a family support group meeting or that the "old man's wife was annoyed that so few wives showed up at the last family council meeting." When soldiers see the Army as infringing on their turf they will react negatively, regardless of how well intended the underlying idea behind the meeting. This is likely to happen unless such family support group meetings and programs are made truly voluntary and information about the program is provided directly to the soldier's spouse. By ensuring that the spouse receives the information yet retains the option to attend or not based on personal desires or preferences these programs will be enhanced.

In spite of the fact that the officers and NCOs "know" that they will be with their troops for the duration (at least for the NCOs this is true and certain) few of the personnel I talked with have as yet begun to appreciate the opportunities for accretive training that the personnel stabilization of COHORT provides. My own view is that this situation will likely change over the next several months as the cadre complete their "event" training (Platoon level Army Training and Evaluation Programs (ARTEPs), Company level ARTEPs and finally Battalion level ARTEPs) This issue, however, will have to monitored as one of the underlying purposes of the COHORT program is to provide the unit cadre with the time to conduct planned training that increasingly tasks the abilities and capabilities of themselves and their troops.

We turn now to an overview of life in the Battalion I have been studying from the perspective of the first term soldiers a view from the bottom of the totem pole (or mountain) as it were.

I have now had discussions with over twelve different squads from different platoons in each of the line companies of the battalion I have been learning about. The soldiers have impressed me tremendously with their openness, candor and downright "smarts." I have been studying Army soldiers for the past fourteen years and I can state without hesitation that these soldiers are really different from those I used to talk with and listen to back in the early 70's the mid-70's, and the early 80's. These soldiers are highly motivated and they are bright. Their motivations, however, may surprise some among you. The vast majority have entered the Army in order to take advantage of the Army College Fund Program. These young men really "have their stuff together." They recognize the importance that a college education will mean for them in terms of advancement and they are excited that they will be able to attend college after completing a tour of duty with the Army. Were it not for this program they would probably not have enlisted. Moreover, these young men appear to have decided to defer certain issues until they have completed their first tour and ETS at the end of three years. The vast majority of these soldiers are single and report that they will remain so until they have finished their tours. While almost all have one or more girl friends, the notion of getting married as an enlisted man is not high on their agenda.

Another feature that impressed me was the physical condition of these men. They are sound of body and are proud of it. They enjoy demanding physical training and, with some exceptions, really get a kick out of "humping a ruck"and climbing up and down the hills of Georgia. Another thing that these soldiers tell me they really like is to see their officers and their NCOs right out there with them going through the same rigors and difficulties and eating the same lousy food. This latter point speaks well for the Light Leaders Program and to the dedication of the officers and NCOs who, despite numerous problems, make it a point of leading by example. Apparently this is paying big dividends. What some have called vertical bonding - the degree to which soldiers identify with and positively relate to their officers and NCOs - is really happening within the 2nd Brigade of the 10th Mountain Division.

The degree and extent of "horizontal" bonding - a "we" feeling among the troops is also very prevalent. Soldiers in the squads I have talked with do things as a squad or in some cases as an entire platoon, not as

individuals. For example, in July the entire Brigade went on "block leave." A two week vacation had been declared by the Brigade Commander during which time both he, his staff and just about every other officer, NCO and troop took off for two weeks of R & R. This concept of "block leave" merits further attention in that it provides a way for a unit to give people time off and also controls such leave so that at other times (all other things being equal) the vast majority of personnel will be available for duty when needed. But back to the troops and to horizontal bonding. If, in a given squad or platoon, one soldier is not able to go home or has no place to go, then the other squad members get together and arrange to "take him along with them." This same situation applies in a different, but equally telling situation - attendance at jump school. Recall that many of these troops indicated that they liked being physically challenged and accordingly the majority want to become jump qualified. Currently, however, the Infantry School at Fort Benning is only able to offer jump qualification courses to individual soldiers from within the 2nd Brigade due to other fill requirements. This situation results in one or two soldiers in any given platoon being offered the opportunity to go to jump school- which they dearly want. The reaction, however, is that these soldiers turn down the offer to attend as individuals. They have indicated to their Platoon Sergeant that they want to go to jump school but that they will go as a platoon or a squad but not as individuals. These troops are "tight" with each other. Recall as well that these soldiers have been together, in the Army, in the same platoon since they started Basic Training. They "know" who can be trusted and who cannot. They know who is "shamming" and who is working to the utmost. The stabilization program, COHORT, has really worked for these young first termers.

In addition, housing may also be contributing to this high level of horizontal and vertical bonding that appears to be operating within the soldiers of the 2nd Brigade of the 10th Mountain Division.

The soldiers of this Brigade have, for reason beyond their control, been assigned to Fort Benning, Georgia, which as most of you may know is a training center and school not an active duty Division installation. Accordingly, the physical facilities that the soldiers in the 2nd Brigade live in are barracks designed for basic trainees- not for active duty or permanent party personnel. There are no semi-private rooms for the soldiers of the 2nd Brigade. All of the soldiers who "live on post" (the first termers) live in platoon sized open barracks bays. Now this is really different then the situation that almost every other first term soldier in the US Army in the Continental US is faced with. Way back in the early 1970's as the Army was gearing up to deal with the abolition of the draft and the start of the volunteer force some ostensibly smart folks (maybe they were civilian maybe not) decided that in order to attract and retain soldiers under a volunteer program the Army would have to change its policies regarding barracks for single soldiers and build nice semi-private rooms. The appeal of this argument was based on the notion that if a young person of 18 or 19 was going to leave home and go off to college then this person would expect - nay demand -that he be housed in an environment that respected his privacy and individuality. Not to be outdone by college campuses, the Army embarked on a major program of construction, did away with barracks for single soldiers, and built nice, semi-private rooms. Now for those of you who remember WWII and Korea, this may sound like coddling the troops, and many a senior NCO said the same. But the policy was set. Barracks bays for enlisted soldiers with an NCO "living down

the hall" were a thing of the past - the "brown shoe Army"- not the Army of the 70's.

Now a funny thing seems to be happening to the first term (single) soldiers in the 2nd Brigade of the 10th Division. All of these troops live in platoon sized open bay areas. The bays are neat, clean and new. In fact all of the facilities for the 2nd Brigade at Fort Benning are new - they call them "star ships." They are brick, air conditioned, three story buildings connected by exterior and interior walkways and stairways. There is a new mess hall and the Brigade HQ occupies an equally new and impressive building just across the road from both Battalions. But the bays, for troops, are the rule. And aside from some grousing about music being played too loud on stereos and a general lack of individual privacy, the troops seem to be dealing with the situation rather well. Moreover, within each platoon bay the troops have done a little bit of their own redecorating. Wall lockers have been moved around and bunks have been moved so that generally speaking while each bay holds roughly 40 men, islands of four to six men each have been created. These islands are made up of soldiers from the same squad within the platoon and they like living that way. When I have talked with these troops and asked them whether or not they felt put upon by being made to live in barracks or the extent to which they "perceived" other soldiers who lived in semi-private rooms down the road - in the 197th Brigade, an active duty unit assigned to Fort Benning to provide "school support" - as having a better deal, their almost unanimous response intrigued me. These soldiers tell me that first of all they have been living in barracks since day one in the Army and that is all they really know about living quarters for single soldiers. As far as "those guys down the road in the 197th" the soldiers I have talked with don't really see them as soldiers. At least not as fighters and mountainmen. The guys in the 197th, you see, need to have semi-private rooms because their mission is to "support the school, to pick up trash, and to pull guard duty." The soldiers's in the 2nd Brigade of the 10th Division, on the other hand are lean, mean fighting machines ready to deploy at a moments notice. For them, barracks living is "no-o-o problem."

Another feature of living in the barracks is that in just about every platoon there is a squad leader who lives right there in the barracks along with the troops. Now he may not "live in the open bay" but has a small room at one end of the bay. But he still lives there. Generally, this squad leader is single or in the process of separating or divorcing his spouse or may be married but his family lives elsewhere for many different reasons. The important thing here, however, is that there is generally an NCO - a few years older then the troops - living right there in the barracks with them. And the troops seem to like this. This squad leader is viewed by many of the troops almost as a big brother. He is there to break up fights when and if they occur. He is also there to guide, to instruct and to set a role model for the soldiers. It kind of reminds one of the way the Army used to be like in them "old days " and the way the U.S. Marines Corps still operates - a military element with reportedly the highest levels of esprit de corps, motivation, morale, and cohesion of all of the the U.S. forces. It appears as if living in a barracks with the other guys from your platoon may foster cohesion. What I don't know, and what will need to be assessed/monitored is the extent to which living in the barracks begins to "turn soldiers off" after they stop spending so much time out in the field training and begin to spend more of

their time doing what most other soldiers in the U.S. Army do - time in garrison pulling base operations, special details, ash and trash, etc.,etc. It may well be that soldiers will take to barracks life if their duty time is characterized by challenging and meaningful training. This issue, however, cannot be answered at this time.

As was the case with my initial interviews and discussions with officers and NCOs not everything from the perspective of the first term troops is rosy either. These troops are really turned off by having to go through double and triple checks of their equipment and gear prior to a field training exercise. The person responsible for checking these troops out is their squad leader. He checks them and corrects deficiencies. The problem, however, is that the checking doesn't stop there. The Platoon Sergeant checks the gear after the squad leader has done his thing and the First Sergeant often follows this up with a check of his own. Now few will argue that some redundancy in a system is necessary. The question - challenge -for the leadership of the unit, however, is deciding when enough is enough. Bear in mind, these soldiers are bright, articulate and motivated. They want to do well. They want to excel. They don't really appreciate multiple checks because this tells them that "the guys up there" (at Platoon or Company) don't really trust me. And trust is a very important component in the cohesion equation. It may also be the linchpin that ties a cohesive unit to a high performing unit. This issue will require more thought and assessment.

While I have not, thus far, tried to compile all sorts of "hard data" regarding the performance of the troops in the battalion, so as to avoid comparing data with units in the U.S. Army, I have come up with the following anecodotal information. The AWOL, Dropped From the Rolls (DFR), Driving Under the Influence (DUI) and theft statistics across the 2nd Brigade are really quite low. We are talking about four or five cases within any given company over a period of five months. Maybe the way in which the 10th Division has been training its cadre and its first termers (emphasis has been on squad and platoon level training), the living accomodations for the first termers (barracks as opposed to semi-private rooms) and the block leave practice in the 10th may be responsible for the apparent high levels of unit cohesion I have witnessed as operating within the 2nd Brigade of the 10th Mountain Division.

What this "research" has taught me thus far goes beyond the little story you have now read. My approach to learning about training, group formation, interpersonal interaction, cohesion, etc. was not to drop in on the Brigade with one of any number of surveys I have used over the years and ask several hundred troops to mark off their answers to several hundred questions on an answer sheet. Rather I have tried to learn about what's really happening by playing non-participant observer by spending time talking with soldiers and listening to them talk to me.

Thus far, my view of the mountain is a little blurred. I hope, in time to have a clearer picture and understanding of the processes involved in building a high performing unit from the ground up.

INFLUENCE OF ENVIRONMENT, ABILITY AND TEMPERAMENT ON PERFORMANCE IN ARMY MOS

Darlene M. Olson
U.S. Army Research Institute[1]

Walter C. Borman
Personnel Decisions Research Institute

Job performance has been conceptualized as a product of abilities, skills, and personal characteristics that individuals bring to the Army, of environmental experiences that influence a soldier after enlistment, and of the person's motivation to perform. Although a substantial portion of the total variability in performance criteria can be explained by individual difference factors, work environment variables related to support, training opportunities, and perceived job importance have been found to have weak, but consistently significant relationships with supervisory ratings of soldier effectiveness, Army-wide rating factors (e.g., Personal Discipline) and measures of hands-on task proficiency (Olson & Borman, 1986).

The impact of cognitive abilities, temperament, work environment and their possible interactive effects on job performance should be investigated. Peters & O'Connor (1980) have proposed that environmental factors may moderate the relationships between ability and performance. In contrast, Schmidt and Hunter (1977) have contended that the prediction of performance from ability is stable across situations and over time for various jobs. More current research (e.g., Staw & Ross, 1985) has found dispositional effects for job satisfaction criteria. Hence, research suggests that both person and environment factors should play a role in explaining the variability in soldier performance.

The model of soldier effectiveness advanced here assumes that performance is influenced by a soldier's abilities and temperament, which are measured when entering the military, and individual perceptions of the work environment developed through experience with the Army job setting. In this context, the purpose of this research was to investigate potential moderating effects of work environment dimensions on the relationship between individual differences and job performance in four clusters of Army jobs.

Method

Subjects. The sample contained 5080 first-term Army enlisted personnel in 9 different jobs. There were 673 infantrymen, 629 cannon crewmen, 485 armor crewmen, 351 radio operators, 618 light-wheel vehicle mechanics, 659 motor transport operators, 500 administrative specialists, 485 medical specialists, and 680 military police. These MOS were sampled at 11 continental United States and four European Army installations. These jobs were grouped into one combat (11B, 13B, and 19E MOS) and three non-combat clusters [Clerical (71L MOS), Operations (31C, 63B, and 64C MOS), and Skilled Technical (91A and 95B MOS)]. Previous empirical research (McLaughlin, et. al., 1984) demonstrated that the above clusters are sufficient to group Army jobs on the basis of aptitudes measured by ASVAB.

Performance Measures. Criterion development work was conducted by the Project A contractors and included construction of the following measures: 1) Army-wide rating scales relevant for evaluating soldiers in any first-tour

---

[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.

Army job, 2) job-specific rating scales, 3) hands-on task proficiency measures, and 4) job knowledge tests. The Army-wide rating scales were developed using a variant of the behaviorally-anchored rating scale methodology, and emphasize performance dimensions relevant to any MOS (e.g., maintaining equipment). The job-specific scales, which were also 7-point behavior summary scales, focus on narrow performance areas relevant to a designated job (e.g., transporting personnel for the motor transport operator job). The hands-on tests consisted of 15 MOS-specific tasks. Hands-on scores were computed for each soldier by averaging the proportions passed across the tasks tested. Multiple choice tests were developed to assess job knowledge relevant to important and representative tasks in an MOS. A total job knowledge score for each research participant was derived as a percentage of the number of items answered correctly. Factor-analysis of the performance ratings resulted in an interpretable solution: 1) Effort and Leadership 2) Personal Discipline and 3) Military Bearing (Campbell, Hanser, & Wise, 1986). Factor scores for the performance ratings, along with an overall soldier effectiveness composite based on the unit weighting of ratings on the Army-wide dimensions were used in subsequent analyses.

Work Environment Measures. The Army Work Environment Questionnaire (AWEQ), a revised 53 item multiple choice questionnaire measures the following Army environmental constructs: 1) Resources, 2) Supervisor Support, 3) Training/Opportunities to Use MOS skills, 4) Job/Task Importance, and 5) Cohesion/Peer Support. AWEQ items are answered using a 5-point frequency rating scale (e.g., 1 = Very Seldom or Never to 5 = Very Often or Always). Respondents are asked to indicate how often each environmental situation described in an item occurs on their present job. Items consist of statements such as "You get recognition from supervisors for the work you do" (Supervisor Support). Five standardized unit weighted factor scores are derived for the AWEQ.

Cognitive Ability. A composite measure of four subtests from the Armed Services Vocational Aptitude Battery (ASVAB), known as the Armed Services Qualifications Test (AFQT), was used as an assessment of general cognitive abilities.

Temperament Measures. The Assessment of Background and Life Experiences (ABLE) inventory (Peterson, Hough, Ashworth, & Toquam, 1986), which includes ten tempera-ment/biodata scales, was administered as a self-report measure of soldier temperament. From factor analysis of the ABLE a three factor solution emerged: 1) Achievement, 2) Dependability and 3) Adjustment. The Achievement factor has items loading from the Self-Esteem, Work Orientation, Dominance and Energy-Level scales. The Dependability factor contains items from the Non-delinquency, Traditional Values, Conscientiousness, Cooperativeness, and Internal Control scales. The Adjustment factor has items loading from the Emotional Stability scale.

Procedures. The rating scales were administered to groups of 15 or fewer peers or supervisors of the target ratees after they were trained using a combination error and accuracy training program. During the peer rating sessions, raters (who were in addition ratees and members of the research sample) also responded to the AWEQ. The ABLE inventory was administered in separate small group sessions. Task proficiency measures were admininstered to each soldier by experienced job incumbents or supervisors, who were trained to evaluate and score each hands-on task. MOS-specific job knowledge tests were given to groups of 15-30 soldiers.

Results

Regression Analyses. Moderated regression analysis was used to estimate the relationships of ability, temperament, perceptions of the work environment, and their interactions to typical performance ratings and more objective performance criteria. A series of four separate regression models were built for each of the four performance measures nested in each job cluster. First, the separate performance variables were regressed on an individual differences model, which contained AFQT and three temperament factor scores to determine the contribution of individual differences at the time of enlistment to subsequent job performance. Second, an environmental model, which contained the five work environment constructs was used to predict the separate performance measures to examine the amount of variance explained by these variables. Third, a full model containing both individual differences and environmental factors was tested. Finally, a set of interactions among the predictors (ability X temperament, ability X environment, and temperament X environment factors) was added to the full model and the separate performance criteria were regressed on it to determine the post-enlistment interrelationships among environmental/organizational influences on soldier performance and the expression of individual differences in ability and temperament on the job.

The regression analyses are presented in Table 1. In each of the four job clusters, the highest multiple correlations were observed for the prediction of job knowledge, with R ranging from .37 to .57, $p < .05$. Ability explained the largest amount of variance in job knowledge scores. Generally, the full model of individual differences accounted for more variance in the performance measures than was explained by the environmental model. However, for both the Operations and Skilled Technical job clusters, higher multiple correlations (Rs = .32 and .26, respectively) were obtained for the prediction of task proficiency from environmental models as compared with the individual differences model (R = .14 and .20, respectively).

In the clerical and combat jobs, soldier ability and temperament characterized by Dependability accounted for the most variance in the performance criteria. For the Operations and Skilled Technical MOS, the temperament factors (particularly Dependability and Achievement) explained significant variability in the rating measures, and soldier ability tended to account for significant variance in the more objective performance measures. The environmental model accounted for 3-10% of the variability in criterion measures for the separate job clusters. The largest standardized regression coefficients were observed for the prediction of ratings from Supervisor Support and Job/Task Importance factors. Training had a strong main effect for the prediction of task proficiency and job knowledge measures for the MOS clusters. Further, those variables with the largest standardized beta coefficients in the separate individual differences and environmental models were retained in the full model of main effects for the clusters.

Table 2 shows that ability X Job/Task Importance interaction effects tended to be significant across MOS clusters (except for Operations) and performance measures (except for hands-on). The majority of interaction effects were concentrated between individual differences related to soldier temperament and work environment constructs. Specifically, temperament factors related to Dependability and Adjustment interacted with soldier perceptions of Job/Task Importance, level of Supervisor Support, and available

Table 1

Standardized Regression Coefficients in the Multiple Regression Models for the MOS Clusters

Clerical	1	2	3	4	5	6	7	8	9	R^2	Adj. R^2	R
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.12*	.19**	.14**	.11*						.09	.08	.30
Discipline	.12*	.06	.17**	.12*						.07	.06	.26
Hands-On	.34**	0	.16**	.08						.16	.15	.40
Job Knowledge	.51**	.02	.18**	.06						.32	.31	.57
2. ENVIRONMENT[2]												
Effectiveness					-.06	.27**	-.05	.09	-.05	.07	.06	.26
Discipline					-.03	.25**	-.04	.07	0	.07	.06	.26
Hands-On					-.01	.06	.21**	.06	.01	.06	.05	.24
Job Knowledge					-.02	.07	.11*	.08	.05	.04	.03	.20
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.12*	.15**	.10*	.09	-.04	.25**	-.04	.01	-.06	.13	.11	.36
Discipline	.11*	.02	.14**	.10	-.04	.25**	-.05	.01	-.01	.12	.09	.35
Hands-On	.33**	.01	.14**	.07	-.02	.06	.19**	-.02	-.02	.20	.18	.43
Job Knowledge	.51**	.02	.16**	.06	-.03	.06	.10*	-.02	.01	.33	.32	.57

Combat	1	2	3	4	5	6	7	8	9	R^2	Adj. R^2	R
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.11**	.15**	.21**	.14**						.11	.10	.33
Discipline	.13**	-.01	.33**	.16**						.15	.14	.39
Hands-On	.21**	.08**	-.07*	-.01						.06	.06	.24
Job Knowledge	.48**	.02	.10**	.05						.25	.24	.50
2. ENVIRONMENT[2]												
Effectiveness					-.08*	.17**	0	.09**	.05	.05	.05	.22
Discipline					-.04	.20**	-.05	.11**	0	.06	.06	.24
Hands-On					-.14**	.01	.20**	.03	.01	.05	.04	.22
Job Knowledge					-.12**	.03	.10**	.09**	.03	.03	.02	.17
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.11**	.13**	.17**	.12**	-.07*	.12**	.02	.04	.02	.12	.12	.35
Discipline	.14**	-.03	.29**	.13**	-.06	.14**	-.03	.07*	-.02	.17	.16	.41
Hands-On	.23**	.06*	-.08**	-.02	-.10**	0	.22**	.05	0	.11	.10	.33
Job Knowledge	.50**	0	.08**	.04	-.09**	-.02	.14**	.12**	0	.28	.27	.53

Operations	1	2	3	4	5	6	7	8	9	R^2	Adj. R^2	R
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.04	.13**	.22**	.08**						.07	.07	.26
Discipline	.03	.01	.28**	.11**						.09	.09	.30
Hands-On	.15**	0	.04	-.02						.02	.02	.14
Job Knowledge	.39**	-.05	.09**	.02						.16	.16	.40
2. ENVIRONMENT[2]												
Effectiveness					-.08*	.10*	0	.16**	0	.04	.04	.20
Discipline					-.01	.15**	-.02	.10**	0	.04	.04	.20
Hands-On					-.12**	.05	.20**	.21**	.08**	.10	.09	.32
Job Knowledge					-.15**	-.05	.16**	.14**	.14**	.07	.07	.26
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.03	.10**	.20**	.08*	-.08*	.04	.01	.12**	0	.09	.08	.30
Discipline	.03	-.02	.25**	.09**	-.03	.10**	-.01	.08*	0	.11	.10	.33
Hands-On	.14**	-.04	.02	-.05	-.11**	-.05	.20**	.22**	.09**	.12	.11	.35
Job Knowledge	.38**	-.08**	.09**	.01	-.13**	-.07	.15**	.15**	.12**	.22	.22	.47

Skilled Technical	1	2	3	4	5	6	7	8	9	R^2	Adj. R^2	R
1. INDIVIDUAL DIFFERENCES[1]												
Effectiveness	.05	.22**	.22**	.11**						.11	.11	.33
Discipline	.05	-.02	.28**	.13**						.10	.10	.32
Hands-On	.19**	.03	.05	.04						.04	.04	.20
Job Knowledge	.30**	-.01	.13**	.06						.11	.11	.33
2. ENVIRONMENT[2]												
Effectiveness					-.10**	.16**	.09*	.16**	.05	.10	.09	.32
Discipline					-.09*	21**	.12**	.06	.06	.09	.08	.30
Hands-On					-.12**	-.02	.26**	0	.03	.07	.06	.26
Job Knowledge					-.01	.11*	-.15**	-.04	.05	.03	.02	.17
3. INDIVIDUAL DIFFERENCES + ENVIRONMENT												
Effectiveness	.05	.19**	.17**	.09**	-.09*	.14**	.11**	.08*	.04	.17	.16	.41
Discipline	.05	-.04	.24**	.12**	-.12**	.17**	.13**	.03	.04	.16	15	.40
Hands-On	.19**	.02	.05	.04	-.12**	-.03	.27**	-.01	.02	.11	.10	.33
Job Knowledge	.29**	0	.14**	.06	-.02	.08	-.14**	-.06	.04	.14	.13	.37

Note. [1]Individual differences model contains the set of temperament factors and ability.
[2]The Environment model consists of the five work environment factors.
Variables in the regression models are 1=Ability, 2=Achievement, 3=Dependability, 4=Adjustment, 5=Resources, 6=Supervisor Support, 7=Training, 8=Job/Task Importance, and 9=Cohesion/Peer Support.
* $p < .05$, ** $p < .01$.

Table 2

Summary of Significant Interactions Among Ability, Temperament, and Work Environment in the Prediction of Performance

Interactions	Overall Effectiveness			Personal Discipline		Hands-On				Job Knowledge		
	CL	CO	ST	CL	CO	CL	CO	OP	ST	CL	CO	OP
Ability X Achievement							A					
Ability X Dependability	A				A							
Ability X Adjustment								A				
Ability X Resources								B				
Ability X Support			B									
Ability X Job Importance	A		A	A							B	
Achieve X Support	A											
Achieve X Training							A					
Depend X Resources						B						B
Depend X Support		A										
Depend X Job Importance						B				A		
Depend X Cohesion/Peer Support												B
Adjust X Resources								A				A
Adjust X Support	A							A				
Adjust X Job Importance		B			A							
Adjust X Cohesion/Peer Support									A			

Note. The job clusters are CL = Clerical, CO = Combat, OP = Operations, and ST = Skilled Technical. Significant interaction effects were not found for the rating criteria in the Operations job cluster. Significant interaction effects were not found for the Personal Discipline rating factor and job knowledge test for the Skilled Technical job cluster. Significant interactions are A = $p < .05$; B = $p < .01$.

organizational Resources. Fewer significant interactions were observed between cognitive ability (AFQT) and temperament in the prediction of job performance. Training X Achievement and Cohesion/Peer Support X Adjustment interactions significantly predicted task proficiency in Combat and Operations clusters respectively. Further, for the Operations jobs, several significant interaction effects between soldier perceptions of Resources and individual differences were found to predict maximal performance criteria.

Generally, when designated interactions are added to the full model of main effects, only about 1% of the variance in performance beyond that explained by main effects can be attributed to interactions. However, for the Clerical MOS, interaction effects accounted for an additional 3-7% of the variability in soldier performance, with higher percentages of explained variance associated with the more objective performance criteria.

Discussion

This research examined relationships among individual differences in ability and temperament, perceptions of the Army work environment, and the performance of first term enlisted personnel. Findings revealed that individual differences and environmental perceptions have independent effects on performance in the four job clusters. Some differential effects were found across job clusters with maximal performance (e.g., job knowledge and task proficiency) predicted best from cognitive ability (AFQT) in the Clerical and Combat jobs.

Significant effects for the work environment indicate that both types of typical performance ratings are predicted from the more climate-oriented constructs of Supervisor Support and Job/Task Importance; particularly in the

Combat and Operations clusters. In contrast, soldiers' perceptions of Training and their opportunities to utilize MOS skills, as well as the availability of Resources (e.g., tools and equipment) tended to predict both job knowledge and task proficiency measures for all job groups. Interaction results show that both temperament and work environment factors moderate the relationship between ability and performance. In addition, work environment factors related primarily to Supervisor Support, Resources, and Job/Task Importance, and to a lesser extent Training tended to moderate the relationships between individual temperament factors and performance.

These findings tentatively indicate that job performance is influenced not only by individual differences in ability, but also by the dispositions that soldiers bring to the Army and their perceptions of the environmental context encountered after enlistment, regardless of how jobs are clustered. Further, findings suggest that pre-enlistment differences among soldiers in ability and temperament interact with their environmental perceptions in the prediction of various performance outcomes. Considerable variance in soldier performance can be attributed to the main effects of individual differences and environmental perceptions, and generally significant interactions among these factors explain little meaningful variance.

References

Campbell, J., Hanser, L., & Wise, L. (1986, November). The development of a model of Project A criterion space. Paper presented at the 28th Annual Conference of the Military Testing Association, Mystic, Connecticut.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current armed services vocational aptitude battery (ASVAB) composites. (Technical Report No. 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Olson, D. M., & Borman, W. C. (1986). Development and field tests of the Army Work Environment Questionnaire (Working Paper RS-WP-86-06). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Peters, L. H., & O'Connor, E. J. (1980). Situational and work outcomes: The influences of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Peterson, N., Hough, L., Ashworth, S., & Toquam, J. (1986, November). New predictors of soldier performance. Paper presented at the 28th Annual Conference of the Military Testing Association, Mystic, Connecticut.

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.

Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. Journal of Applied Psychology, 70 (3), 469-480.

CHARACTERISTICS OF COGNITIVE PERFORMANCE IN STRESSFUL ENVIRONMENTS

L.E. Banderet, Ph.D., B.L. Shukitt, B.A., E.A. Crohn, B.A.,
R.L. Burse, Sc.D., D.E. Roberts, Ph.D., and A. Cymerman, Ph.D.
US Army Research Institute Environmental Medicine
Natick, MA 01760-5007

Stressful environments sometimes impair performance with critical consequences in work situations. Military operations, space missions, mountain-climbing expeditions, and rescue efforts illustrate performances with life and death consequences in which it would be useful to know the characteristics of such impairments. This paper examines data from six studies conducted with similar psychometric instruments and procedures under different environmental stressors.

METHOD

Subjects

A total of 87 men served as fully-informed research volunteers in the six studies. Eighty were military personnel and seven were civilians.

Assessment Metrics

Cognitive performance was assessed with nine tasks. The Computer Interaction, Tower, and Map Compass tasks were developed in our laboratory (Banderet, Benson, MacDougall, Kennedy, & Smith, 1984; Jobe & Banderet, 1984); the other six tasks were adapted from the Navy's Performance Evaluation Tests for Environmental Research (PETER) Program (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984; Carter & Sbisa, 1982). All tasks were generated by computer and printed, off-line, on a laser copier. Each task had 15 alternate forms for repeated assessment. Task descriptions and sample items are found elsewhere (Banderet, Lieberman et al., 1986; Banderet, MacDougall et al. 1986; Banderet, Shukitt, Kennedy, Houston, & Bittner, in review).

Procedure

Experimental conditions, number of subjects, and scheduled times for cognitive assessment for the six studies were as shown in Table 1. Except for the Dehydration Study, all were repeated-measures experiments. The Inspired Air, Operation Everest II, and Tyrosine studies investigated high altitude exposure in a hypobaric chamber. The Dehydration studies investigated the effects of hypovolemia, with and without cold; the Atropine Study, the effects of varied doses of atropine (vs. placebo) in a hot-dry environment.

Testing procedures and methods in all studies were similar to those for the PETER Program (Bittner et al., 1984; Jobe & Banderet, 1984). Initially, subjects were trained and given extensive practice with performance feedback. To insure performance was stable and near-maximal, each task was completed 12-18 times before experimentation. The Tower, Computer Interaction, and Map Compass tasks were given typically for 5-6 min; all other tasks, for 3-4 min. The actual duration and number of practice administrations for each task were described in the publications cited in Table 1.

Table 1. Conditions for studying the effects of environmental stressors upon cognitive performance.

STUDY	N	CONDITIONS	ELAPSED TIME OF REPORTED MEASURES	REFERENCES
INSPIRED AIR	23	4600 M 23 °C + 20% RH	1 OR 6, 14 OR 19, 24 OR 29, 38 OR 43 H	BANDERET & BURSE, 1984
ATROPINE	7	2 MG ATROPINE 40 °C + 20% RH	2.0 TO 2.5 H	BANDERET & JOBE, 1984
COLD & DEHYDRATION	36	-24 °C + 4 MPH WINDS RESTRICTED FLUID INTAKE	50 & 54 H	BANDERET, MACDOUGALL, ROBERTS, TAPPAN, JACEY, & GRAY, 1986
DEHYDRATION	18[1]	2% DEHYDRATION (BODY WEIGHT) 20 TO 27 °C	9 H	BANDERET, MACDOUGALL, ROBERTS, TAPPAN, JACEY, & GRAY, 1986
OPERATION EVEREST II	7	4600, 5500, 6400, 7600, 600, 600 M (23 °C + 75% RH)	8, 15, 24, 31, 39, & 41 DAYS	BANDERET, SHUKITT, KENNEDY, HOUSTON, & BITTNER (IN PRESS)
TYROSINE EVALUATION	24	4700 M + 15 °C (50% RH) PLACEBO	1.0 TO 4.5 H	BANDERET, LIEBERMAN, FRANCESCONI, SHUKITT, GOLDMAN, SCHNAKENBERG, RAUCH, ROCK, & MEADORS, 1986

NOTE: THE PREDOMINATE STRESSOR IN EACH STUDY IS LISTED FIRST IN THE CONDITIONS COLUMN.

[1] THESE SUBJECTS WERE ALSO IN THE COLD AND DEHYDRATION STUDY.

OUTPUT (number of problems attempted per minute) and ERRORS (number of problems wrong per minute) were determined for each task. On tasks with limited response alternatives, ERRORS were weighted to discourage careless responses. A third measure, CORRECT, was calculated to reflect both problem-solving and error rates (CORRECT = OUTPUT - ERRORS). Thus, CORRECT incorporated the weighting for errors.

Product-moment correlation coefficients were calculated to explore the relationship between cognitive performance in stressful conditions and baseline performance. Specifically, changes in CORRECT and baseline performances were analyzed. Analysis of Variance and one-tailed Student's t statistics were computed. Significance levels were $p \leq 0.05$.

RESULTS

Cognitive performance was found to be sensitive to a variety of stressful conditions. OUTPUT and ERRORS for all stressors investigated are shown in Figure 1. CORRECT, the measure influenced by both OUTPUT and ERROR rates, is not shown; however, it decreased significantly from baseline in all studies with the exceptions of Grammatical Reasoning (Dehydration, Cold, and Atropine Studies) and Pattern Comparison (Atropine). All nine tasks were not used in each study; bars are shown for those that were. Slower problem-solving rates were usually responsible for the observed performance impairments. Compared to such changes, ERRORS contributed little.

At 4600 meters altitude, seven cognitive tasks were significantly impaired 1 or 6 hours after ascent, as demonstrated by the CORRECT measure of performance. Figure 2 shows the performance tradeoffs associated with these impairments. After ascent, OUTPUT decreased and ERRORS increased from baseline values. At 14 or 19 hours ERRORS decreased significantly on all seven tasks from the earlier altitude values. On two tasks, Pattern Comparison and Number Comparison, OUTPUT was slower than that observed after ascent. This suggests that the performance impairments were different after

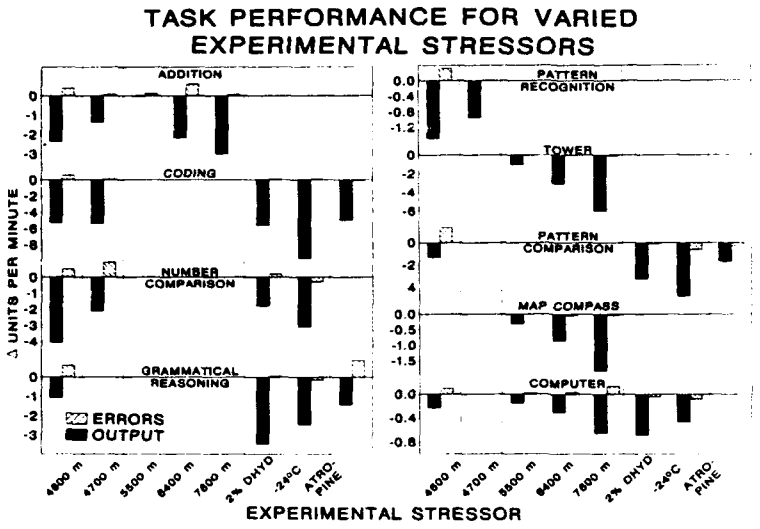


Figure 1. Changes in task output and errors on nine cognitive tasks for varied stressors.

subjects had been at altitude 14 or 19 hours than they were soon after ascent. Grammatical Reasoning, Pattern Recognition, and Computer Interaction task performances returned to baseline levels (CORRECT) as soon as 14 or 19 hours after ascent; performance on all tasks eventually recovered after 38 or 43 hours, except for Number Comparison and Addition. ERROR and OUTPUT changes were greatly decreased after 24 or 29 hours since performance on many tasks had recovered. In the Tyrosine Study (Banderet, Lieberman, et al., 1986) we did not find the same dramatic increase in errors as we found in the Inspired Air Study after ascent to a similar altitude.

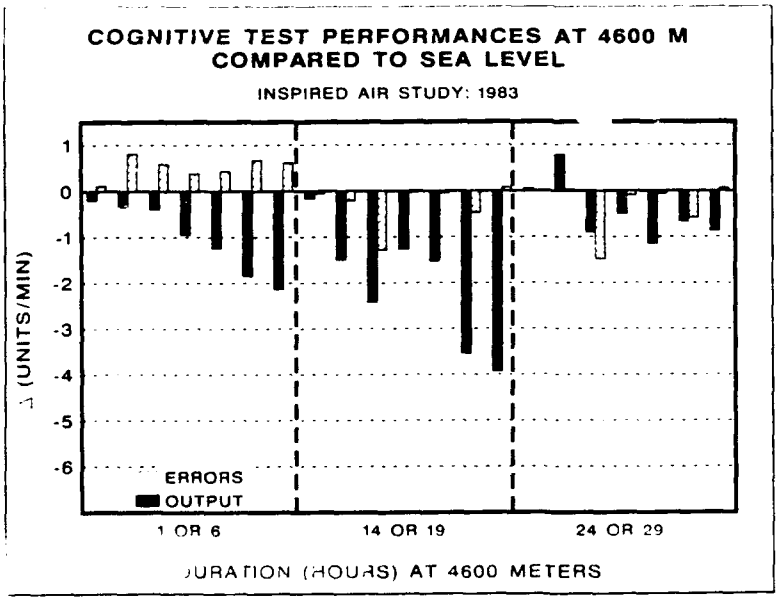


Figure 2. Changes in OUTPUT and ERRORS for varied durations after ascent to 4600 meters simulated altitude. Displayed values are changes from baseline performances for the first test administration in each time interval. For each duration, data (left to right) are for the Computer Interaction, Grammatical Reasoning, Pattern Comparison, Pattern Recognition, Addition, Number Comparison, and Coding Tasks.

Similar data are shown in Figure 3 for the Operation Everest II Study. Performance impairments resulted from a slowing of OUTPUT. Higher altitudes produced more slowing of OUTPUT; return to sea level did not eliminate these

impairments. Increased ERRORS were observed, but they accounted for less than 50% of the performance impairment, even at extreme altitudes (above 6000 m).

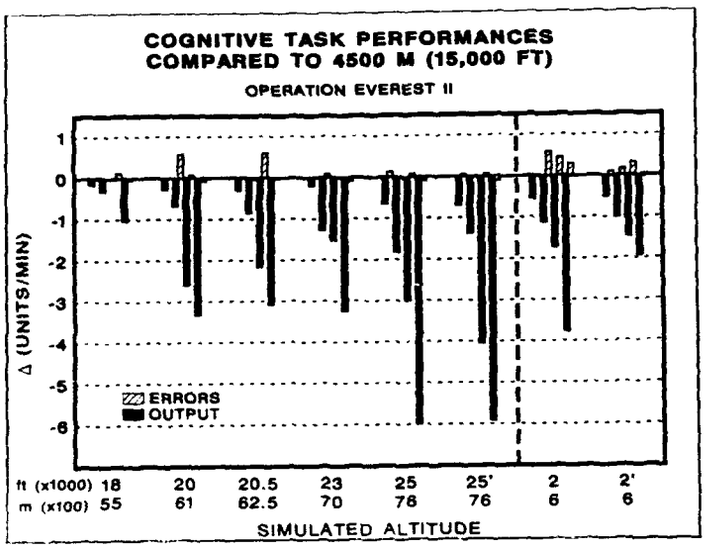


Figure 3. Changes in OUTPUT and ERRORS for varied altitudes and following return to sea level. Displayed values are changes from 4500 meter performances. Data (left to right) are for the Computer Interaction, Map Compass, Addition, and Tower Tasks at each altitude.

Table 2 shows the correlation coefficients calculated for cognitive performance in a stressful condition (change in CORRECT from baseline) with baseline performance for the various stressors. Statistically significant relationships are indicated with asterisks. Negative correlations resulted in all instances, 15% being statistically significant. In these instances, usually involving spatial tasks, decreases in CORRECT during the stressful challenge were negatively associated with baseline performances. That is, subjects with the largest impairments during the challenge were those with the highest baselines. This was observed on Tower (hypoxia), Pattern Comparison (hypoxia or dehydration and cold), and Computer Interaction (hypoxia). This relationship was not observed with Pattern Recognition or the verbal tasks.

COGNITIVE TASK	INSPIRED AIR N = 23	2% DEHYD-RATION N = 18	2% DEHYD. & -24°C N = 18	-24°C & WIND N = 18	4700 M & 15°C N = 23
ADDITION	-0.37	—	—	—	-0.16
CODING	-0.31	-0.12	-0.28	-0.41	-0.01
NUMBER COMPARISON	-0.38	-0.13	-0.14	-0.17	-0.06
GRAMMATICAL REASONING	-0.26	-0.37	-0.02	-0.37	—
PATTERN RECOGNITION	-0.26	—	—	—	-0.37
TOWER TASK	—	—	—	—	-0.40*
PATTERN COMPARISON	-0.56 **	-0.37	-0.57**	-0.25	—
MAP COMPASS	—	—	—	—	-0.04
COMPUTER	-0.33	-0.20	-0.51*	-0.41	—

Table 2. Correlation coefficients are shown for changes in cognitive performance under stressful conditions and baseline performance. Not all tasks were used in all studies; coefficients are shown for those that were. An asterisk by a correlation coefficient indicates a statistically significant relationsnip.

DISCUSSION

Impairments in cognitive performance on most tasks were found for all of our environmental stressors. Performance was impaired on all tasks at altitudes from 4200 - 7600 m, at least initially. Impairments in performance at altitude resulted from decreased OUTPUT rather than increased ERRORS. This latter finding also generalized to the other stressors of dehydration, cold, and atropine with heat. It was also robust since our error adjustment exaggerated ERRORS.

Increased errors were more prevalent early in the exposure to 4600 meters altitude than they were later. We suspect such increased errors resulted because of the variability of the euphoric and somewhat disruptive behavior induced by altitude the first few hours after ascent. In the Operation Everest II Study, we continued to see substantial impairments when subjects descended to sea level after several days' exposure to 7600 meters. This probably resulted because physiological adaptations to very high altitudes by the pulmonary, renal, and circulatory systems may require days to reestablish normal physiological values.

Subject baselines differed as much as 10-fold on some tasks. We evaluated data from these tasks to determine if subjects who responded rapidly in baseline conditions were more likely to exhibit decrements when tested under stressful conditions. Performance decreases during stress were negatively correlated with baseline performance on spatial tasks that encouraged "intuitive" judgments. A simple psychomotor task, Computer Interaction, also showed this relationship in the Dehydration and Cold Study, which suggests high baseline values were difficult to sustain while subjects were cold and hypovolemic (conditions which impair blood flow to the fingers).

Cognitive performance deteriorated to some degree under all environmental stressors evaluated. That such impairments occur with relatively simple and highly overlearned tasks suggests the vulnerability of even simple performance when task demands require rapid responses. It is clear that in these stressful environments impairments resulted from a slowing of performance rather than increased error rates.

SUMMARY

Characteristics of performance impairments, e.g. output-accuracy tradeoffs and individual performance styles, are not always described or even investigated in performance studies. This effort explored cognitive performance decrements from six studies of environmental stressors such as heat and atropine, dehydration, cold, and high altitude. Cognitive performance was assessed using tasks based upon the Performance Evaluation Tests for Environmental Research (PETER) methodology.

Performance impairments were encountered for all stressors on most tasks but some recovered with continued exposure. Impairments were due to a slowing of performance rather than increased errors. Performance in stressful environments was negatively correlated with baseline performances on spatial tasks requiring intuition. On such tasks, subjects with impaired performance during the stressor had higher baselines before the challenge. These findings imply the need for training or job design considerations for reducing the impact of adverse environments upon work tasks.

REFERENCES

Banderet, L.E., K.P. Benson, D.M. MacDougall, R.S. Kennedy, & M. Smith. (1984). Development of cognitive tests for repeated performance assessment. Proceedings of the 26th Annual Meeting Military Testing Association, Munich, Federal Republic of Germany, 375-380.

Banderet, L.E., & R.L. Burse. (1984, August). Cognitive performance at 4600 meters simulated altitude. Paper presented American Psychological Association, Toronto, Canada.

Banderet, L.E., & J.B. Jobe. (1984). Effects of atropine upon cognitive performance and subjective variables (Technical Report No. T15/85). Natick, MA: U.S. Army Research Institute of Environmental Medicine.

Banderet, L.E., H.R. Lieberman, R.P. Francesconi, B.L. Shukitt, R.F. Goldman, D.D. Schnakenberg, T.M. Rauch, P.B. Rock, & G.F. Meadors III. (1986, in press). Development of a paradigm to assess nutritive and biochemical substances in humans: A preliminary report on the effects of tyrosine upon altitude- and cold-induced stress responses. The Biochemical Enhancement of Performance: Proceedings of a Symposium, Lisbon, Portugal.

Banderet, L.E., D.M. MacDougall, D.E. Roberts, D. Tappan, M. Jacey, & P. Gray. (1986). Effects of hypohydration or cold exposure and restricted fluid intake on cognitive performance. Predicting Decrements in Military Performance Due to Inadequate Nutrition: Proceedings of a Workshop, Washington, DC: National Academy Press, 69-79.

Banderet, L.E., B.L. Shukitt, R.S. Kennedy, C.S. Houston, & A.C. Bittner, Jr. Cognitive performance and affective responses during a prolonged ascent to 7600 m (25,000 ft) simulated altitude. (Submitted for review).

Bittner, A.C. Jr., R.C. Carter, R.S. Kennedy, M.M. Harbeson, & M. Krause. (1984). Performance evaluation tests for environmental research: Evaluation of 112 measures (Report NBDL84R006 or NTIS AD152317). New Orleans, LA: Naval Biodynamics Laboratory.

Carter, R.C., & n. Sbisa. (1982). Human performance tests for repeated measurements: Alternate forms of eight tests by computer (Report NBDL82I3003). New Orleans, LA: Naval Biodynamics Laboratory.

Jobe, J.B., & L.E. Banderet. (1984). Cognitive testing in military performance research. Proceedings of a Workshop on Cognitive Testing Methodologies, Washington, DC: National Academy Press, 181-193.

ADDENDUM

Human subjects participated in these studies after giving their free and informed voluntary consent. Investigators adhered to AR 70-25 and USAMRDC Regulation 70-25 on Use of Volunteers in Research. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

SIMULATION BASED TESTING: A new Approach in Selecting ATC-Applicants

Juergen H. Haettig
STREITKRAEFTEAMT, Dezernat Wehrpsychologie

In the last few years much attention was spent to the problem of selecting personnel for duties with high mental work-load, because traditional tests fail in predicting success. A typical example for this problem is the selection of air-traffic-control (atc) applicants. A statistical analysis of the wastage-rate of military atc-students reveals that about fifty per cent of the novices did not meet the requirements of the basic atc-training. A detailed analysis of the data of the entrance-ecamination-tests, the data of the ATC-School and the final examination tests revealed, that the selection test battery in use can predict academic achievement but fails in predicting job-related achievement.
Until yet we use paper-pencil-tests which measure verbal abilities, numerical abilities, spatial orientation and learning abilities. Maybe that these are neccessary abilities to become a well performing air-traffic-controller, but they are not sufficient. We think that additional requirements are problem-solving-techniques, stress-resistance and Gestalt-perception (as a cue: WITKINs field-dependency; HOPKIN, 1982). One of the most important ability may be the ability to control a situation which changes continously.

SPOHRER (1984) has developed a job-related test, the "Approach Control Test" which simulates time-dependent situations asking for decision-making under time stress. The applicant must guide two or more planes to a landing position in seven seconds cycles. In his work he was able to show that this test gives additional information for selection purposes. As a pilot project we had tried to adapt the ATC as a computer-administered-test. But very soon it was obvious that a paper-pencil test cannot be transposed to a computer-test without changing the psychological implications. So we have designed a new test which includes the control area, that is the net of ats-routes (i.e. air-traffic-service-routes), from the SPOHRER-Test.

I would like to give you a brief sketch of this new test:
The test consists of two parts, a learning part and a test part. In the learning part, the subject has to learn the names of report points and the connection between the report points, the ats-routes. The test part is started, when the subject can reproduce all the report points. In this way we can control knowledge acquisition and preknowledge.
When a task is presented the subject is informed about fuel, speed, runway-in-use and first report point in the control area. The subject has enough time to make a plan of the flight-move-

ments, because the task is started by the subject itself. When the task is running, it is time-triggered (because a plane cannot stop in the air). The report points are listed as a menue table. If the command is correct, the name of the next report point is inverted.

While the test is running, the applicant cannot see the names of the report points and the ats-routes, but he can call names and routes by a help function.
The plane is decending continously 1000 feet from reportpoint to reportpoint. The flightlevel in landing position must be 10 (1000 feet).
There is one plane to guide.
When the control-command is wrong or too late or when there is no control-command at all the plane is searching the next report point by a fixed rule set.
The number of flight-movements is restricted by fuel. The time to think about the next destination is limited by the speed of the plane.

The test was administered to 24 military atc-applicants with a mean age of 22 years, who took part at an entrance-examination test of the civil aviation administration. In this test 25 tasks are presented. A sitting lasts fifty minutes on the average.

As parameters to describe the subject's atc-test behaviour we defined following measures:

- Efficiency of Control — that is the number of correct commands in relation to the number of flight-movements;
- Accuracy of Control — that is the number of correct commands in relation to the total number of commands;
- Time of Decision-Making — that is the time between task-presentation and starting the task;
- Stability of Planning — that is the autocorrelation of the "time of decision-making".

RESULTS

Item-difficulty

Item-difficulty is an empirical measure of traditional psychometric tests. In our definition efficiency of control is related to item-difficulty. We had defined a theoretical item-difficulty which reflects speed, limitation of the number of movements by fuel and number of operational report-points.
Picture 1 shows the results of the comparison of theoretical defined item-difficulty and empirical item-difficulty. There is no correlation between empirical and theoretical item difficulty with the excention of the very difficult items 14 up to 17. The theoretically defined difficulty of this items depends on the speed of the plane. In this tasks a decision must been made up in the time between 3.2 seconds (item 17) and 6.3 seconds (item 14). In the difficulty-class 3 and 4 there is an increase of the empirical item-difficulty dependent on the serial

position: an effect of acquisition of system-management capabilities.

Efficiency of Control and Accuracy of Control

This measure reflects control over the system. Over all tasks there is no significant difference between applicants who failed and applicants who passed the entrance-examination-test. (Pic. 2 and 3) Eleminating items 14 up to 17 we did an ordinary least square (OLS) approximation to find the initial system-management capability and the increase by learning (see Tab. 1). Successful applicants had a better initial system-management-capability but less increase in learning. Therefore at last both groups reach the same management capability. Notice the increase of learning: successful applicants had a continous learning-increase, applicants who failed had a discontinous increase. Therefore the approximation (expressed in mean deviation from the function) is better for applicants who passed.

Table 1: Parameters of the OLS-Approximations

	Alpha	Beta	Mean Deviation	
Efficiency				
	54.44	0.60	6.86	passed
	42.43	1.00	12.44	failed
Accuracy				
	67.33	0.51	6.99	passed
	56.06	0.73	13.12	failed

Decision-Making Time and Stability of Planning

The subjects started the tasks themself. This feature was implemented to see how prior experience of the subject would influence the planning of a new task. Picture 4 shows the mean time of decision-making for both groups.
Successful applicants are more consistent in planning than failing applicants. This fact is supported by the results of the autocorrelation of the mean time of decision-making (Pic. 5). There are significant differences in lag one and two: high correlation ($r1=0.8$; $r2=0.64$) in the group of successful applicants, zero correlation in the group of the unsuccessful applicants. Applicants, who passed the examination, are planning their guidance similar to the last two tasks. Applicants, who failed, varied time of planning in an unsystematical manner.

DISCUSSION

This experiment was performed to get experience with simulation based tests. We are interested in selection of personnel for duties with high mental workload. Thus we compared the system-management behaviour of applicants who passed an elaborated entrance examination test with the behaviour of applicants who failed. Over all there are some findings of interest.

First we saw that successful applicants had a better initial management capability but a slow increase in learning. In contradiction applicants who did not pass the over-all-assessment of the civil aviation administration started with a low value but had a faster learning function. It must be pointed out that the learning function of loosers has steps. This is in accordance with findings of KLUWE et al. (1986). In their research about learning in complex systems they found, that learning is not incremental but a leap-function. Because in traditional testing (and I think CAT is traditional testing,too) there is only one value,which reflects the number of correct responses, the capability of system-management and the progress in learning to manage a complex system cannot be measured by traditional tests. Maybe the test stops a few items before a sudden increase of the performance will occure.
Second we found, that successful applicants were continous in planning the guidance of the plane. There was a variation, but this variation was a smooth one. Applicants, who failed, varied in an unsystematical way. DÖRNER et al. (1983) found that good problem-solvers are more consistent in information seeking, information acquisition and information evaluation than bad one. Maybe this is the influence of a personality trait, because bad problem-solvers tended to neurotic reactions. In situations with unsufficient circumstances they often started crash-programs.

In this atc-simulation-based test efficiency and accuracy of control were limited by psychomotoric processes. We had choosen the "mouse" as input-medium but most of the applicants had problems with the mouse, especially then they were nervous. When the speed of a plane was very high (tasks 14 to 17) there was no real chance to be conscious of the command given. Therefore the difficulty of this items refers to psychomotoric rather than to cognitive difficulty.

Simulation based testing is a new approach in selecting special personnel because this kind of tests can evaluate learning behaviour, system-management-performance and not at least the influence of personality traits on performance. In fact there remains a serious and so far unsolved problem: to find an appropriate psychometric measure for selection purposes.

References

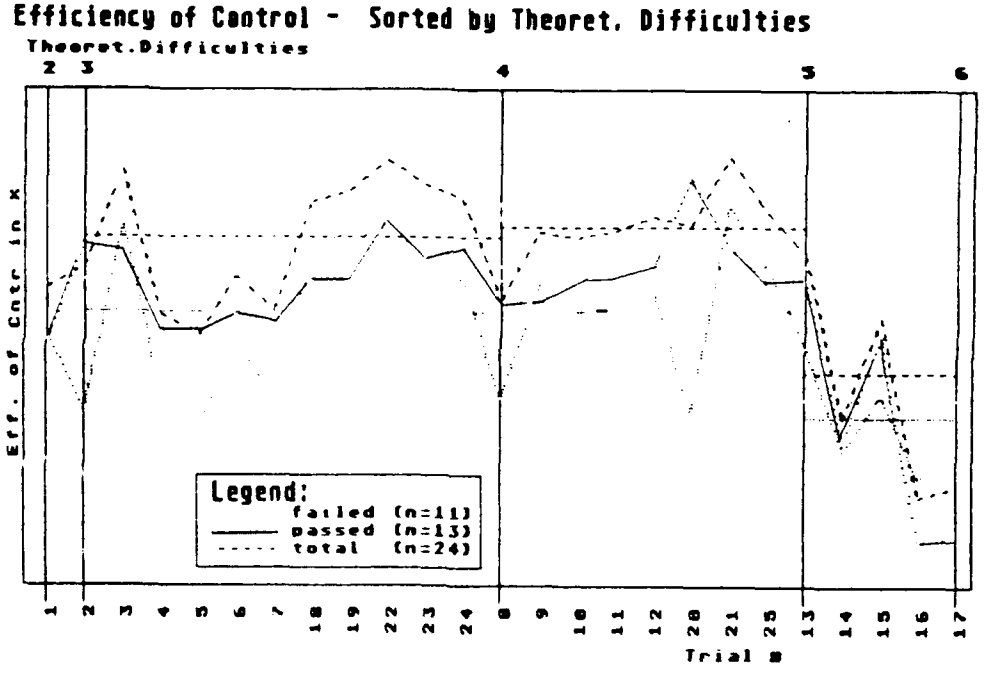
DÖRNER, D., KREUZIG, H.W., REITHER, W., STÄUDEL, Th. (eds.), 1983. Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität.
Bern: Huber.

HOPKIN, V.D., 1982. Human Factors in Air-Traffic-Control.
Neuilly Sur Seine: AGARD - AGARDograph No. 275

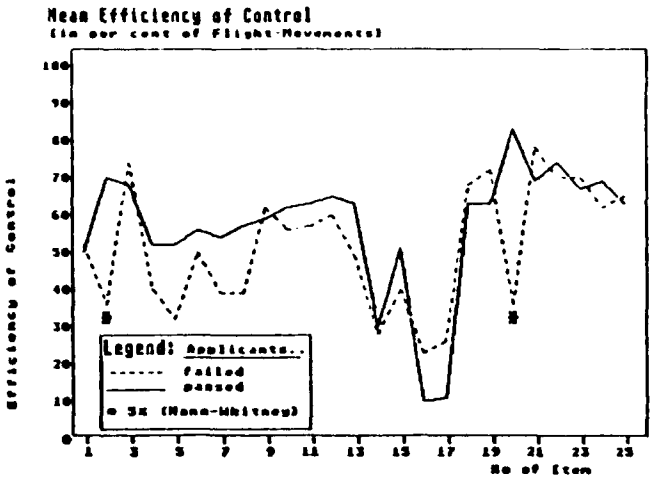
KLUWE, R., MISIAK, C., RINGELBAND, O., HEIDER, H. 1986. Lernen durch Tun. Eine Methode zur Konstruktion von simulierten Systemen mit speziellen Eigenschaften und Ergebnisse einer Einzelfallstudie. In: M. Amelang (ed.). Bericht über

den 35. Kongreß der Deutschen Gesellschaft für Psychologie in Heidelberg 1986. Band 1.
Göttingen-Toronto-Zürich: Hogrefe

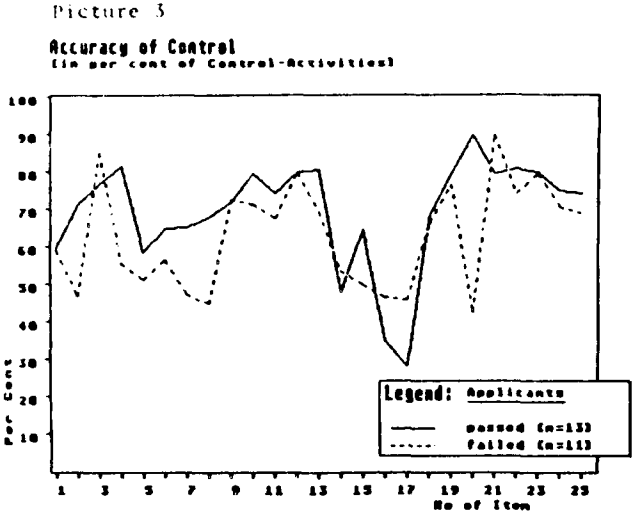
SPOHRER, Th., 1984. Konstruktion eines tätigkeitsspezifischen Tests zur Auswahl von Fluglotsenbewerbern.
Hamburg: DFVLR - FB-84-81



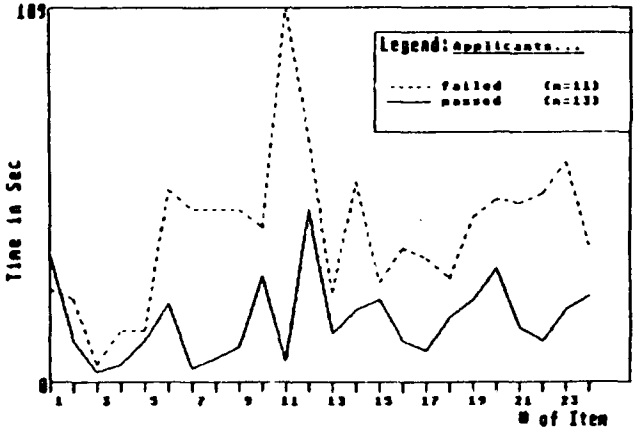
Picture 2



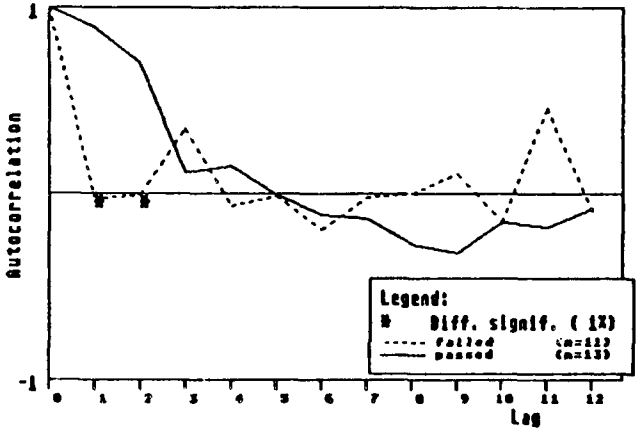
Picture 3



Picture 4



Picture 5



THE APPLICATION OF A HUMAN FACTORS DATABASE IN ARMY AIRCRAFT ACCIDENT AND INCIDENT INVESTIGATION

A J W Feggetter, Human Factors Unit, HQ DAAC, Middle Wallop,
H M McIntyre, Stockbridge, Hants,
L Mortenson, UK
B Pritchard TRC, Brackley, Northants, UK

INTRODUCTION

1. Background. The Army aircraft accident and incident rate fluctuates from year to year, however the percentage attributable to human error has remained at about 76% over the last 10 years. Associated with these catastrophic events is an increasing cost in terms of the hardware and in terms of human life and limb. In 1984 the Human Factors Unit at the Headquarters Directorate Army Air Corps (HFU HQ DAAC) set up a human factors database to assist in Army aircraft accident and incident investigation. The aim was to assist in understanding the underlying reasons for a particular accident in order to prevent future similar ones. As a long term aim such studies contribute towards the development of a theory of human error. Thus the overall accident rate may be improved and the proportion of accidents attributed to human error reduced. A full description of the database is given in Feggetter and McIntyre 1984.

2. Theoretical Underpinning Research into accidents, including those in aviation, (McFarland 1953, Wansbeek 1969, Rolfe 1972, Allnutt 1976.), nuclear power (Kemeney 1979, Reason 1986c.) and maritime transport shows that accidents are rarely due to a single cause but rather to a host of interacting contributary factors. Existing theories of human error were considered. The authors designed the database so that it adhered to no one theory but rather took into account many which have been proposed, thus enabling them to be tested thoroughly in an applied setting. The database consists of information from selection, training, accident investigations and normative studies. This structure is shown in figure 1.

AIM

3. The aim of this paper is to show the value of a database in terms of the effective and efficient management of individual aircrew. Studies are described which compare selection and training data of accident and non accident aircrew. They show how strengths and weaknesses already highlighted during selection and training may be pertinent to individual accidents.

DESCRIPTION OF DATA (See Table 1)

4. Psychomotor data from the Biggin Hill pilot selection tests were obtained for all personnel. They consist of scores on three sensory motor tests (Bartram et al 1984).

5. Selection group * This group consisted of 246 entrants to the Army Pilots Course all of whom had taken the Catell 16 PF. 233 had completed the Eysenck Personality Inventory forms A and B (EPI.) *On data collected by Hull University 1982/1983. Normative group This group consisted of 13 sq dron pilots who

had not at the time of testing had a flying accident. In addition to the psychomotor tests they undertook an Eysenck Personality Questionnaire (EPQ). Data were also obtained from their training records. This group also answered, as far as possible, the same questions which would be put to aircrew during a human factors accident investigation. Accident group This group consisted of 10 pilots who had been involved in flying accidents. In addition to the psycho-motor tests data were obtained from their training records and from the human factors accident investigation.

Table 1 Data Sources

	psychomotor			Personality		Training	Database
	INSB	SMA	CVT	16PF	EPI		
Selection	*	*	*	*	*	-	-
Accident	*	*	*	-	-	*	*
Normative	*	*	*	-	*	*	*

STUDY 1 AN EXAMINATION OF THE PSYCHOMOTOR TESTS FOR EACH GROUP

6. Aim The aim of the study was to determine whether the normative group and the accident group were representative of the general population of AAC at the selection stage.

7. Method. The means and standard deviations were calcuated for each test. A T test for independant samples was applied.

8. Results No significant difference was found amongst the three sample groups.

9. Conclusion It can be seen that the groups are hommogeneous

STUDY 2 AN EXAMINATION OF PERSONALITY FACTORS

10. Background Personality has been cited in 76% of human error accidents (Feggetter and McIntyre 1986).

11. Aim To determine the personality profile of the typical AAC pilot.

12. Method The 16 PF results for the selection group were analysed.

13. Results The typical person who passes the selection test is of average intelligence, emotionally stable, fairly dominant, happy go lucky, imaginative, conscientious, venturesome, tough minded, trusting, fairly controlled and relaxed. Those who pass the course and become pilots tend to be less happy go lucky, less venturesome, less trusting and more controlled. It is interesting to note that those who fail during the final phase of the training course have a profile which is very similar to that of those who pass. The points of difference are that those who fail are less conscientious, less tough minded, less self assured and less well controlled. Officer pilots tend to be at the

extremes on many factors. They are the most venturesome and the most happy go lucky, more self assured and more controlled than other groups.

14. Conclusion A personality profile of the "typical" pilot, as he presents himself during the selection process has been established. The 16 PF is a self rating test and open to faking. It appears that the profile presented by the AAC entrants is very similar to that found in studies on airline pilots. It may be that pilots as a group are under strong social pressure to conform to a stereotype. It was therefore considered that a more reliable method of obtaining information would be a content analysis of the training records.

STUDY 3 AN EXAMINATION OF TRAINING RECORDS

15. Background During the initial flying training each sortie is written up on a "sortie report" form. Space is provided for the instructor's comments and each sorties is given an overall grade ranging from green which is a pass, through brown which is a marginal pass to red which is a fail. At the end of each fortnight a summary report of progress is made.

16. Aim The aim of this study was to compare the training records of the accident group and the normative group. Firstly the colour codes were examined and secondly the instructors comments were reviewed. It was hypothesised that there would be a difference between the two groups with the accident group experiencing more difficulties, personality featuring prominently amongst them. It was anticipated that this difference might be reflected in the contributary causes of accidents.

17. Method For the purpose of this study selected samples from both the normative and accident groups were used. It was essential that the records included in the samples should be complete and contain no missing data. Within this limitation the samples were matched so that they covered approximately the same time period, 1978 to 1984. A sub group of training records for each person was formed. This comprised each fortnightly report and any sortie report in that time which had a handwritten comment reflecting either well or badly on the sortie. Each brown sortie report and each red sortie report were also included. The hand written comments in the sub group were then coded in terms of 16 PF, EPQ and psychomotor abilities.

18 Results The gross frequency count showed that there was a significant difference between the two samples in terms of the red (fail coding) ($Chi2 = 5.33$, $p < 0.25$). The accident group had significantly more sorties coded red. The difference between the two was most significant during the early stages of training. Significant differences were found between the two groups for overconfidence, underconfidence and hand eye coordination. There are also differences in spatial ability, accuracy, tenseness and tendency to be self critical See Table 2

19. Conclusion The training records give useful information on the individuals in terms of personality and skills highlighting

strengths and weaknesses. These characteristics have been assessed on an almost daily basis during a 10 month period. It would seem likely that they are reliable assessments of the student. It would appear from the analysis that those individuals described as overconfident or as having difficulties with hand/eye coordination, who are relaxed and not very self critical tend to be those who have accidents. The results suggest that an individual's weaknesses may contribute to the accident and may have been already identified during training.

Table 2 Analysis of personality and skill factors

Percentage of total sorties receiving adverse comments shown

CODE	ACCDNT	NORM	CODE	ACCDNT	NORM
Planning	10.20%	6.47%	Spatial ability	4.05%	2.22%
Hand/eye	22.70%	10.32%	Self critical	0.27%	1.82%
Overconfident	10.20	4.85%	Underconfident	4.86%	23.07%

ACCDNT - Accident group (N=10) NORM - Normative group (N=13)

STUDY 4 AN ANALYSIS OF THE HUMAN ERROR ACCIDENTS

20. Aim To identify the major contributary causal factors in the human error accidents and to relate them to the personality or performance factors identified during training.

21. Method The contributary cause factors for the accident were derived from the database. These are shown in table 3. Table 4 illustrates the major characteristics of the pilots in the accident group as assessed by their training records.

22. Results Overconfidence was cited as a contributary cause in 5 of the 10 accidents. In 4 of these overconfidence was highlighted in the individual's training records. In one case such comments constituted 60% of all the comments made during his training. 4 of the accidents involved inexperienced pilots who were inadequately supervised. In accident A inadequate planning was cited as a cause. This weakness was picked up during training comprising 33% of all comments made. It is interesting to note that the 3 individuals whose casual attitude was felt to have in part caused their accident had the highest proportion of adverse comments made about accuracy.

23. Conclusions This study is descriptive and the sample size small, however the results highlight trends worthy of further consideration. It would appear that there is a relationship between causal factors of an individual's accident and his training record. At present it is not possible to predict from training records whether or not an accident will occur. When those individuals do experience an accident factors identified as weaknesses during training seem also to be contributary factors to the accident. This suggests that these may be necessary conditions for the accident rather than sufficient.

OVERALL DISCUSSION AND CONCLUSIONS

24. The majority of the Army Aircraft accidents are attributable to human error. The underlying premise of this research is that there is no single cause of an accident. Previous research has shown that it is the human element which may be the weak link in the chain. Accidents happen when a number of events occur and somehow accumulate to give rise to the ultimate catastrophe.

25. Within the military environment considerable effort and cost goes into the selection and training of the aviator. Information on an individual's skills and personality during selection and during the training is gathered systematically but its full potential is rarely realised.

Table 3 Major Contributary Causes of Accidents in Accident Sample

ACCIDENT	CONTRIBUTARY CAUSE FACTORS
A	Inexperience, Supervision, Planning, Workload, Lack of crew cooperation
B	Inexperience, Supervision, Overconfidence, Distraction Fatigue, Low arousal, End-of-tripitis, Visual cues
C	Inexperience, Supervision, Planning, Get-home-itis
D	Inexperience, Supervision, Overconfidence, Visual cues End-of-tripitis,
E	Slow reaction, Cockpit gradient,
F	Inexperience, Overconfidence, Casual attitude
G	Casual attitude, Distraction, End-of-tripitis, False assumption
H	Overconfidence
I	Casual attitude, Sortie stress, Physiological stress
J	Overconfidence, fatigue, Visual cues, End-of-tripitis

26. Training Decay It is expected that skills will deteriorate from a peak at the end of training. This may coincide with a peak in the accident figures at the 500 flying hours stage. There may also be specific skill defecits associated with problems in training. Skills that are difficult to acquire are more vulnerable to decay. Skills associated with sorties graded red are therefore most at risk. To reduce the likelihood of this leading to an accident supervisors should be aware of the specific skill defecits identified in an individual's training record. A further peak of accidents occurs at the more experienced levels. A system which analyses career experience and continuation training is likely to similarly identify potential high risk pilots.

Table 4 Analysis of Training records for Accident Group

	A	B	C	D	E	F	G	H	I	J
SPATIAL	-	4	5	-	2	-	3	-	-	-
HAND/EYE	-	30	17	-	20	18	20	-	-	4
PLANNING	33	5	8	9	2	-	14	3	7	7
ACCURACY	-	5	17	5	12	54	34	5	20	13
O/CONFID	-	5	-	60	-	9	-	30	18	-
UN/CONFD	-	-	1	-	5	-	6	-	-	11

Figures are percentage of adverse comments found
O/CONFID - Overconfidence UN/CONFID - Underconfidence

STRUCTURE OF THE DATABASE

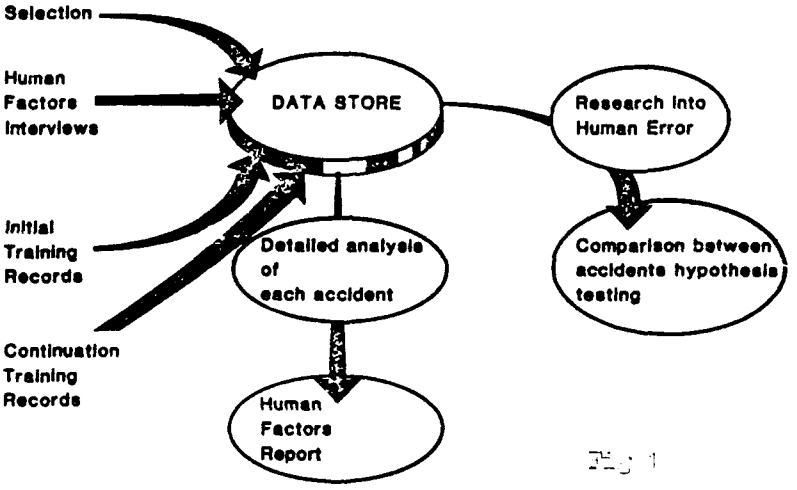


Fig 1

REFERENCES

Feggetter A &McIntyre H (1984) A Microcomputer Based Database Processing System to investigate H F aspects of A/C Acc. and Inc. Proc. WEAAP conf. Helsinki. Allnutt M (1976) H F in Pilot Error (Ed R Hurst) London: Crosby Lockwood Staples. Kemeny J (1979) The Need for Change: The Legacy of TMI. Rep. of Pres.Comm. on Acc. at 3 Mile Island. Washington: Govt Print. Off. McFarland R (1953) H F in Air Trsp. New York: McGraw-Hill. Reason J T (1986c) An Interacticnist View of System Pathology. NATO Adv. Res. Workshop on Failure analysis of Information Systems. Bad Windsheim, Germany. Rolfe J M (1972) Ergonomics and air safety. App. erg.3,75. Wansbeek G C (1969) H F in airline inc. 22nd Annual Int. Air Safety Seminar at Montreux. Bartram et al (1984) MICROPAT, Proc of 26th Annual Conf. of MTA: Munich. Wagenaar W A (1986) The Cause of Impossible Accidents. The 6th Duiker Lecture. Uni. of Amsterdam. Feggetter and McIntyre (1986) The Military Pilots Behavioural Overlay. Proc. of 16th Helico-Baltic Conf. Stockholm.

Job Knowledge Test For Navy and Marine Corps Jet Engine Mechanics

Patricia A. Alba
MAXIMA
8301 Broadway Suite 212
San Antonio, Texas 78209

Herbert George Baker, PhD
Navy Personnel Research and Development Center
San Diego, California 92152-6800

The Department of Defense is currently conducting a Joint-Service Job Performance Measurement Project designed to link performance on the job to the Services' selection and classification processes. Hands-on work sample tests and newly developed surrogate instruments are being used to measure the performance of first term enlistment personnel. Hands-on testing is being developed because it provides a reliable measurement of a set of representative tasks. However, hands-on testing is expensive to administer in terms of time and money and often results in the elimination of tasks due to time constraints or the possibility of personnel injury or equipment damage. In an attempt to save dollars and man hours and to measure the ability to perform all tasks regardless of practicality, less expensive and easier to administer surrogate measuring devices are being developed and compared to the hands-on testing methodology.

The Air Force developed the interview component of Walk Through Performance Testing which is a surrogate conducted one on one at the work site in a show and tell fashion (Hedge 1984). The Navy developed a paper and pencil job knowledge test surrogate which can be administered one on one or in groups of any size. The Navy's Job Knowledge Test is different than most job knowledge tests in that it uses photographs as reference points and requires the examinee to identify components in the photographs and then select, from a list, procedures that would normally be followed when performing various tasks. Both services developed a set of rating forms to be administered to the incumbents, their immediate supervisors and at least one of the incumbents' peers.

In 1985, the Air Force transferred its entire Jet Engine Mechanic Job Performance Measurement System and a Jet Engine Mechanic Job Knowledge Test patterned after the Navy's methodology to the Navy for Navy and Marine Corps use. The system was designed to measure the performance of jet engine mechanics assigned to the J-79 intermediate or organizational maintenance activities (IMA or OMA). The inclusion of the job knowledge test, as noted in Blackhurst and Baker (1985) allowed for a direct comparison of surrogate performance techniques developed by different Services on the same sample of incumbents.

Jet Engine Mechanic Job Knowledge Test Development

Normally, before a test can be developed, it is necessary to select and analyze tasks that are representative of the job domain to be evaluated.

Neither of these steps were required for the development of the Jet Engine Mechanic Job Knowledge Test since the seven common and three IMA or OMA unique tasks were previously selected and analyzed for the interview components of Jet Engine Mechanic Walk Through Performance Testing.

Photographs

Photographs for inclusion in the job knowledge test were taken at Naval Air Station Dallas. The subjects of the photographs were a J-79 jet engine, individual engine components, and materials used to remove or replace parts and sections of the engine. The components were photographed on and off the engine to allow for flexibility when developing the job knowledge test items. Sizing of the photographs was conducted on site by the photographer, under the direction of an active duty jet engine mechanic subject matter expert and the test developer. The team effort was necessary to ensure that each photograph contained the appropriate component or section of engine that the examinee was to identify during the testing situation, as well as, at least two other components to serve as test distractors for each task.

Job Knowledge Test Items

The relevant Navy technical publications were reviewed by the test developer and two subject matter experts assigned as trainers at Marine Corps Air Station Beaufort, South Carolina. A list of all appropriate steps was constructed for each of the ten tasks. False steps were identified to serve as step distractors for each task. The respective step distractors were mixed with the true steps by either replacing a true step or adding the distractor at a logical position in the task's step sequence.

The test items were arranged in the test book with the photograph(s) on the left page and the steps and distractors on the page to the right so the examinee could view the photographs and respond to the questions without being required to turn the page. A short verbal scenario was provided at the top of each right hand page to accompany the respective photo reference point and to set the stage for each task. Instructions indicated that the examinees were to record their responses on an answer sheet provided by the test administrator. Table 1 on the next page shows an example of the right page layout.

Pilot Test

Air Force and Marine Corps subject matter experts reviewed the completed job knowledge test book for accuracy and understandability. A pilot test was conducted at Naval Air Station Dallas, by two contractor test administrators who received extensive administration training and had collected data for the Air Force Jet Engine Mechanic Specialty. The test was administered to three first term jet engine mechanics, one from the organizational activity and two assigned to the shop maintenance areas. The individuals were allowed as much time as necessary to complete the test. All three jet engine mechanics completed the job knowledge test within an hour. Minor editorial changes were made and the job knowledge test was printed and compiled for administration to jet engine mechanics assigned to the intermediate and organizational maintenance activities.

Table 1

Example of Job Knowledge Test Right Page Layout

SCENARIO FOR TASK R 1: You have been instructed to install a pressurizing and drain (P&D) valve on the engine. The main oil cooler, compressor rear frame bracket and lines are already on the engine.

3. (I016) Which component on the picture to the left is the pressurizing and drain (P&D) valve? Write the matching letter on the answer sheet.

4. From the list below, select the actions or checks that you should take when installing the P&D valve. Place a check mark in the yes column of the answer sheet if you should perform the action. Place a check mark in the no column if you should not perform the action.

(I017) Lubricate the O-rings and seal with graphite grease prior to installation.
(I018) Install elbows, jam nuts, and O-rings on the large outlet and rear ports.
(I019) Leave the jam nuts finger tight until after the fuel manifolds are installed.
(I020) Install a clamp bracket between the P&D valve and the main oil cooler.
(I021) Install the valve in the correct position with the ports facing the appropriate lines.
(I022) Torque the four bolts holding the valve to the main oil cooler.
(I023) Safety wire all four bolts together in one series.
(I024) Install two bolts to secure the valve to the rear mounting bracket.
(I025) Install an O-ring, drain tube, O-ring, and connector bolt (in that order) in the drain port.
(I026) Torque the two bolts securing the valve to the rear mounting bracket.
(I027) Position the two large elbows to the rear and use a common screwdriver to align the fittings when installing the fuel manifolds.
(I028) Torque the jam nuts and manifolds, and lockwire.
(I029) Position the rear elbow and install the reference fuel pressure manifold tube.
(I030) Torque the jam nut and tube coupling and lockwire.

Data Collection

The Jet Engine Mechanic Job Knowledge Test was administered one on one to an aggregate 44 first term jet engine mechanics at five air stations in CONUS and Hawaii. Approximately one half of the incumbents took the test before Walk Through Performance Testing. The remaining incumbents attempted Walk Through Performance Testing first.

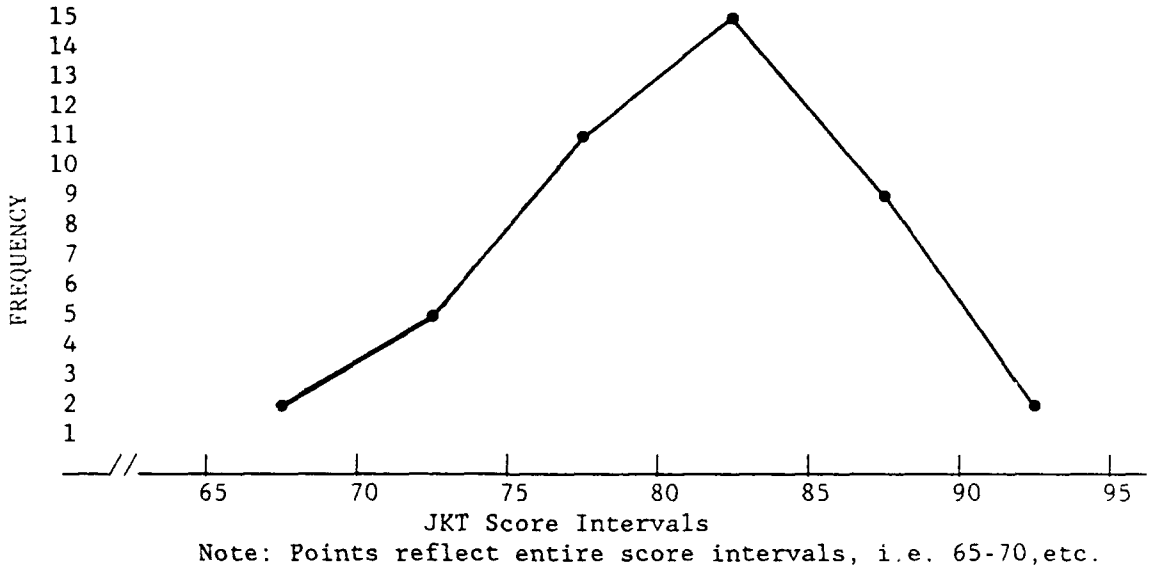
The Job Knowledge Test used step pass/fail scoring. Each of the ten tasks had a ten point value. A subject's total JKT score was obtained by summing the ten scores after deductions were made for step errors. Error deduction points were calculated for each task by dividing ten by the total number of steps in the task.

Results

The Job Knowledge Test scores ranged from 66 to 93 with a mean of 80 and a standard deviation of 5.85. Figure 1 shows the distribution of the 44 Job Knowledge Test scores.

Figure 1

Job Knowledge Test Score Distribution



Note: Points reflect entire score intervals, i.e. 65-70,etc.

The scores were correlated with pay grade, whether or not an individual received a high school diploma or merely sat through 12 years of high school, J-79 school final course grade, the Armed Forces Qualifying Test (AFQT), and the General Technical and Mechanical Maintenance sub scores of the Armed Services Vocational Aptitude Battery (ASVAB). Significant positive relationships were found between the Job Knowledge Test scores and the AFQT and Mechanical Maintenance Scores. Insignificant relationships were found for the remaining variables.

Disclaimer

The opinions expressed here are those of the authors and do not necessarily express those of the Department of the Navy.

REFERENCES

Blackhurst, J. L. & Baker, H. G. (1985) Inter-service transfer of job performance measurement technology. A paper presented at the 27th Annual Conference of the Military Testing Association, San Diego, CA, 21-25 October 1985, pages 327-330.

Hedge, J. W. (1984) The methodology of walk-through performance testing. A paper presented at the annual meeting of the American Psychological Association, Toronto.

INTER-SERVICE TECHNOLOGY TRANSFER:

PERFORMANCE TESTING OF JET ENGINE MECHANICS

Herbert George Baker, PhD
Navy Personnel Research and Development Center

Major Jack L. Blackhurst
Air Force Human Resources Laboratory

Abstract

Tri-service cooperation capitalized on a unique opportunity to address some facets of the feasibility of the inter-service transfer of job performance measurement technologies. A test package developed by the Air Force to measure the job proficiency of first-term J-79 jet engine mechanics was modified to allow for aircraft design differences, and administered to a sample of first-term J-79 jet engine mechanics in the Marine Corps. This paper discusses test package content, the processes of adaptation and data collection, and further technology transfer potentials of this test package.

INTRODUCTION

In an endeavor to improve the classification of individuals into jobs for which they are optimally suited, the Armed Services are investigating the feasibility of linking enlistment standards directly to on-the-job performance. Within this global effort, each Service is developing measurement technologies to collect valid, accurate and reliable hands-on job performance information for selected occupational specialties.

In hopes of lessening the fiscal impact of the JPM studies, surrogate measures, which are less expensive and easier to administer than hands-on measures, are also being developed. Technology transfer is another possible cost reduction strategy. Hands-on and surrogate measures can be developed for specialties which are comparable across the services. Sharing testing technologies could reduce the costs eventuating from in-depth task analyses and test development. The Air Force was the first to develop a potentially transferable testing methodology, the Jet Engine Mechanic Job Performance Measurement System (JPMS) (Alba, 1986; Blackhurst & Baker, 1985).

THE AIR FORCE JOB PERFORMANCE MEASUREMENT SYSTEM (JPMS)

The Air Force Jet Engine Mechanic JPMS consists of Walk Through Performance Testing (WTPT), four rating forms, four questionnaires, and administration instructions for all instruments. The WTPT is a task specific procedure that combines observation of hands-on performance and interview about intended performance for a set of representative tasks. Interview testing allows the inclusion of tasks which would otherwise be precluded because of safety considerations, equipment availability, time, and costs. The WTPT is individually administered and requires an experienced administrator to observe and evaluate an incumbent's hands-on performance and verbal responses. To determine if the job measurement via interviewing is equivalent to hands-on performance a subset of tasks is assessed by both observation and interview.

Four rating forms are group administered, being completed at self, peer, and supervisor levels: (1) the Global Rating Form (technical and interpersonal competence); (2) the Dimension Rating Form (occupational specialty behavioral items); (3) the Task Rating Form (task specific measures, e.g., installing or removing components); and (4) the Air Force Wide Rating Form (military related performance factors (e.g., leadership and integrity). In addition, four questionnaires obtain background and experience information as well as incumbent, supervisor and peer opinions on the quality of the entire JPMS.

The development of the Air Force JPMS included task selection based on extensive task analysis and expert judgment. Complete administration procedures were developed and a comprehensive evaluator training program instituted. Thorough pretesting preceded data collection (Alba & Dickinson, 1985; Bierstedt, 1985b). Complete details and training materials can be found in Alba and Wilcox (1985), Bierstedt and Hedge (1985), and Bierstedt (1985c, 1986). Detailed pretest description and results can be found in Bierstedt (1985a). For data collection results see Bierstedt, 1986).

J-79 JPMS TRANSFER TO THE NAVAL SERVICES

The transfer of the Air Force JPMS to the Navy was conducted in two phases. Phase I was a feasibility study to determine the practicality of the transfer. A team consisting of two senior Marine Corps jet engine mechanics and three contractors who were familiar with the complete Air Force test package concluded that, except for two tasks, the Air Force jet engine mechanic instruments were, with slight modification, transferable to the Naval services. The team used the Air Force Task Selection Plan, the Navy and Marine Corps Training Outline and the Jet Engine Occupational Data Summary to identify two replacement tasks.

During that time, it became known that the J-79 engine was being rapidly phased out of the active duty Navy system. Because Marine Corps jet engine mechanics receive training identical to that of their Navy counterparts, and because a small sample of first-term J-79 mechanics were still available as test subjects, assistance was requested from Headquarters, U.S. Marine Corps. The agreement of the Marines to supply test subjects created, in effect, a tri-service effort.

Phase II of the transfer included a task analysis of the two replacement tasks, development of the new items, modification of the remaining instruments, and pilot testing to ensure that the methodology was applicable to the testing of naval personnel. The technology transfer resulted in the following products:

(1) Rating Form Booklet for Navy/Marine Corps J-79 Jet Engine Mechanics (Intermediate Maintenance Activity [IMA]);
(2) Rating Form Booklet for Navy/Marine Corps J-79 Jet Engine Mechanics (Organizational Maintenance Activity [OMA]);
(3) Rating Forms Rater Training Program Trainee Booklet for Navy/Marine Corps Jet Engine Mechanics;
(4) Rating Forms Rater Training Program Administrator's Guide for Navy/Marine Corps Jet Engine Mechanics;
(5) Rating Form Questionnaire;
(6) General Background Questionnaire;
(7) J-79 IMA Task Experience Rating Questionnaire;
(8) J-79 OMA Task Experience Rating Questionnaire;
(9) General Utility/Acceptability questionnaire for Navy/Marine Corps J-79 Jet Engine Mechanic Performancce Assessment Systemm;
(10) Walk Through Performance Testing for Navy/Marine Corps J-79 Jet Engine Mechanics (IMA);
(11) Walk Through Performance Testing for Navy/Marine Corps J-79 Jet Engine Mechanics (OMA);
(12) Walk Through Performance Testing Administration Manual for Navy/Marine Corps J-79 Jet Engine Mechanics (IMA and OMA);
(13) Answer Sheets for WTPT and the Rating Forms.

Table 1 compares the tasks included in the Air Force and the Navy/Marine Corps WTPT. Table 2 lists an extract of items included in the Air Force and Navy/Marine Corps rating forms. Table 3 summarizes the steps taken to develop, modify and transfer the Air Force jet engine mechanic JPMS to the Navy.

Cost

Excluding the pre test, the Air Force job performance measurement methodology required 18 months to develop and cost

approximately $143,000 for contractor services and travel. An additional $107,300 was paid for civilian employee and active duty labor and travel.

Transfer of the Air force JPMS took an additional 6 months and cost approximately $53,500. Contract salaries, travel and other expenses totaled $46,000. Civilian and active duty man hours and travel accounted for approximately $7,500. Table 4 provides a break down of these costs. Note that table 4 only includes civilian and active duty manhours, salaries and travel. Other costs such as civilian and active duty administrative support and employee benefits are excluded. Table 4 includes all contractor charges.

Table 1. TASKS INCLUDED IN AIR FORCE AND NAVY/MARINE CORPS WTPTs

TASK (ITEM)	Item TYPE	Air Force SHOP	Air Force FL	N/MC IMA	N/MC OMA
Complete Forms	I/H	X	X	X	X
Inspect Engine Plumbing	H	X	X	X	X
Inspect Trailer	H	X	X	X	X
Install Lockwire	I/H	X	X	X	X
Install Starter	I/H	X	X		
Install Constant Speed	I	X	X		X
Install Anti-Icing Duct	I/H	X	X	X	X
Install EGT Harness	I/H	X	X	X	X
Install Exciter Box	H	X	X	X	X
Rig Inlet Guide Vane Components	H	X	X	X	X
Install Bleed Air System Component	H	X	X	X	X
Rig Afterburner Components	H	X	X	X	X
Install Bearings	I	X		X	
Install Oil Seals	I	X		X	
Remove Rotor Assembly	I	X		X	
Isolate Fuel Malfunction	I		X		X
Determine Source of High Oil Consumption	I		X		X
Isolate Starter Malfunction	I		X		
Install P&D Valve	I/H			X	X
Install AB Control Valve	I			X	X
Total # of Tasks included in WTPT		15	15	15	15
Total # of items in WTPT with 5 I/H overlap		20	20	20	20

I = Interim H = Hands-on

The transferred instruments were administered to 44 first term Marine jet engine mechanics at 5 air stations over a 7 month period (minus 2 months of downtime). The total cost for training the test administrators and collecting and analyzing the Marine Corps data was $132,000. The cost reflects fringe benefits, overhead, G&A, fees, communication, reproduction, and salary for training, trip preparation, trip reports, scoring, scheduling, planning, data entry, data analysis, and other technical and administrative procedures. The direct cost for administering the performance evaluation instruments to 44 incumbents was $29,000. This amount includes only travel and direct salary for 2 test administrators and one task monitor while on site. Table 5 shows the per-incumbent instrument administration costs.

Table 2. EXTRACT OF ITEMS INCLUDED IN AIR FORCE AND NAVY/MARINE CORPS RATING FORMS

ITEM	AIR FORCE SHOP	AIR FORCE FL	N/MC IMA	N/MC OMA
GLOBAL RATING FORM				
Technical Proficiency	X	X	X	X
Interpersonal Proficiency	X	X	X	X
DIMENSIONAL RATING FORM				
Completion of Forms	X	X	X	X
Remove/Replace Components	X	X	X	X
Inspect Engines	X	X	X	X
Quality Control	X	X	X	X
Maintenance	X	X	X	X
Trouble shoot		X		X
Prepare Engine for Shipment	X		X	
TASK RATING FORM				
Complete Maintenance Forms	X	X	X	X
Inspect Plumbing				
Install lockwire	X	X	X	X
Inspect Trailers	X	X	X	X
Inspect for FOD Matter	X	X	X	X
Inspect Compressors	X	X	X	X
Install Protective Covers	X	X	X	X
Transport Engines	X	X	X	X
Install Tachometer Generators	X	X	X	X
Inspect Splines	X	X	X	X
Install Afterburner Flaps	X	X	X	X
Install EGT Theromocouple Harnesses	X	X	X	X
Install Number 3 Oil Seals	X		X	
Remove Turbine Rotor Assemblies	X		X	
Remove Gear Boxes	X		X	
Inspect Turbine Nozzles	X		X	
Wrap Engines for Shipment	X		X	
Inspect Engine Bearings	X		X	
Service Engine Starters	X	X		
Install Starters	X	X		
Install Starter Adapter Pads	X	X		
Rig Inlet Guide Vane System	X	X	X	X
AIR FORCE OR NAVY/MARINE CORPS WIDE RATING FORMS				
Technical Knowledge/Skill	X	X	X	X
Initiative/Effort	X	X	X	X
Knowledge of and Adherence to Regulations	X	X	X	X
Integrity	X	X	X	X
Leadership	X	X	X	X
Military Appearance	X	X	X	X
Self Development	X	X	X	X
Self Control	X	X	X	X

Table 3. STEPS TAKEN TO MODIFY AND TRANSFER THE JPMS

- Conduct transfer feasibility study (Air Force to Navy)
- Request assistance from USMC
- Obtain Marine SMEs
- Analyze Navy/Marine Corps tasks
- Develop two replacement items
- Modify Air Force Instruments for Navy/Marine Corps use
- Pilot test Navy/Marine Corps instruments
- Refine instruments
- Conduct data collection and analyses

Table 4. JPMS DEVELOPMENT COSTS

	AIR FORCE DEVELOPMENT MANHOURS	AIR FORCE DEVELOPMENT DOLLARS	TRANSFER TO NAVY MANHOURS	TRANSFER TO NAVY DOLLARS
ACTIVE DUTY	3,797	$27,000	260	$ 5,200
CIVILIAN	2,509	80,300	102	2,300
CONTRACTOR	4,164	143,000	1,860	46,000
TOTAL	11,470	$250,300	2,222*	$53,500*

* Includes administration of an additional job knowledge test.

Table 5. JPMS ADMINISTRATION COSTS

ELEMENT	DOLLARS
TRAVEL	$536
RATING FORMS AND QUESTIONNAIRES	5*
WALK THROUGH PERFORMANCE TEST	74
TOTAL	$616

* Cost based on an average of 12 individuals and 2 test administators per session.

CONCLUSION

The transfer of testing technology from the Air Force to the Naval services can be considered a successful prototypic venture. Because the effort was overtaken by events and the sample is very small, data are not comparable nor generalizable to any great degree. Nevertheless, inter-service cooperation has resulted in significant temporal and fiscal economics in terms of test development. It has been demonstrated that technology transfer is one potentially important source of JPM program economy in the overall DoD effort, and should be evaluated in further studies.

In addition to possible use by the Marine Corps for training assessment or reserve skills testing, the test package shows high potential for implementation in the Navy's surface fleet where the J-79 engine is used in several classes of ships. Also, the J-79 engine is employed in a number of planes used by several allied forces, which could result in inter-allied technology transfer.

REFERENCES

Alba, P. A. (1986). Transfer of Air Force performance measurement technology to the Navy: Final report. Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Alba, P.A., & Dickinson, T. L. (1985). Walk through performance testing documentation for jet engine mechanic (AFS 426X2). Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Alba, P.A., & Wilcox, T. R. (1985). Walk through performance testing manual for jet engine mechanic (AFS 426X2). A test developed for AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Bierstedt, S. A. (1985a). Pretest data summary for jet engine mechanic specialty. Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Bierstedt, S. A. (1985b). Rating form development for jet engine mechanic specialty (AFS 426X2). Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Bierstedt, S. A. (1985c). Rating forms rater training program administrator's guide for jet engine mechanic specialty (AFS 426X2). A rater training manual developed for AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Bierstedt, S. A. (1986). Collection and analysis of performance data for jet engine mechanic specialty (AFS 426X2). Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Bierstedt, S. A., & Hedge, J. W. (1985). Job performance measurement system (JPMS) trainer's manual. Report submitted to AFHRL, Brooks AFB, TX: Air Force Human Resources Laboratory.

Blackhurst, J. L., & Baker, H. G. (1985). Inter-service transfer of job performance measurement technology. Proceedings, 27th Annual Conference of the Military Testing Association. San Diego: Navy Personnel Research & Development Center.

Patterns of Skill Level One Performance in Representative Army Jobs: Common and Technical Task Comparisons

Roy C. Campbell
Charlotte H. Campbell
and
Earl L. Doyle
Human Resources Research Organization

In the project for Improving the Selection, Classification and Utilization of Army Enlisted Personnel, commonly known as Project A, nine jobs or military occupational specialties (MOS) were covered intensively in the concurrent validation. The coverage included, among other measures, hands-on tests and written tests based on task samples for each MOS. The MOS, along with the number tested for each method, are shown in Table 1.

Table 1

MOS and Number Tested

MOS	SL1 Title	Written N	Hands-On N
11B	Infantryman	678	682
13B	Cannon Crewman	639	619
19E	Armor Crewman	459	474
31C	Single Channel Radio Operator	326	341
63B	Light Wheel Vehicle Mechanic	596	569
64C	Motor Transport Operator	668	640
71L	Administrative Specialist	501	494
91A	Medical Specialist	483	496
95B	Military Police	665	665

Army doctrine specifies that all skill level one soldiers are responsible for being able to perform all tasks in their MOS skill level one Soldier's Manual (SM) as well as the tasks listed in the skill level one Soldier's Manual of Common Tasks (SMCT). This latter document lists those tasks, known as Common Tasks, that every soldier, regardless of job or location, must be able to perform to survive in a hostile combat environment.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

For Project A, the domain definition for each MOS consisted of these two types of tasks--those that were included because they were dictated by the soldier's job (MOS-specific or Technical tasks) and those that were included because Army doctrine requires all soldiers to perform minimum essential tasks dictated by exposure to wartime conditions (Common tasks). During the final process in which tasks from each domain were selected for testing, the process was structured so that the selection would represent the full range of task requirements in an MOS. Thus, for each MOS, the tasks tested include both Technical and Common tasks in both the hands-on and written components.

To be sure, the distinction between Technical and Common tasks is sometimes artificial. The skill level one soldier being trained probably does not discriminate between the two categories. And in many MOS, such as 11B, 95B, and 91A, there is little actual job distinction between MOS-specific and Common tasks. In these, and in some other MOS, if a task did not already exist in the SMCT, the job requirements would dictate the task be included as an MOS-specific task.

Yet much is made over Common Task requirements. The specific task concept for Common tasks began emerging in 1976 but is based on the long established Army tradition and concept that all soldiers, in combat, may be called upon to fulfill certain survival functions. The complexity of the modern battlefield has compounded, not diminished, this requirement. SMCT tasks receive as much attention and revision emphasis by TRADOC as do any of the MOS-specific technical tasks. Units are required to test selected common tasks annually. Army Training and Evaluation (ARTEP) and field exercises for all type units emphasize combat survival along with unit mission performance. But there are differences in emphasis as well. The 11B Infantryman literally lives with his M16 rifle; the 71L Administrative Specialist may only draw his/her M16 for maintenance and quarterly or semi-annual training. Yet by doctrine, each is equally responsible for certain M16 tasks. The question then, is whether there are distinctions among Army jobs in the performance of Common tasks and also whether there are significant distinctions between performance on Technical tasks and Common tasks. Project A, with the test results from over 5000 soldiers, provided an opportunity to examine this issue.

Method

In the 9 MOS, a total of 290 individual tasks were tested. Table 2 shows a breakout of these tasks by Technical and Common category and by test component. Almost all tasks tested in the hands-on component were also tested in the written component; however, there were some tasks tested by written component only.

Table 2

Distribution of Observations by Test Component

	Hands-On	Written
Technical	89	158
Common	60	123

The first analysis considered all the MOS combined (Table 3). There was only a slight and insignificant difference on hands-on results between the Common and Technical domains--the apparent difference being accounted for by the larger variance in performance in the Technical tasks. In the written tests, however, the difference in performance is significant, with higher performance levels reflected in the Common task performance. It should be noted however, that this difference may be the result of test difficulty. As yet, no overall item analysis of the written tests has been performed to identify difficulty patterns.

Table 3

Comparison of Technical and Common Task Performance on Hands-On and Written Tests For Nine MOS Combined

		Test Component	
Tasks		Hands-On	Written
Technical	N of Tasks	89	158
	Mean %	68.2	57.6
	S.D.	19.2	12.8
Common	N of Tasks	60	123
	Mean %	73.3	63.7
	S.D.	15.1	12.9
Test of Difference Between Common and Technical		$t = 1.721$ $p < .09$	$t = 3.948$ $p < .001$

Although the nine MOS were carefully selected to represent the entire domain of Army jobs, the Technical/Common tasks analysis continued by looking at the nine MOS broken down into families. These family classifications followed the groupings developed by McLaughlin, Rossmeissl, Wise, Brandt, and Wang (1984). Three families are represented: Family I is Combat (11B, 13B, 19E), Family II is Operations (31C, 63B, 64C), and Family III is Skilled/Technical (71L, 91A, 95B). (The 71L MOS actually belongs in a fourth job family--Clerical--but we have grouped it with the Skilled/Technical MOS for the analyses reported here.)

Table 4 shows the results by this family breakout. For the written tests there are no significant differences in performance between families, that is, where family membership affects outcome. In the hands-on tests, however, there appears to be a significant difference by family--Families I, II and III being each separated by about 5 points in performance. Closer examination however reveals that much of this difference by family is due to interaction between Common and Technical tasks within the family. Common task performance across families is quite consistent. The difference between families is accounted for almost solely by the Technical tasks, with 17 points difference in mean performance between the two most separated families.

Table 4

Performance Results Based on Family Membership

		Job Family		
		I - Combat	II - Operations	III - Skilled/ Technical
Hands-On Component Tasks				
Technical	N of Tasks	36	24	29
	Mean %	61.4	78.4	68.3
	S.D.	23.3	12.0	14.7
Common	N of Tasks	19	22	19
	Mean %	72.8	73.7	73.3
	S.D.	16.9	15.8	12.9
Written Component Tasks				
Technical	N of Tasks	63	49	46
	Mean %	55.1	58.1	60.4
	S.D.	13.0	11.2	13.0
Common	N of Tasks	44	39	40
	Mean %	63.0	63.0	64.2
	S.D.	12.9	13.7	12.3

Analysis of Variance: Job Family x Technical/Common

Hands-On Component

Source	SS	df	MS	F	p
Job Family	1910.16	2	955.08	3.28	.04
Technical/Common	543.78	1	543.78	1.86	.17
Family x Technical/Common	1561.08	2	780.54	2.68	.07
Error	41694.06	143	291.57		

Written Component

Source	SS	df	MS	F	p
Job Family	508.83	2	254.41	1.55	.21
Technical/Common	2316.63	1	2316.63	14.14	.00
Family x Technical/Common	205.18	2	102.59	.63	NS
Error	45062.95	275	163.86		

Within families however, there is always a significant difference between Technical task performance and Common task performance. However, this performance difference is not entirely consistent--in Families I and III, Common task performance is better than Technical performance. In Family II, the opposite is true for the hands-on test although the trend shown in Families I and III holds true for the written tests.

Conclusions

For a variety of reasons, relative differences in performance between Army jobs were expected. These differences can be variously attributed to innate task difficulty, assignments, training emphasis and even entrance requirements into the MOS. However it would appear that the Army policy regarding Common task proficiency appears to be working. While differences in performances between groups of MOS showed up as expected, these differences were almost entirely attributable to technical tasks within each group. Common task performance is remarkably uniform between Family groups. Based on Project A results it would appear the Army Common Task Management has produced its desired results.

References

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 FY1981 and FY1982 enlisted accessions (ARI Technical Report 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Effect of Practice on Soldier Task Performance*

Paul Radtke
Dorothy S. Edwards

American Institutes for Research

One of the forms administered in the Army's Selection and Classification study, usually known as Project A, was a Job History Questionnaire. For each of nine Military Occupational Specialties (MOSs) the form listed all of the tasks covered by paper and pencil knowledge tests and by hands-on performance tests. These tasks were selected from the domain of tasks for an MOS by a panel of experts because they were done frequently and were important to overall job performance. About thirty tasks were selected for each MOS; all were measured with performance based knowledge tests; about half were also measured with hands-on tests.

In the Job History Questionnaire soldiers were asked to indicate how often during the past six months they had performed each task, using a scale of "Not at all, 1-2 times, 3-5 times, 6-10 times, or more than 10 times." Next, soldiers indicated how recently they had performed each task, using a scale of "Never, during past month, 1-3 months ago, 4-6 months ago, or more than 6 months ago."

The frequency and recency ratings were correlated with the scores on the knowledge tests and with the hands-on tests for each MOS. The results for two sample MOSs, one combat and one support MOS, are shown in Tables 1-2. The number of cases for these correlations varies, but in every case is substantial. The minimum and maximum N is given at the top to reduce the number of columns in the tables. When there is a wide range in the number of cases it reflects a smaller N on one or two tests and nearly maximum Ns on the others. The size of the N makes a rather small correlation significant statistically; the rather small correlations probably have little practical significance. Note that the recency correlations should be negative, because of the way the scale was written.

The tables have some items of interest, however. Recency appears to be more closely associated with test performance than does frequency of practice, in that more of these correlations attain statistical significance. Recency and frequency are correlated, as shown in the last column of the tables.

There is a tendency for the more complex tasks to be more highly correlated with frequency and recency, though there are some exceptions in both directions -- complex tasks not correlated or easy tasks correlated.

Performance on MOS-specific tasks tends to be more highly correlated with frequency and recency of practice than performance on the common

---

*This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Table 1. Correlation between Job History Questionnaire Scales (Frequency and Recency of Performance) and Task Test Scores for 11B (Infantryman)

Decimal points omitted; * = significant at P = .01

	K Tests (N=495-697)		HO Tests (N=496-696)		Frequency & Recency
	Freq.	Rec.	Freq.	Rec.	
11 B Knowledge					
Perform CPR	04	-17*			-61
Adm Nerve Agent Antidote	04	-13*			-48
Put on Field/Pres. Dressing	03	-06	09	-09	-41
Perform OP Maint. on M16A1	06	-08	08	-04	-40
Load/Reduce/Clear M60	12*	-09	25*	-20*	-50
Engage w Hand Grenades	08	-13*	05	-08	-46
Prepare Dragon for Firing	14*	-16*	18*	-18*	-57
Prepare Range Card for M60	00	-10	24*	-19*	-54
Call for/Adjust Indirect Fire	16*	-16*			-60
Navigate on the Ground	16*	-20*			-52
Id Terrain Features on MAP	10*	-10*			-37
Put on M17 Mask	02	-05	07	-08	-48
Put on Protective Clothing	02	-13*			-39
Collect/Report Info	-06	-06			-38
Camouflage Self/Equip	07	-07			-42
Id Armored Vehicles	15*	-13*			-45
Move under Direct Fire	00	-03			-63
Estimate Range	-02	04			-58
Move over Obstacles	04	-01			-52
Operate Radio Set AN/PRC-77	09	-10*	04	-02	-50
Install/Fire Claymore Mine	07	-07	22*	-19*	-45
Tech of Urban Terr Movement	-06	03	-03	-04	-49
Select Hasty Urban Firing Pos	-02	-03			-64
Establish Obv Post	01	-08			-64
Sel Fire Team/Overwatch Pos	02	-03			-72
Zero AN/PVS-4 to M16A1	-02	-02	-06	-17*	-65
Place AN/PVS-5 into operation	09	-04			-60
Set Headspace/Timing on .50	37*	-37*	41*	-35*	-72
Engage Target w LAW			-04	-03	-52

Table 2
71L (Administrative Specialist)

71L Knowledge	K Tests (N=498-508)		HO Tests (N=494-508)		Frequency & Recency
	Freq.	Rec.	Freq.	Rec.	
Adm Nerve Agent Anti-Self	07	-20*			-49
Load/Clear M16A1	-01	-01			-32
Oper Maint M16A1	06	-06	02	-10	-36
Det Magnetic Azimuth	-02	-08			-34
Det Grid Coordinates	08	-17*	09	-16*	-40
Put on M17 Mask	-05	03	07	-21*	-34
Maintain M17 Mask	04	-09			-46
Put on Protective Clothing	18*	-18*			-45
Know Rights as POW	09	-15*			-63
Camouflage Self/Equip	14*	-12*			-55
Prac Noise/Light/Litter Disc	-16*	-16*			-67
File Documents/Corresp	18*	-13*	10	-12*	-80
Est Functional Files	04	-06			-72
Control Supplies	07	-09			-88
Rec/Contl Office Equip	02	00			-84
Dispatch Outgoing Dist.	11*	-10*			-79
Type Military Orders	01	-02	10	-11*	-77
Type 2nd Comment to DF	32*	-30*	12*	-11*	-80
Type Jt Message Form	15*	-18*	08	-14*	-79
Type a Memo	14*	-19*	08	-14*	-76
Type a Basic Comment to DF	20*	-20*	20*	-24*	-80
Assemble Correspondence	16*	-13*			-79
Type Military Letter	25*	-24*	18*	-15*	-80
Safeguard FOUO Material	03	-04			-84
Rec/Trans Classified Material	10	-06	16*	-12*	-82
Put on Field/Press Dressing			-02	-13*	-34
Prep. Requisition/AUTODIN			17*	-14*	-75

Table 3. Correlations Between Job History Questionnaire Scales and Scores on Common Soldiering Tasks
Decimal points ommitted: * = significant at P = .01

A. Frequency - Knowledge Tests

K Tests	11B	13B	19E	31C	63B	64C	71L	91A	95B
CPR	04	-02		03		09		00	-11*
Nerve agent	04	02	01			07	07		
F/P dressing	03		10	15*	09	07		-03	10*
LRC M16		-09		-01	01	-02	-01	07	00
Op/Mtn M16	06			-03	05	03	06		
LRC M60	12*					18*			20*
Mag. Azim.					04		-02		12*
Grid Coor.			15*	16*	13*	13*	08	18*	08
Put on mask	02	07			05	05	-05		08
MOPP	02	07	12*	09	16*	15*	18*	14*	
CEOI			38*						24*

B. Recency - Knowledge Tests

CPR	-17*	02		-09		-10*		00	-13*
Nerve agent	-13*	-12*	-11*			-06	-20*		
F/P dressing	-06		-04	-12*	-07	-08		-05	-08
LRC M16		00		-12*	-02	-01	01	-16*	-02
Op/Mtn M16	-08			00	-02	-10*	-06		
LRC M60	-09					-15*			-22*
Mag. Azim.					-06		-08		-05
Grid Coor.			-12*	-23*	-18*	-18*	-17*	-20*	-04
Put on mask	-05	-04			00	-04	03		-06
MOPP	-13*	-04	-12*	-15*	-11*	-12*	-18*	-12*	
CEOI			-29*						-27*

C. Frequency - HQ Tests

CPR		12*				15*		13*	17*
Nerve agent		01				11*			
F/P dressing	09		06	04	02	01	-02	07	14*
LRC M16		01		-03	-01	01			06
Op/Mtn M16	08					05	02		
LRC M60	25*					14*			07
Mag. Azim.					01				07
Grid Coor.			09	22*		13*	09	18*	17*
Put on mask	07	04			-04	08	07		09
MOPP		05		01		04			
CEOI			31*						

D. Recency - HQ Tests

CPR		-19*				-18*		-09	-15*
Nerve agent		-01				-10*			
F/P dressing	-09		-10	-15*	-06	-08	-02	-11	-13*
LRC M16		01		-07	-02	02			-12*
Op/Mtn M16	-04					-07	-10		
LRC M60	-20*					-30*			-03
Mag. Azim.					-12*				-02
Grid Coor.			-06	-21*		-17*	-16*	-19*	-13*
Put on mask	-08	00			-02	-06	-21*		-06
MOPP		-06		-07		-08			
CEOI			-27*						

tasks. It may be that common tasks have been subject to more practice during the soldier's enlistment. This hypothesis is consistent with the generally higher mean scores on the common tasks. If true, the common tasks may have been "overlearned." and thus less subject to forgetting or to decrement through lack of practice.

Some common tasks were tested in more than one MOS. This allows us another way to look for consistency in association of test scores and frequency or recency of practice. Table 3 shows these data for the common tasks. One task, "Determine grid coordinates" shows significant correlations with frequency and recency in six of the seven MOSs in which the knowledge test was given. It also showed significant correlations in the hands-on tests in most of the MOSs in which it was given. It is the consistency of the findings rather than the magnitude of the actual correlations that makes us believe that competency in this task is indeed related to frequency and recency of practice. The test was very similar in both measurement methods: soldiers had to read grid coordinates using a protractor. They had an advantage in the written mode in that the correct answer appeared as one of four choices, whereas they had to report the coordinates to the test administrator in the hands-on mode without the recognition advantage afforded by the multiple choice item.

A second test that has a similar pattern of significant correlations with the knowledge tests is "Put on and wear protective clothing." This test, however, does not correlate with the hands-on measure. Since the soldier must put on the clothing required at four progressive levels of protection, over-dressing at phase 1, or MOPP Level 1, as it is called, could keep the soldier from correctly reaching the higher levels.

Naturally we looked for characteristics that these two tasks have in common that are not present in other tasks that do not show this pattern of correlations. We found only one. Each of the tasks requires a specific procedure that terminates in an objectively verifiable product or result. Exact grid coordinates are determined and reported, and certain garments are worn at each MOPP level. This means that the "right answers" are totally unequivocal and readily observable by even a careless scorer in the hands-on mode. These tests had reliability estimates that were among the highest in the MOSs in which they appeared, which is probably also a function of the clarity and observability of the response.

Another test that is fairly consistent in correlations with frequency and recency is "Load, reduce, and clear the M60 machinegun." It was given in only three MOSs, so the consistency cannot be as pronounced as with the grid coordinates and protective clothing tests. Table 4 shows the correlations for this task as well as those for a similar task: "Load, reduce, and clear the M16A1 rifle." Performance on the M16 tests is not as highly correlated with frequency, probably because it is the soldier's main weapon and is more often practiced and proficiency is maintained at a high level. The task is also somewhat simpler than the matching task on the M60.

At the bottom of Table 4 we have shown the mean percent passing the knowledge tests and the mean percent "GO" on the hands-on test for all MOSs in which the M60 and M16 tasks were covered. Note that performance on the hands-on test is higher than on the knowledge test for both tasks,

but the performance on the M16 weapon is superior to performance on the M60. The M60 task is somewhat more complex, and has more steps, but the M16 is almost certainly practiced more often. Soldiers do appear to be able to load, reduce, and clear their primary weapon, as indicated by the mean of 85% GO on the hands-on test.

Table 4. Correlations between frequency and recency of practice and test scores on two weapons, the M60 machinegun and the M16 rifle

	Infantry	Cannon Crewman	Radio Operator	Auto Mechanic	Truck Driver	Admin. Specialist	Medic	Military Police
Frequency								
LRC M60 K	12*			18*				20*
LRC M60 HO	←	25*		14*				07
LRC M16 K		-09	-01	01	-02	-01	07	00
LRC M16 HO		01	-03	-01	01			06
Recency								
LRC M60 K	-09			-15*				-22*
LRC M60 HO	-20*			-30*				-03
LRC M16 K		00	-12	-02	-01	01	-16*	-02
LRC M16 HO		01	-07	-02	02			-12*

	LRC M16	LRC M60
Mean % correct, K	72.79	61.80
Mean % GO, H-O	85.84	68.35

A final test that shows substantial correlation with both frequency and recency of practice is "Use automated CEOI" (Communications Electronics Operating Instructions). It was given in only two MOSs, and is similar to

grid coordinates in that it results in an objectively observable result. The correlations were as shown below:

	Tank Crewman	MP
CEOI K-test & freq.	38*	24*
CEOI K-test & recency	-29*	-27*
CEOI H-O test & freq.	31*	Not given
CEOI H-O test & recency	-27*	Not given

This test requires memory of procedures for looking up information in a table and reporting call signs, radio frequencies, and authentication data. A number of soldiers taking the hands-on test reported on how easily the procedures for reading the table are forgotten.

Conclusions

The ratings on frequency and recency of practice of tasks tested in Project A show very low correlations with test performance. There are, however, some tasks that show a significant relationship, and in a consistent enough manner to suggest that we are not dealing with chance results.

Tasks that are related to practice seem to be those that produce objectively observable results, that are relatively complex, and related to the MOS specific parts of the job rather than to the common soldier tasks.

Reference

Campbell, J.P. 1986, August. Project A: When the textbook goes operational. Paper presented at the 94th Annual Convention of the American Psychological Association. Washington, D.C.

Effects of Test Programs on Task Proficiency

Patrick Ford and R. Gene Hoffman
Human Resources Research Organization

The general purpose of Project A is to predict job performance by establishing the relationship between entry measures and performance on a sample of job tasks in nine selected MOS (Eaton, Goer, Harris, & Zook, 1984). At a conceptual level the relation between applicants' ability and the tasks on a job ought to be stable so long as the job does not change. In practice, however, there are several mediators between ability and performance. Among the potential mediators are test programs that focus individual training in units. In these programs a central agency establishes a set of tasks that are to be tested and, presumably, trained in units. Data collected for Project A during June to November 1985 provide an opportunity to look at the effect of these programs on soldier performance.

This paper considers three programs that may affect task proficiency:

- Common Task Test (CTT). This is a hands-on test that all soldiers are to take each year. The Training and Doctrine Command selects a subset of tasks from the skill level one Soldier's Manual of Common Tasks. During the Project A data collection the operative CTT had 19 tasks. Across the nine MOS, the test samples for Project A included 14 of them. For comparison, 25 other non-CTT common tasks were also in the Project A data base.

- Expert Infantry Badge (EIB). This is a hands-on test that is administered to eligible infantrymen (MOS 11B). During the data collection it included 21 tasks of which 8 were included in the Project A 11B sample. The 11B test battery included 21 other tasks.

- Expert Field Medical Badge (EFMB). This is a written and hands-on test that is administered to medical specialists (MOS 91A). During the data collection the hands-on section included 32 tasks of which 10 were included in the Project A 91A sample. The 91A test battery included 20 other tasks.

The criterion measures for looking at the effect of the test programs are results from tests administered as part of Project A. There are two types of criterion measures:

- Hands-On Tests - These tests were based on direct observation of a soldier's performance of a job task. The tests were developed to provide consistent conditions for performance.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Scores were percent of steps performed correctly or, in some cases, percent of product prepared correctly. There was a separate score for each task.

- Written Tests - These tests were in a multiple-choice format. Items were organized into subtests with each subtest corresponding to a job task. The score was percent correct by task.

During the data collection (which was the Project A Concurrent Validation), the tests had been administered to over 5000 skill level one (SL1) soldiers in nine MOS. The MOS covered along with the number of soldiers tested for each method are shown in Table 1.

Table 1

MOS and Number Tested

MOS	SL1 Title	Written N	Hands-On N
11B	Infantryman	678	682
13B	Cannon Crewman	639	619
19E	Armor Crewman	459	474
31C	Single Channel Radio Operator	326	341
63B	Light Wheel Vehicle Mechanic	596	569
64C	Motor Transport Operator	668	640
71L	Administrative Specialist	501	494
91A	Medical Specialist	483	496
95B	Military Police	665	665

Approach

The CTT analyses were limited to the SL1 common tasks (defined as tasks included in the SL1 Soldier's Manual of Common Tasks). Performance on Project A tasks that were also on the CTT was compared with performance on Project A SL1 common tasks that were not on the CTT. The comparison was made on two levels--across all nine MOS and by MOS family. The MOS families were based on previous work (Rossmeissl, Wise, Brandt, & Wang, 1984) that identified four families: combat (11B, 13B, 19E); operations (31C, 63B, 64C); clerical (71L); and skilled technical (91A, 95B). The CTT analyses combined the clerical and skilled technical families. The analyses were conducted separately for hands-on and written criteria.

The analysis of specific MOS programs included all Project A tasks for MOS 11B and 91A respectively. Two comparisons per method were conducted for each program: (1) MOS program (EIB or EFMB) tasks and CTT tasks with Project A only tasks and (2) MOS program tasks with tasks not covered by the MOS program (including CTT tasks that were not in EIB or EFMB, respectively).

Results

The CTT comparisons are summarized in Table 2. Whether the differences are statistically significant depends on the orientation of the interpreter. If the question is simply "Does performance on this particular set of CTT tests differ from performance on this particular set of non-CTT tests?" essentially all of the differences would be statistically significant. That is, with test scores as percents, the extremely large number of soldiers tested yield standard errors of the mean for most tests at approximately .9. A more conservative standard is required, however, if the tests are treated as samples of their domain and the pertinent question is "Does performance on all tasks in the CTT domain differ from performance on all tasks in the non-CTT domain?" For the second question, the N is number of tasks sampled within task categories (e.g., CTT/Non-CTT) rather than soldiers.

The CTT comparisons were analyzed by means of a two way analysis of variance using tasks as subjects, with program membership as independent variables. Following the conservative interpretation (N as number of tasks), none of the differences are significant.

Table 2

Summary of Results on CTT Tasks and Non-CTT Common Tasks

Test Mode	Task Type	Family	N of Tasks	Mean	S.D.
Hands-On (60 cases)	CTT	All	28	76.73	8.31
		Combat	8	79.67	7.99
		Operations	11	74.92	7.48
		Skilled Tech. & Clerical	9	76.01	9.68
	Project A Common	All	32	70.40	18.84
		Combat	11	67.87	20.16
		Operations	11	72.49	21.61
		Skilled Tech. & Clerical	10	70.88	15.43
Written (123 cases)	CTT	All	56	65.77	12.54
		Combat	18	66.13	13.04
		Operations	18	67.93	12.34
		Skilled Tech. & Clerical	20	63.51	12.54
	Project A Common	All	67	61.91	13.01
		Combat	26	60.87	12.64
		Operations	21	60.37	14.20
		Skilled Tech. & Clerical	20	64.87	12.32

The EIB comparisons are summarized in Table 3. Here both hands-on comparisons are significant: Special program (EIB or CTT) with no special program ($F=8.022$, $P<.02$); and EIB with non-EIB ($F=6.21$, $P<.05$). Neither written comparison approaches significance.

Table 3

Summary of Results on 11B Special Program Tasks

Test Mode	Task Type	N of Tasks	Mean	S.D.
Hands-On (13 cases)	EIB & CTT	8	80.19	12.70
	Project A Only	5	57.26	16.50
	EIB	6	79.70	14.22
	Non-EIB	7	64.23	18.49
Written (28 cases)	EIB & CTT	13	61.58	9.51
	Project A Only	15	59.18	12.37
	EIB	8	62.21	9.80
	Non-EIB	20	60.03	11.69

The EFMB comparisons are summarized in Table 4. None of the differences are significant.

Table 4

Summary of Results on 91A Special Program Tasks

Test Mode	Task Type	N of Tasks	Mean	S.D.
Hands-On (16 cases)	EFMB & CTT	6	75.27	6.25
	Project A Only	10	70.58	12.49
	EFMB	5	76.43	6.24
	Non-EFMB	11	70.49	11.86
Written (30 cases)	EFMB & CTT	14	68.72	9.27
	Project A Only	16	65.66	13.06
	EFMB	11	70.33	9.40
	Non-EFMB	19	65.21	12.21

Discussion

Our reluctance to call the CTT differences significant ought not to be interpreted to mean that the CTT program makes no difference. All it means is that there is so much variation among the hands-on means for Project A only and so few cases overall that we can not say with confidence that hands-on performance on any set of tasks selected for CTT will be better than performance on tasks not selected. It is possible, for example, that some of the tasks not selected are more complex or require greater coordination than

tasks selected for CTT. Besides the possible sampling error among tasks we must also remember that the Project A results are a snapshot of a wide range of units at different points in their training cycles. Since the CTT effect could weaken over time, any evaluation that does not minimize the delay understates the effect.

The results do suggest that the portion of the CTT captured by Project A during the summer of 1985 had a positive association with hands-on scores. It is somewhat surprising that the difference was strongest in the MOS in the Combat family.

The EIB appears to be a very powerful program and must be considered when interpreting criterion data on 11B. It is less clear, however, that similar programs would achieve comparable results in any MOS. The impact of the EFMB on 91A, for example is not nearly as dramatic. Among the myriad of explanations for the difference, two seem to be especially appropriate. First, the EFMB has not had time to develop the credibility that the EIB has. The credibility of the program affects the number of people who are tested and, probably more important, the intensity of training that precedes the testing. Second, there may be a ceiling effect for 91A. Performance of medical specialists may be high enough without the program that any increment is small.

Conclusion

The impact of test programs on soldier performance is ambiguous. No program considered in this paper had a meaningful effect on performance as measured by written tests. The 1985 CTT apparently affected hands-on results in the Project A data but we cannot generalize that a comparable effect will occur every year. The effect was to equalize performance on a subset of common tasks across MOS mainly by increasing hands-on performance of soldiers in combat MOS. The EIB program had a strong effect on hands-on performance of infantrymen and should be considered as a moderator of 11B performance. However the EFMB program for medical specialists, though parallel to the EIB, did not have a comparable effect.

References

Eaton, N. K., Goer, M. H., Harris, J. H. and Zook, L. M. (Eds.). (1984). Improving the selection, classification, and utilization of army enlisted personnel: Annual report, FY1984 (ARI Technical Report 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

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 FY1981 and FY1982 enlisted accessions (ARI Technical Report 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

CLARIFYING EXPECTATIONS THROUGH AUDIOVISUAL JOB PREVIEW

John P. Wanous, PhD
The Ohio State University

Herbert George Baker, PhD
Navy Personnel Research and Development Center

Abstract

Four audiovisual realistic job preview segments were developed for use in the placement of applicants in the trades apprentice program at the Long Beach Naval Shipyard. This paper discusses the results and details further research needs.

Introduction

Unacceptably high turnover among naval shipyard tradesmen indicates that current placement procedures need sophistication and refinement. Of particular importance is the high turnover among workers in the skilled trades apprentice programs because of costly training investments (Baker, 1986). Shipyard management desires to refine the placement process in order to more optimally classify new workers and improve worker tenure. One method may be through the use of job preview materials.

The objective of this effort was to develop prototype audiovisual materials that will provide a broad orientation to the shipyard work milieu and focused realistic job previews (RJPs) for three specific shipyardd trades. The products of this effort should ultimately conduce to better informed job applicants, thereby contributing to an improved person-job match.

The RJP

The RJP is designed as a recruitment procedure that increases the percentage of newly hired employees who "survive" (i.e., do not quit) during a specified period, e.g., the first 6 months on a new job. The theory underlying the RJP can be summarized as follows. Job candidates who are given the RJP have more realistic expectations than those who are recruited in the more traditional method of "selling" the organization to the recruit. As a result of these more realistic expectations, the recruit: (1) is better able to make an informed decision whether to accept or reject a job offer, (2) feels a greater commitment to a job choice since it was based on more complete information, and (3) is better able to cope with the stress of a new job because the RJP has "inoculated" expectations, i.e., there are fewer surprises and disappointments (Wanous, 1980).

A variety of methods have been used to present realistic information to job candidates, e.g., booklets, films, videotapes, as well as "live" oral presentations -- sometimes followed by

"question-and-answer" sessions. Indeed, the medium used may make a difference in RJP effectiveness. For example, Popovich and Wanous (1982) suggested that audio-visual methods might be more effective than written booklets, based on their assessment of the research literature in social psychology concerned with persuasive communication. Furthermore, there is some empirical support for this contention (Premack & Wanous, 1985).

Regardless of the medium used to present the RJP, all RJPs have one common element -- the accurate depiction of the most important factors influencing a newcomer's decision to remain on the job or to quit. Thus, RJP is based to a great degree on job analysis. Prime targets for inclusion in the RJP are those aspects of the work itself and the working environment, which newcomers identify as their most surprising and disappointing discoveries after beginning a new job. This aspect of RJP has to do with instilling appropriate expectations.

Even though a particular job factor is very important to most employees, this does not necessarily impel its inclusion in the RJP. Pay, for example, is important, but expectations about pay may be more realistic than other job facets (e.g., supervision, working conditions, or promotions) because most organizations advertise pay rates and discuss pay with job candidates.

To summarize: Job factors included in the RJP should be important, but they should also be those that job candidates frequently misperceive. Because the naivete of job candidates is almost always biased toward inflated expectations (Wanous, 1980), the RJP must challenge unwarranted expectations, often serving to deflate misperceptions about job factors.

The first field evaluation of experimental RJPs was conducted in the life insurance industry 30 years ago. Since then there have been over 20 more field experiments of the RJP. A review by Premack and Wanous (1985), using a meta-analysis technique (see Hunter, Schmidt, & Jackson, 1982, for an explanation of the method), is the most comprehensive and current review of this research. That review found that RJP lowers initial job expectations, increases the tendency to reject a job offer (self selection), and increases the job survival of those who do accept the job offer.

The meta-analytic calculations revealed that the above mentioned effects of RJPs are stable across situations when one removes that portion of the between-study variation due to artifactual sources (the principle source being sampling error). Thus, an organization using the RJP is likely to achieve results similar to the average results found in the Premack and Wanous review if the RJP is used on a large number of people.

The fiscal savings accruing to an organization can be considerable (Premack & Wanous, 1985). Generally speaking, the

lower an organization's job survival rate for newcomers, the more the RJP can help by increasing the rate of job survival.

Approach

Construction of the RJP calls for methods that will yield a "content valid" preview. The basic procedure used here was one of obtaining job-relevant information from multiple sources and looking for areas of convergence among the sources. This reduces the chances of biasing the RJP (including material that should not be included and omitting relevant information). Based on the research evidence to date (Popovich & Wanous, 1982; Premack & Wanous, 1985), video cassettes were the medium of choice for this RJP project. The global approach included interview-based job analysis, videotaped interviews, use of file footage, and management review.

Results

Interviews were conducted on-site at the shipyard with those responsible for conducting the apprentice training programs and with groups of job incumbents in the three trades of concern. Within each of the three trade groups, personnel were selected for interviews based on their experience with the trade -- from apprentices to journeymen. A variety of experience levels is highly desirable, since research on employee perceptions of jobs (Wanous, 1980) shows there are systematic changes as the new employee gains more experience.

The initial interviews were conducted in March, 1986, and formed the nucleus of general information about the shipyard as a place to work and specific information about each of the three trades. In June, 1986, a second wave of personnel from the three trades was interviewed, and a cross-section of job experience was obtained. The second round of interviews was videotaped and formed the raw material for the RJPs.

Material included in the edited versions of the RJPs was that which was mentioned by both groups of interviewees. This was done to protect the validity of the RJPs; i.e., if a particular aspect of the work was mentioned by several groups, it was assumed to be a more significant and general facet than something mentioned only once.

Approximately 3 hours of videotaping were required for interviewing personnel from the three trades. The videotape was then reviewed several times and detailed notes of the content were taken. These notes were then analyzed and the content dichotomized into trade-specific and more general comments about the shipyard). Notes from both sets of interviews were compared to identify areas of overlap. These comparisons conduced to a decision about what to include in the final RJPs.

Each of the three trade-specific RJPs are approximately 16 minutes long. The viewing time of the general shipyard RJP is about 25 minutes. This means that about 40% of the raw interview material was used. To supplement the interviews, file footage of the shipyard itself and of personnel performing tasks germane to each of the three trades was also used. This footage was used as a visual aid to enhance a particular point being made in an interview. For example, when a painter speaks about working in a small, closed space, a scene similar to that is shown on the screen, with the individual's voice-over.

The RJPs were shown to a cross section of managers from the Long Beach Naval Shipyard to obtain constructive criticism for possible revisions. Minor changes were then incorporated into the contractor's final product deliverable.

Conclusion

Both the shipyard orientation and the trade-specific RJPs should prove useful in challenging applicant job perceptions and instilling appropriate expectations. However, they may require shortening in view of time constraints associated with applicant processing. In addition, while highly accurate and revealing, the three trade-specific RJPs may need further work to enhance their ability to hold viewer attention (e.g., using more on-screen job incumbent narration). All of the RJP segments developed in this effort should be revised to the point where they can be field tested, perhaps at a job fair or during placement interviews. NAVPERSRANDCEN researchers are continuing to explore ways to enhance and field test the RJP materials.

References

Baker, H. G. (1986, April). Improving the placement and retention of naval shipyard workers. Proceedings of the Psychology in the Department of Defense Tenth Symposium (USAFA-TR-86-1). Colorado Springs, CO: Department of Behavioral Sciences and Leadership, U.S. Air Force Academy.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage Publications.

Popovich, P., & Wanous, J. P. (1982). The realistic job preview as a persuasive communication. Academy of Management Review, 7, 570-578.

Premack, S. L., & Wanous, J. P. (1985). A meta-analysis of realistic job preview experiments. Journal of Applied Psychology, 70, 706-719.

Wanous, J. P. (1980). Organizational entry: Recruitment, selection, and socialization of newcomers. Reading, MA: Addison-Wesley.

MICROCOMPUTER-BASED OCCUPATIONAL INFORMATION AND COUNSELING FOR NAVAL SHIPYARDS

Lila Norris, Educational Testing Service

INTRODUCTION

The Long Beach Naval Shipyard has an extensive apprentice training program that is designed to develop highly-skilled journeyworkers; it provides a pool of trained workers for the shipyards. The program is a four-year planned course of study that includes both classroom instruction and on-the-job learning.

To be eligible for the program applicants must apply to the Office of Personnel Management and then pass a test administered by that office. Eligible applicants are then invited to a Job Fair that is conducted by the shipyard to help applicants decide which trades they want to be considered for. Typically, several hundred applicants attend the Fair, where they are given brief write-ups about the trade options and can speak to representatives from the trades.

While the Job Fair may well serve the purposes of some applicants, others might need more help, either because they haven't given adequate thought to what they want from their work or because they don't have a good understanding of the nature of the trades or the training programs. Still others may need help weighing the pros and cons of several trades they are considering.

Problem

Since training an apprentice involves a large investment in time and money, it is especially important to get as good a match as possible between the person and the job. One possible approach to improving this match is to provide pre-employment counseling that helps people assess their work preferences and gives them comprehensive information about jobs. One promising method of enhancing pre-employment counseling is the use of a microcomputer-based occupational information system.

Objectives

(1) design and develop a prototype of an occupational information system that runs on the IBM PC;
(2) design and develop an on-line Preference Elicitation Instrument (PEI) to assess the work preferences of applicants to the Long Beach Naval Shipyard apprentice program;
(3) develop comprehensive information for three shipyard trades - Electronics Mechanic, Marine Mechanic, Painter.

OVERVIEW OF SEA-LECT

SEA-LECT (which is the name given to the system) was designed as a comprehensive program that covers the major steps in the career

decision-making process. Its overall intended purpose is to help applicants to the training program decide which trades they want to be considered for, and to teach them a process for making that decision. Its specific purposes are to help applicants (1) assess their work preferences as they relate to the apprentice trades, (2) identify trade alternatives that might best satisfy what they want, (3) get information about trade alternatives, and (4) get help evaluating which choice or choices might be best for them.

SEA-LECT has a brief introduction and three sections. The introduction describes the typical steps in the career decision-making process and lists the trade options that are available. The list of trades can easily be revised to reflect the options currently being offered. The first section, About You, helps users assess their work preferences as well as things they might want to avoid in their work. This section is especially useful for applicants who are not familiar with the trades and have had little or no relevant work experience. It should also help to broaden the options of someone who has focused on one particular trade because he/she is unaware of important similarities among the trades. So, for example, a person who really likes troubleshooting might consider trades as diverse as electronics mechanic or plastics molder.

The second section, About Work, gives detailed information about the trades and the training programs. The information is divided into topics so that an applicant can get as much, or as little, information as desired. This section also provides a link to videotape presentations about the trades.

The third section, Selecting, draws on users' experiences from the first two sections and gives help with evaluating and choosing from among trade alternatives. At the end of Selecting, applicants can get a printout of their evaluation of their top three choices and a rank ordering, based on their preferences, of all the trade options (that is, if a printer is available).

In SEA-LECT users have the option of starting at the beginning and going through the sections in their intended order (the recommended way), or starting in any other section they prefer.

Another important feature of SEA-LECT is that personnel at the shipyard can determine which trades should be included in the system at any time. So, for example, if one year's program does not include training for boilermakers, that trade can easily be eliminated from the listing and, therefore, from consideration by users.

SEA-LECT was designed to run on the IBM-PC with a color monitor. It's written in C language and is contained on a single floppy disk. It requires no more than 256K RAM. A printer is desirable but not required.

DEVELOPMENT OF SEA-LECT

The Introduction to SEA-LECT

There is a brief introduction to the system which tells users about the print key (F1) and the quick exit key (F10) and lists the trade options available at that time. It also describes the steps in the career decision-making process and relates them to the sections in the system.

About You - Preference Elicitation Instrument (PEI)

The PEI, which is the About You section of SEA-LECT, was designed with a highly specific purpose in mind, namely, to assess the work preferences of applicants to the Long Beach apprentice program. Since these applicants have already made the decision to work at the shipyard, their choices have already been narrowed to the trade options available at the shipyard. Consequently, rather than querying users about the world of work in general, the PEI focuses instead on the activities and conditions of work found in the trades at the Long Beach Naval Shipyard.

The PEI is taken on-line. The first part looks at preferences. Users are presented with a work feature and asked if they want it in their work. If they want it, they are shown a list of trades that have the feature; if they don't want it or are unsure, no list is presented. If users are unclear about what a feature means, they can ask for examples.

The list of features included in the first part of the PEI grew out of an analysis of the 23 Long Beach apprentice trade options. In creating the list, the intention was to (1) identify common features, those shared by several trades; (2) cover all of the major work activities for the 23 trades; and (3) make the list as short as possible. These considerations helped determine how specific or how general to make an activity. For example, "bending and shaping metal" is an important activity for sheetmetal workers while "bending and shaping insulation" is important for insulators. This led to the activity "bending and shaping materials."

The list of features that were identified fell naturally into four categories. The broadest category was WHAT IS WORKED ON OR WITH; the next broadest category was MAJOR WORK ACTIVITIES; and the least broad was SPECIFIC WORK ACTIVITIES. The fourth category included statements about HOW WORK IS DONE. The full menu includes 45 preferences.

Each trade was evaluated on all preferences. The evaluation was broad - a trade either satisfied a preference or failed to satisfy a preference.

After users see all the work features, they are shown a list of

all those they have said they want in their work. Then they are asked to rate every one according to how important it is to have that preference in their work. A three-point scale is used: 3 = very important; 2 = important; 1 = less important. From a user's ratings, the computer generates scores for the trades. An example of how scores are computed is shown in EXHIBIT A.

Users are then shown a listing of the trades in two groups - highly desirable trades (at and above their median score) and less desirable (below their median score). Within each group, trades are listed in rank order based on score.

EXHIBIT A: Example of Computation of Scores for Trades

User's important work preferences	(A) User's Ratings	(B) Ratings for Elect. Tech.	(A)x(B)
work with electronic, etc.	3	1	3
repairing, maintaining	3	1	3
producing finished article	2	0	0
doing extensive troubleshooting	2	1	2
doing complex math	1	1	1
being physically active	3	0	0
		Score for Electronics Technician is:	9

The second part of the PEI looks at work conditions and physical demands that people might want to avoid in their work. These are presented in the form of menus, with 13 work conditions and 6 physical demands. If a work condition or physical demand is selected from the menu, users are shown which trades on their lists from the first part of the PEI have that condition or physical demand. Users are given the option of removing those trades from their lists or leaving them on.

Thus, when applicants complete the About You section, they have an evaluation of all the trade options. At a later time, if they are offered a trade that is not one of their top choices, they and their counselors can refer to this evaluation to help judge the attractiveness of the trade option being offered.

At the end of the About You section users are encouraged to use the next section of the system - About Work - to find out more about those trades on their lists that appeal to them.

About Work - Information About Trades

Information was developed for three trades - Electronics Mechanic, Marine Mechanic, and Painter. The major sources of

information were job descriptions and descriptions of apprentice training programs, both of which were provided by the personnel department of the Long Beach Naval Shipyard. Additional information was obtained from the Guide to Military Occupations, from 10-minute videotapes about the trades which were produced by the shipyard, and from conversations with training program personnel.

After the information was developed it was reviewed by appropriate staff at the Long Beach Naval Shipyard and revised accordingly.

The information, which is presented on line in the About Work section, is divided into topics. These topics are presented to the user as a menu so that they can choose which ones they want to see. One of the topics, Training Program, is further subdivided into two major categories - Classroom Instruction and Work Experience. In reviewing these topics, the shipyard asked to have one added that dealt with the level of math aptitude demanded by the trade. Unfortunately, that information was not available in time for this project.

Selecting

The last section of SEA-LECT, which is called Selecting, has users choose three trades of interest to them. Having selected three trades, users are then asked to evaluate them in terms of the rewards they offer (e.g., salary, work activities) and the chances that they can meet the demands of the trades they're considering. Finally, users select one of their three choices as a possible best choice. If a user selects a trade that is clearly not a best choice (based on his/her self-assessments) the computer responds with an appropriate message.

Because shipyard personnel felt that it was important to highlight the importance of high math aptitude for some trades, the computer reminds users to take math ability into account when they assess their chances (i.e., for high math aptitude trades).

When users complete SEA-LECT they can get a copy of the screen that summarizes their assessments of the three trades they were considering and a copy of how all the trades were ranked (on the basis of their ratings in the About You section).

CONCLUSIONS

The demonstration version of SEA-LECT was well received by shipyard personnel. The concept of the system was deemed appropriate; the database was considered relevant, though all agreed that the information needed a thorough review if and when a decision was made to complete the system; and the design was found to be attractive. It was generally felt that applicants to the apprentice

training program would find SEA-LECT both interesting and useful, though there was general agreement that the system should be field tested with applicants.

However, there were two major concerns - time and hardware. Since several hundred applicants typically attend the Job Fair, how can this large group be accommodated? For example, for each of 300 applicants to spend half an hour going through SEA-LECT requires two full days and 10 microcomputers. One solution is to screen applicants and decide which ones might best profit from using SEA-LECT. Applicants with high test scores who know which trades they want to be considered for might not need SEA-LECT. On the other hand, applicants with low test scores who aspire to trades that are not likely to be open to them would be good candidates to use the system. Another possibility is to advise some applicants to use only part of SEA-LECT, perhaps just the About Work section.

In any case, concerns about time and hardware need to be addressed. If these concerns are insurmountable, a solution might lie with changing the delivery mode of SEA-LECT from a microcomputer to a pencil and paper workbook.

Developing a Microcomputer-based Assignment System for Shipyard Apprentices [1]

by
Joyce D. Mattson

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

Overview

The annual task of matching apprentice candidates to a large number of Navy shipyard trades has been a one-at-a-time manual process. This paper describes the development of a computerized optimal assignment system which considers information about all apprentice candidates at a shipyard simultaneously to generate a job specialty recommendation for each individual. The system applies a sophisticated mathematical assignment algorithm, yet is user-friendly and operates within the constraints of an IBM-XT personal computer.

This paper will describe: (1) the mathematical basis for the system, (2) its operation from a user's perspective, and (3) the quality of its assignments when compared with the present manual assignment system.

Background

The military services have utilized large-scale mainframe-based assignment systems since the late 1960's to optimally allocate their entry-level enlisted personnel to training programs. These assignment systems have typically considered such factors as the cost of personnel moves, prediction of school success, prediction of service tenure, preferences, and job quotas and have been uniformly successful in improving personnel assignments compared with the previously-utilized manual methods (Hendrix, Ward, Pina & Haney, 1979; Kroeker & Folchi, 1984; Kroeker & Rafacz, 1983; Schmitz, Nord & McWhite, 1984; Schmitz & McWhite, 1986; Ward, Haney, Hendrix & Pina, 1978).

The shipyards' personnel allocation problems are similar to the military's, but on a smaller scale. Assignments are made by a manual system, which begins with administration of an aptitude battery to the several thousand individuals applying for 50-300 vacancies at a particular shipyard.

Individuals passing the battery are rank-ordered on a "register" based on a combination of their test scores and veteran's preference status. Individuals toward the top of the register receive job information, express their

---

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 author would like to express her appreciation to Drs. Jeffery Kennington and Richard Helgason (who developed the assignment algorithm and a portion of the system software), to Carol Chatfield, who developed the remainder of the software, and to Glenn Gustitus, who aided in the analyses.

trade preferences, and are interviewed to assess their suitability and motivation for different trades. Assignments are made sequentially, starting at the top of the register, until all vacancies are filled.

Despite some successful placements using this system, there are several problems: (1) candidates part-way down the register have restricted job choices and often know little about the trades which actually become available to them, (2) there is no systematic matching of ability levels to job requirements, and (3) veterans' status accounts for a disproportionate share of the placement decision.

The system described in this paper ameliorates these problems by optimizing placements across all individuals simultaneously using raw rather than veteran's-preference-adjusted aptitude scores and by assigning equal weight to preferences regardless of an individual's position on the register.

Mathematical basis for the system

Three principle steps, which will be described in more detail, were taken to develop the mathematical basis for the computerized optimal assignment system:

Step 1. The variables used to determine the quality of different person/trade matches were selected and measured.

Step 2. An objective cost function (OCF) was developed to combine these variables and compute a single "cost"[2] for each person/ trade combination , and

Step 3. An assignment algorithm was selected to extract the set of assignments with minimum overall cost across people.

Step 1: Assignment variables. Three elements were used to determine the quality of each person/trade combination:

(1) The fit of the individual's overall ability to the overall ability requirements of the trade.

(2) The fit of the individual's mathematics ability to the mathematics ability requirements of the trade.

(3) The individual's preferences.

Step 2: Objective cost function. A separate objective cost function (OCF) was developed for each assignment variable, and the three functions were weighted and combined to compute a single cost for each person/trade combination. The functions were based on subject matter experts' (SMEs') judgments since empirical validity information was not available.

---

2 "Cost" is used in a general psychological rather than a monetary sense. A good assignment has a low cost, while a poor assignment has a high cost.

To develop the overall ability/overall difficulty OCF, the overall difficulty of each trade and the overall ability of each candidate were quantified, and an equation was formulated attaching costs to different combinations of the two. Trade difficulty was quantified by asking SMEs at the 8 Naval shipyards to rank-order trades in terms of their overall score requirements on the aptitude selection battery and then converting the mean rank-order for each trade to a normalized z-score. Intraclass correlations in the .95 to .98 range across raters characterized the mean rank-orders. Each candidate's overall ability was likewise quantified by converting his/her total score on the aptitude selection battery to a normalized z-score relative to other applicants. The objective cost function for the overall ability/overall difficulty combination was then the difference between these two z-scores.

An identical procedure was used to derive the math ability/math difficulty OCF, except that math ability was measured by the algebra/ math reasoning subtest of the aptitude selection battery and math difficulty z-scores were based on SME judgments of the mathematics requirements of the trades.

The OCF for preferences was the logarithm of the preference level truncated to a value of 1.00 for 10th or less preferred choices and to 1.25 for assignments the individual would not accept.

The three OCFs were then weighted by policy-makers at Long Beach Naval Shipyard. The weights differed for math-intensive versus other trades and depending on which of the 3 objective cost components were to be used for a particular assignment application. For the empirical comparisons in this paper, the overall ability, math ability and preference fit functions were weighted 36, 24, and 40, respectively, for math-relevant trades and 50, 0, and 50 for other trades. These weights were applied in the formula below to yield a final objective cost value for each person/trade combination.

$$\sqrt{\left(\frac{w_1 d_1}{SD_1}\right)^2 + \left(\frac{w_2 d_2}{SD_2}\right)^2 + \left(\frac{w_3 d_3}{SD_3}\right)^2}$$

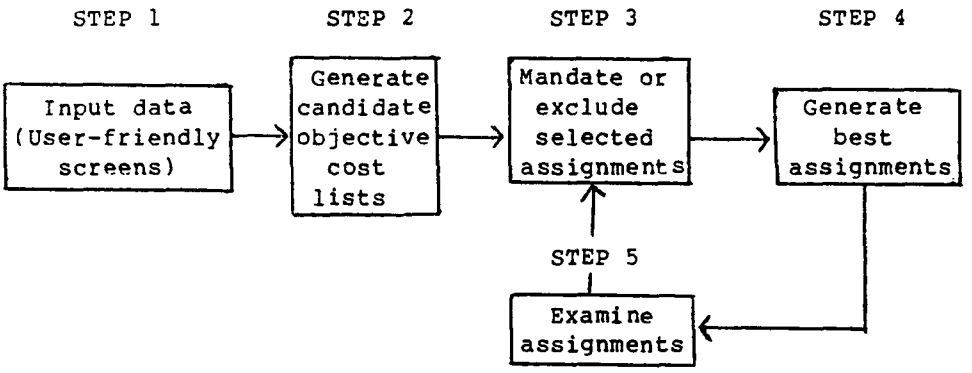
(In this formula: w_1 ,w_2, and w_3 are the weights for the overall, math and preference components, respectively; d_1,d_2, and d_3 are the OCFs (i.e., z-score differences) for overall, math, and preference fits, and SD_1,SD_2, and SD_3 are the standard deviations of d_1, d_2, and d_3 across all candidates and possible trade assignments.)

Step 3: Assignment algorithm. A linear network model assignment algorithm developed by Kennington and Helgason (1980) was chosen to efficiently select the set of person/trade assignments with minimum overall cost.

The system from a user's perspective

Figure 1 outlines the system from the user's perspective.

Figure 1
System from the user's perspective



In Step 1, the user enters trade quotas as well as ability, preference, and identifying information for apprentice candidates. This information is checked, and the costs of all possible assignments for an individual or for the total group are listed (Step 2).

In Step 3, the user mandates or excludes particular assignments -- for instance if someone lacks the physical qualifications for a particular trade -- and generates a recommended assignment for each individual (Step 4).

These assignments are examined in Step 5, and the user interviews or otherwise screens candidates for placements unsuitable for non-algorithm-related reasons.

Steps 3 through 5 can be repeated until an acceptable set of assignments is found. At each iteration, acceptable assignments from previous iterations are retained and unacceptable assignments are excluded. Each new set of recommendations thus reflects a shift in only the previously unacceptable assignments.

As is evident from this description, there is considerable flexibility in the degree of user versus computer control of the assignment process. On the one hand, assignments can be made almost entirely by computer if the full system is used. On the other hand, computer involvement can be terminated at Step 2, and the cost values generated can be used as input to manual assignments. Options are also provided for basing assignments on the overall ability fit alone or for combining it with math fit and/or preferences.

Preliminary empirical results

To determine the system's ability to satisfy preferences and SME judgments of trade ability requirements, 143 individuals operationally assigned under the manual system were reassigned using the computer.

Table 1 shows the degree to which preferred assignments were granted in each case. Thus, 53% of individuals received their first-choice trade under computerized assignment compared with only 31% under manual assignment. This

increase of 22% is matched by a corresponding decrease in individuals receiving their fifth or lower choice assignments using computer methods.

Table 1
Level of preference for assigned trade
under manual versus computer assignment

Preference Level	Percent Assigned	
	Manual	Computer
1st choice	31	53
2nd choice	20	22
3rd choice	11	13
4th choice	8	6
5th - 10th choice	17	5
Below 10th choice	3	1
Would not accept	6	1
No preference expressed	5	1

Preferences were thus much more likely to be satisfied using computer methods, even when weighted only 40%-50% in the OCF. These results were mirrored in additional preference analyses, where 32% of individuals received identical assignments under the two systems, 51% received more preference-consistent computer assignments, and only 17% received less preference-consistent computer assignments. Individuals whose preference fits improved with computer assignment had slightly higher mean overall ability and math scores than those whose fits degraded.

Computer methods also improved the fit of applicant's abilities to rated job requirements, as indicated by the correlations between SMEs' desired rank-order of trade ability means and the actual rank-order of those means. (See Table 2.)

Table 2
Rank-difference correlations between
actual and desired ordering of trade ability score means

Type of Ability	Assignment Method	
	Manual	Computer
Overall	.58	.77
Math	.79	.91

Thus, there was a rank-difference correlation of .77 between the desired and actual rank-ordering of overall ability score means using computer methods and of only .58 using manual methods. Correlations for math ability were .91 and .79, respectively. These comparisons are meaningful, however, only if the

SME judgments on which they are based relate to useful external criteria such as school or trade performance.

Summary

Results of this work indicate the feasibility of developing a user-friendly microcomputer-based apprentice assignment system that improves the placement of apprentices both in terms of satisfying preferences and in terms of generating the desired alignment of abilities with ability requirements. Further work should be directed toward providing an empirical validity-related basis for the components of the OCF and toward validating the system's utility against external criteria such as performance and tenure.

REFERENCES

Hendrix, W. H., Ward, J. H., Pina, M., & Haney, D. L. Pre-enlistment person-job match system (AFHRL TR 79-29). Brooks Air Force Base, TX: Human Resources Laboratory, September 1979.

Kennington, J. L., & Helgason, R. V. (1980). Algorithms for Network Programming. New York: Wiley and Sons.

Kroeker, L., & Folchi, J. Classification and assignment within PRIDE (CLASP) system: development and evaluation of an attrition component (NPRDC Tech.Rep. 84-40). San Diego: Navy Personnel Research and Development Center, May 1984.

Kroeker, L., & Rafacz, B. A. Classification and assignment within PRIDE (CLASP): a recruit assignment model (NPRDC Tech. Rep. 84-9). San Diego: Navy Personnel Research and Development Center, November 1983.

Schmitz, E. J., Nord, R. D., & McWhite, P. B. Development of the Army's Enlisted Personnel Allocation System (TR). Alexandria, VA: U.S. Army Research Institute, August 1984.

Schmitz, E. J., & McWhite, P. B. Evaluating the benefits and costs of the Enlisted Personnel Allocation System (EPAS) (MPPRG WP86-15). Alexandria, VA: U.S. Army Research Institute, April 1986.

Ward, J. H., Haney, D. L., Hendrix, W. H., & Pina, M. Assignment procedures in the Air Force procurement management information system (AFHRL TR 78-30). Brooks Air Force Base, TX: Human Resources Laboratory, July 1978.

A TESTING TIME FOR NAVAL OFFICERS

ALAN JONES, SENIOR PSYCHOLOGIST (NAVAL), UNITED KINGDOM

Charles Spearman

Charles Spearman is best known for his work in mental measurement and the application of statistics (in particular the origination of factor analysis) to the intercorrelations of test scores. Around the turn of the century, he noted that all tests of mental ability tended to correlate positively with one another, suggesting that there was some underlying general ability running through them all. His two-factor theory emphasised a "g" or general ability factor present in all intellectual activities and a number of s or specific factors unique to each test.

As many readers will know the two - factor theory was criticised by other research workers such as E.L.Thorndike and Thurstone in the USA and Thomson and Burt in the UK. Other theories based on group factors (eg verbal, numerical) were put forward and Thurstone discarded the notion of general ability and proposed a number of common factors (Verbal, Inductive reasoning, Spatial, Numerical etc). However, even his critics recognised his contribution. For example, Guilford (1936) wrote "No single event in the history of mental testing has proved to be of such momentous importance as Spearman's proposal of his famous two-factor theory in 1904" (page 459).

The Admiralty

In the Summer of 1919, Spearman and the Admiralty began exploring the potential usefulness of psychometric tests in the selection of career Royal Navy officers. The Admiralty was aware of the developments in psychological testing in education and, in particular, of the application of the US Army Alpha and Beta tests. The perceived benefits of testing were the rapid classification "according to natural ability, irrespective of knowledge and standard of education, provided candidates are acquainted with reading, writing and arithmetic" (unless otherwise indicated all quotations and references are from the relevant Admiralty files). Without yet again going over the nature versus nurture debate, it is obvious that being acquainted with arithmetic and reading and writing in English must be dependent on previous experience and education. What was probably meant was that the tests gave an evaluation of relative "natural ability" amongst individuals from a broadly similar background who had similar educational opportunities and attainments. Spearman's tests also had the advantage that they could be applied to groups in a relatively short time (around 45 minutes for the testing session).

The relevant Admiralty authorities were keen to try out the tests and realised that the value of the tests depended on their relationship with relevant criteria (eg "the eventual classification by trainers"): "I hope that we shall have plenty of data upon which to form a judgement", one officer wrote.

Spearman, not surprisingly, showed an overwhelming interest in measuring the general ability of naval officers. The Admiralty was also interested in a test of aptitude for higher mathematics (for example, in relation to the Gunnery and Torpedo Course). Spearman advised against developing such a test and argued that mathematics problems couched in highly technical terms were largely dependent upon specific aptitudes, while success in non-

technical mathematical questions depended upon general ability. It was therefore agreed that the focus would be on general ability.

The Tests

Spearman's tests were highly speeded and all were verbal. There were no non-verbal (eg matrices) or mechanical items, although 4 items involved arithmetic and 2 had spatial elements. The eight tests are listed below with examples. Some of the tests can be found in Burt (1921).

DIRECTIONS: 12 items, 5 minutes eg. At the end of this sentence write a word contrary to "happy". (Similar to Alpha test 1. See Burt pages 242-245). One item had a spatial element and 2 involved arithmetic.

OPPOSITES: 50 items, 2½ minutes eg. Write the opposite of "shut" (The first 50 questions from Burt, pages 237-238).

ANALOGIES: 25 items, 2½ minutes eg. SAILOR is to SOLDIER as NAVY is to ... (The first 25 questions from Burt pages 238-240). Burt had in fact originated this format around 1910 (Hearnshaw, 1979).

SENTENCE COMPLETION: 50 items, 5 minutes eg One dog can ... a flock of sheep.

RELATIONS: 10 items, 5 minutes eg. A is larger than B, and C is larger than D. What does this prove about the size of A as compared with D?

Two of the items involved arithmetic and one might be loosely described as spatial.

CATEGORISATION: 12 items, 5 minutes eg. Something that both clergyman and shepherds are, but not miners or bakers.

ANSWERING QUESTIONS ON A PASSAGE:

9 items (maximum 3 points per answer), 5 minutes. A passage followed by a question such as: What is the relation taken to exist between freedom of language and simplicity of expression?

PASSAGE COMPLETION: 41 items, 5 minutes eg.... I once ... a student to do his utmost to be popular in the social ... (See Burt pages 245-247).

All items were open-ended and had to be scored by Spearman or his co-workers, unlike some of the Army Alpha tests which were multiple choice.

Arrangements were made for Spearman to administer his tests as follows:

a. All cadets (aged 13) entering the Royal Naval College at Osborne in September 1919 (N = 39). In the event, the enthusiasm of the Osborne authorities resulted in 74 additional cadets from the previous year's entry being tested. Cadets studied for 2 years at Osborne College following a more technical education than at most private secondary schools and then went on to the Naval College at Dartmouth for a more professionally - orientated 2 years of study followed by

a training ship and experience in a ship of the Fleet before promotion to Lieutenant (age 21 to 22).

b. One class of Officers (N = 23) on the Gunnery and Torpedo course at the Royal Naval College Greenwich.

c. One class of officers on the Staff Course (N = 16) plus 9 members of the staff (March 1920). Impressed by the 1920 results the Admiralty arranged for a further class to be tested in June 1921 (N = 42 including members of staff).

The total number of cadets and officers eventually tested totalled 197, with 6 officers being tested twice. The total cost to the Admiralty appears to have been around £20 ($80).

The Results

Correlations between relevant naval officer selection and training variables and Spearman's tests are shown in the table below. It should be remembered that all three predictor variables are likely to have had direct or indirect restriction of range because of existing selection procedures; some 40 per cent of candidates appear to have passed the interview, of whom most (85 per cent) passed the Qualifying Examination.

PREDICTOR-CRITERION RANK DIFFERENCE (Rho) CORRELATIONS

		PREDICTORS		
CRITERION	N	QUALIFYING EXAMINATION	SELECTION INTERVIEW	SPEARMAN'S TEST
a. Osborne Scholastic Order of Merit (December 1919): Regular marking plus end of term examination.	97	0.52	0.22	0.40
b. Osborne Overall Order of Merit (January 1920): Likelihood of making a good officer.	97	0.27	0.20	0.29
c. Greenwich Gunnery and Torpedo Course: Preliminary Examination Order to Merit.	21	-	-	0.33
d. Staff College (March 1920) Intellectual Order of Merit: Mean of 5 assessors.	16	-	-	0.63
e. Staff College (June 1921) Intellectual Order of Merit: Mean of 7 Assessors.	30 to 40	-	-	0.29

The Osborne results show that the tests had a moderate correlation with the scholastic order, but that this correlation was somewhat lower than that for the Qualifing Examination. However, the author has carried out a multiple regression analysis (assuming the Rho coefficients to be roughly comparable to product moment coefficients) and found that the tests did add to the level of prediction offered by the Examination (R = 0.57). The interview did not add significantly to the multiple correlation. The tests were the best predictors of the overall order (0.29) and neither the interview nor the Examination adds significantly to this correlation. These figures might, of course, vary if correlations corrected for range restriction were used. Spearman considered that all three predictors yielded information about candidates. Combining the three sources of information would "be of great assistance not only in selecting them originally but in regulating their treatment subsequently".

The Long Torpedo and Gunnery Course was intended for Lieutenants. Spearman's tests were correlated with the examination at the end of the first 3 months, which was intended to go over and reinforce technical subjects encountered previously. Spearman may have been disappointed with the correlation of 0.33, but he argued that technical examinations depended much less on general ability and more on specific abilities. Since he did not believe in group factors, he may have considered that such abilities were too specific for meaningful psychological measurement.

The RN Staff College had resumed in Greenwich in June 1919. Its students were typically Lieutenant Commanders and Commanders. Marder (1961) notes that it was not too successful in its first years, partly because those officers chosen were sometimes of below average ability. This may have been one of the reasons behind the Admiralty's having Spearman administer his tests to Officers on the course. He was obviously very happy with the obtained correlation of 0.63 for the first sample, although he cautioned about the small sample size. Spearman was also concerned about two outstanding non-corresponding cases and an investigation was made; an officer who had performed unexpectedly badly on the tests had been in some distress because of personal problems (this case is discussed in Spearman, 1927, pages 337-338) and one who showed up well on the tests had evidently been overlooked because of his quietness.

Those responsible for the Staff Course must have been genuinely impressed because they asked Spearman to return a year later to conduct further tests. This time the correlation was lower (0.29), which could in part be ascribed to a lower level of agreement among assessors (an average intercorrelation of 0.40 as against 0.65 for the previous year). The Director of the Staff College concluded that the "tests measure something, and very possibly they measure that something with comparative accuracy, but I doubt if they can be taken as an accurate indication of total mental ability. If an officer stands high in these tests it probably indicates that, in certain directions, he possesses considerable mental ability".

What Happened Next

After the 1919-20 trials, the Headmaster of Osborne and the Admiralty personnel who had been involved met to consider the results. All were agreed on the value of Spearman's tests and that they should be implemented to assist the Interview Board in selecting cadets for Osborne. However, the potential problem of test security was raised; it was considered that school teachers would do virtually anything to secure

copies of tests in order to "cram" their pupils. Therefore, the tests might have to be varied every year.

However, subsequent discussions within the Admiralty showed resistance to the introduction of tests. There was concern that parents and schoolteachers would not find such a "secretive" method of assessment (ie no published syllabus) acceptable, and that, in fact, schoolteachers would find a way around test security. Another objection was that the tests did not measure general ability, but the test protagonists could point to the obtained correlations between the tests and various criteria, whether or not the tests were measuring "general ability".

The proposal to use Spearman's tests was therefore dropped but was resurrected again 3 years later, along with a suggestion for a more structured interview schedule and a Competitive (rather than a Qualifying) Examination. Once again, objections were made against Spearman's claims that the tests measured "natural intelligence". Whatever the influence of these arguments, however, it appears that the major argument against revising the seelction system was that there were hardly any applicants to select from!

The Royal Navy undertook no further experiements with psychometric tests in the period between the two World Wars. In 1942 the RN began using tests to help select the Reserve (RNVR) officers who formed the bulk of the wartime officer corps. However, by now the model of ability in use was that of Burt (see Hearnshaw, 1979) and Vernon rather than that of Spearman; there were now group factors (falling into two main groups: verbal-numerical-educational and practical-mechanical-spatial-physical).

The test batteries used were correspondingly different from the verbally-dominated tests of Spearman; mechanical comprehension, mathematics, arithmetic and spatial ability were included in the basic test battery (validity of around 0.54 with initial officer training), while other experimental tests (clerical instructions and a verbal orientation test) showed promising validities - Vernon and Parry (1949).

Spearman's results from Osborne were, however, used when the Admiralty was considering how the selection procedure for career officers should be organised. Sixteen of 39 first-term cadets tested by Spearman were still serving as Seaman Officers in October 1944 (25 years later) and had complete test and Osborne assessment data available. Obviously such a small sample is unlikely to produce a reliable estimate but a tetrachoric correlation of 0.68 was obtained between the tests and having been promoted Commander; of the 8 top test scorers, 6 had been promoted Commander as against 4 of the bottom 8. The author has recalculated the correlation with an 8/8 split (rather than a 10/6 split) and obtained a correlation of 0.40.

The eventual selection procedure adopted in 1947 was what is today called an assessment centre (see Jones 1984). Psychometric tests are only one of the inputs to the final assessment of officer potential. The underlying approach to psychometrics has remained that of Vernon (1950) with the following tests now in use: Verbal Reasoning (same and opposite, analogies, jumbled sentences, completing sentences), Non-Verbal Reasoning (diagrammatic sequences and matrices), Clerical Instructions and Numeracy (numerical facility, numerical reasoning, and statistical interpretation). A recent study showed a correlation of 0.45 between this battery and professional examinations at the end of the first stage of officer training (N = 657).

An Evaluation

Looking back over the 67 years, we can evaluate what Spearman had achieved. He had shown that psychometric tests could be easily applied to groups of personnel in the relevant ability range. The tests had correlated with a number of important criteria; cadet's scholastic performance (0.40), Gunnery and Torpedo Course preliminary examination (0.33) and the assessed intellectual ability of officers on the Staff Course (0.63 & 0.29). As far as can be ascertained, all the naval officers directly involved with the research had been impressed by the tests and the results obtained. However, he can be criticised for an exclusive concern with general ability and then claiming to measure it adequately by verbal tests alone. This perhaps led him to explain away lower correlations with more technical criteria, and to resist more specific aptitude tests or even work sample tests which have since proved very useful for Royal Navy selection. Nevertheless, his pioneering work within the Navy and in other organisations prepared the way for the widespread use of psychometric tests in the British Forces in World War 2 and subsequently. Spearman's place as a pioneer in military testing, as well as psychometrics generally, should therefore be warmly and gratefully acknowledged.

REFERENCES

1. BURT C — Mental & Scholastic Tests. London, Staples Press, 1921. (See in particular pages 233-247)

2. HEARNSHAW L S — Cyril Burt: Psychologist. London, Hodder & Stoughton, 1979. (See in particular pages 46-71, 87-95 & 154-181)

3. GUILDFORD J P — Psychometric Methods. New York, McGraw-Hill, 1936

4. JONES A — Royal Navy Officer Selection: Developments, Current Procedures & Research. Paper presented at the 26th MTA Conference, Munich, 1984

5. MARDER A J — From the Dreadnought to Scapa Flow: Volume I The Road To War 1904-1914. London, Oxford University Press 1961. (See in particular pages 28-33, 46-52 & 265-266).

6. SPEARMAN C — The Abilities of Man: Their Nature & Measurement. London, MacMillan, 1927

7. VERNON P E — The Structure of Human Abilities. London, Methuen, 1950

8. VERNON P E & PARRY J B — Personnel Selection in the British Forces. London, University of London Press, 1949. (See in particular pages 34-35, 232-233 & 240).

REVISION OF PSYCHOLOGICAL SERVICE OF THE FEDERAL ARMED FORCES

Peter W. Mademann, Federal Armed Forces Recruiting Office, Hamburg
Klaus J. Puzicha, Federal Armed Forces Administration Office, Bonn

At present, 900.000 soldiers from seven allied NATO nations are stationed in the Federal Republic of Germany, a country about the size of Colorado, U.S.A.
Naturally, in the joint efforts of defence, the armed forces of the Federal Republic have been assigned a central task in this region. The numerical strength of the German Federal Armed Forces is 495.000 men, as agreed in the NATO treaties. The majority is made up of Army units totalling 340.000 soldiers. The Air Force counts 110.000, and there are 38.000 service men performing their military duty in the Navy. The remaining 6.000 soldiers are recruited from reservists drafted for military training over relatively short periods.
In the Federal Republic of Germany military service is compulsory, and consequently approximately 45 % of the standing German Armed Forces consist of conscipts who, after their vocational education, serve their 15-month-term and then return to their normal occupation. Professional knowledge already acquired can thus be used to some extent for military activities.

Approximately 42 % have signed up voluntarily for a certain period of time (between two and 15 years, the majority between four and eight years). Only 13 % are professional soldiers. The share of conscripts amounts to 52.0 % in the Army, 32.0 % in the Air Force, and 22.0 % in the Navy. The differing shares of conscripted soldiers in the three branches of the armed forces is due to the different degrees of technicalization and specialization requiring more or less special educational backgrounds.
Altogether, the armed forces have a yearly staff recruitment demand of 225.000 men.
Approximately 100 out of a total of 140 psychologists of the Federal Armed Forces are dealing with problems of staff replacement, selection or classification on the basis of psychological tests and other sources of information. The rest is working on human engineering problems, special problems of aviation psychology, clinical psychological cases or fundamental problems of military psychology.
Only five psychologists are working full-time on basic and applied research (see annex 1).
The emphasis of psychological activities in the Federal Armed Forces is centered on the field of personnel psychology in form of selection and classification.
Test figures from 1985, revealing that 380.000 young men including 36.000 volunteers and 9.200 officer candidates had to undergo psychological test procedures, clearly prove the eminent place value of personnel psychology in the Federal Armed Forces. This is not surprising, if you consider that almost one half of the armed forces is exchanged year by year and suitable replacements must be found. In the past it was more or less easy to solve this question adequately; because of the relatively high potential of young men we were able to pick the elite in most cases.

But, in the future the military psychology will be forced to leave beaten paths and to reconsider its so for successful conception.
The development and distribution of hormonal ovulation inhibitors in the late sixties caused a dramatic decrease in population development which reduced the birth rate to one half within one decade.
The 1975 age class to be drafted into military service in 1994 merely reaches a strength of 250.000 which is only 10 % more than the total number of personnel required by the armed forces (annex 2).
Various legislative measures such as the extension of the service period to 18 months or stricter medical classification criteria are to ensure that there will be a sufficient supply of men to meet the quantitative demand of the armed forces in the nineties. The future will show whether this scheme will work out.
However, there is still no solution for the problem of meeting the demand in qualitative respects.
A basic reorientation of the Psychological Service of the Federal Armed Forces is necessary in order to meet the qualitative requirements of the armed forces at least to some extent.
The elite selection system practiced so far will be substituted by a classification procedure with rather demands on the quality of psychological diagnostics. Henceforth, it must be the ultimate aim of qualification diagnostics to promote the highest possible efficiency of the individual service man at his military job.
In times of relative abundance the traditional aptitude tests were based primarely on existing or acquired abilities, skills, and knowledge in order to ascertain fitness for a military assignment. Determinants having a significant bearing on diagnosing the efficiency of military personnel at its job are, however, certainly more complex and should be included in the psychodiagnostic process in future.
They comprise:
. skills and knowledge to be acquired during the military training including training and learning motivation,
. attitudes towards the Federal Armed Forces, especially the satisfaction soldiers find in their job,
. a precise and constantly revised knowledge of the requirements of military jobs.

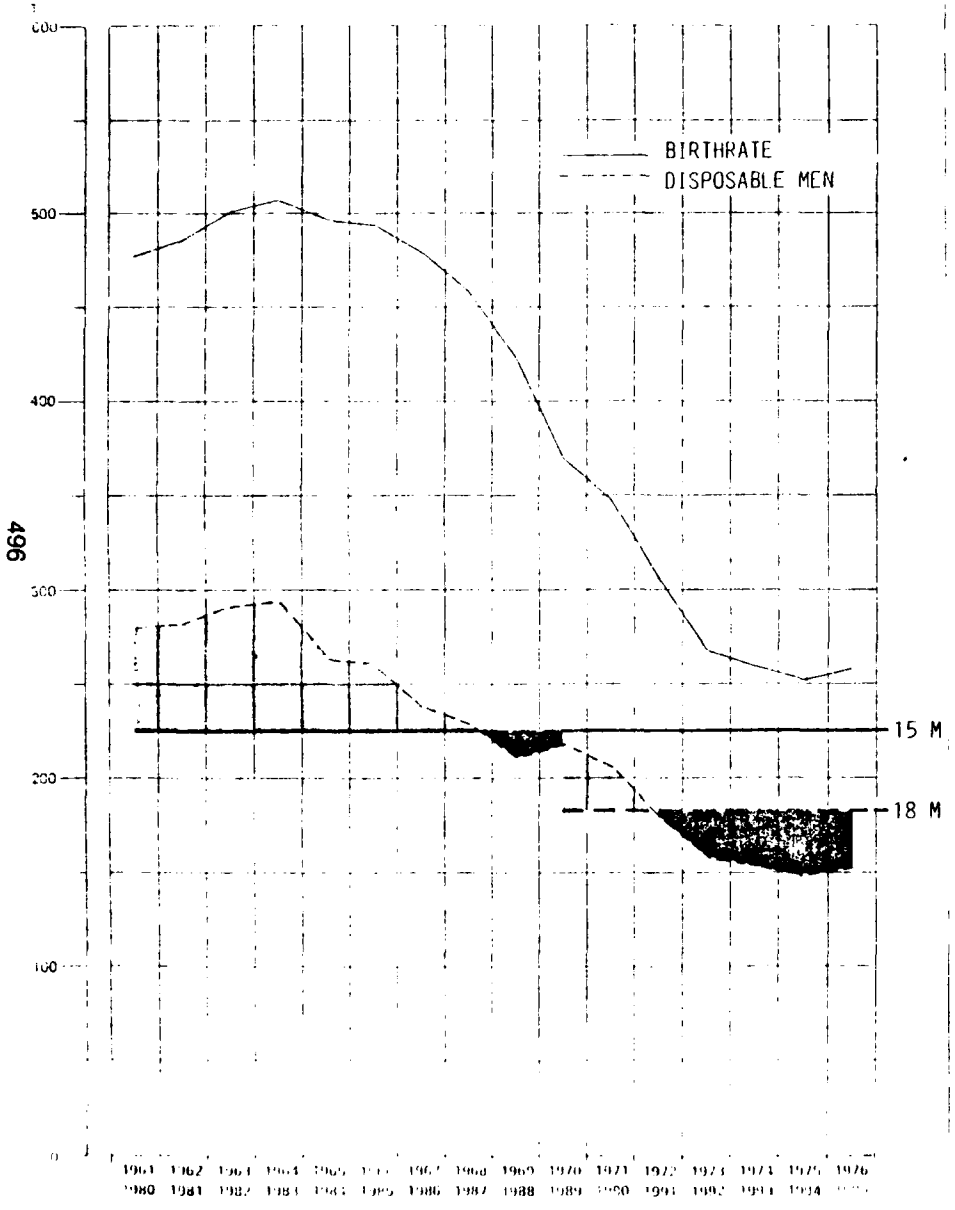
Moreover, we have to consider the efficiency characteristics of military equipment as well as attitudes and motivations influenced by the Federal Armed Forces as additional determinants relating to the efficiency of soldiers (annex 3).
The priority task of psychology in the Federal Armed Forces in order to overcome difficulties in meeting qualitative demands is to decisively increase the quality of the prognoses by more accurate and improved evaluations. This calls for extensive analyses of training and military activity, adequate measuring instruments as well as the validation of applied methods according to characteristics and criteria.
These technical reasons necessitate a considerable increase of personnel in the field of method development and method control.
The second stage plans the extension of psychological methods to a broader basis. The traditional status diagnostics based on aptitudes shall be improved by aspects which consider the learning and development processes, social competence and motivation as well as certain behaviorial habits and lead to process diagnostics based on behavior and decision.
The military psychologists pay considerable attention to the further developments of the assessment center technique and simulation-based qualification

assessment procedures.
According to the present conception of the Psychological Service of the Federal Armed Forces, qualification prognoses are made in every case prior to joining the armed forces. In the case of soldiers signing up for a service-time of up to 15 years, the prognosis covers a period which is unduly long. The skills and the knowledge but especially the motivation of young 18 - 19 year-old men are hardly stable so that long-term prognoses contain a number of uncertain factors which do not allow a sufficiently reliable statement. It is impossible to consider certain personality developments or events in their social surroundings influencing the efficiency or willingness of the soldiers.
So far, after joining the armed forces no psychological interventions or assessments are being made, except in the case of the flying personnel or soldiers transferred to military hospitals on account of devaint behaviour.

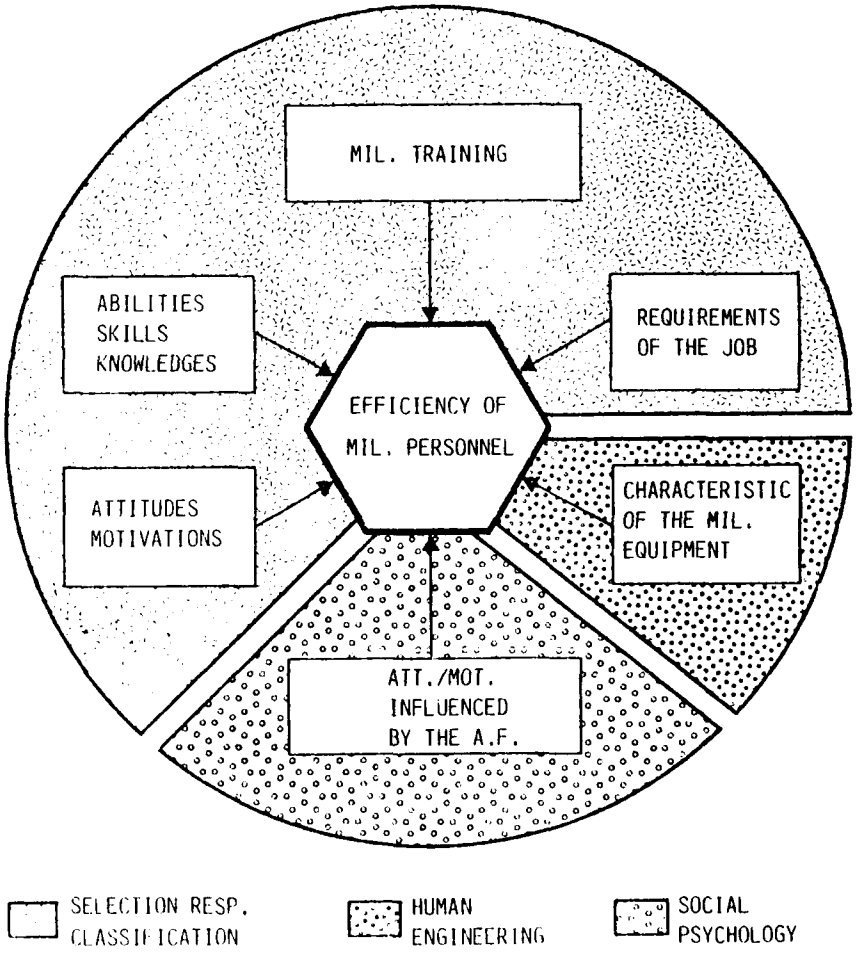
This conception of applying psychological assessment procedures almost entirely before entering the Federal Armed Forces, is based on the historic development of military psychology and also on the fact that the psychologists are exclusively civilians whose competency ends at the barrack-gate.
In the meantime, there has been some reconsideration on the part of the military side and, consequently, there is a greater demand of psychological services in the army, especially with regard to personnel problems. This development offers an excellent opportunity for intensifying the cooperation with the armed forces and for applying qualification assessment procedures inside the barracks, where we can make use of additional knowledge gained by officers' judgments or by the evaluation of data relating to practical performance during military training or at the place of work.
If the soldiers' effiency at their place of work should be increased or optimized, it seems inevitable to reduce the time lag caused by the present procedure or to modify initial prognoses by taking new factors and diagnostic findings into consideration.
This, however, necessitates that psychologists work in the immediate vicinity of the military service or in the army itself in order to ensure a flexible treatment of individual pecularities.
This revised organizational structure would also offer the advantage of carrying out on-the-spot evaluations to be drawn up by a centrally operating team, in addition to the diagnostic work in the army.
The study group "Revision in the Psychological Service of the Federal Armed Forces" with seven military psychologists appointed by Dr. Ermisch, State Secretary in the Ministry of Defence, has arrived at the following conclusion during their 6-months activities in the first half of 1986.
The only way to effectively cope with the difficulties in meeting qualitative demands is an interacting system of basic development, personnel psychology, ergonomics, and organizational psychology. Since an increase of the total personnel strength of the Psychological Service was out of the question for the revision, the study group recommended to replace the traditional qualification assessment outside the barracks walls by a computerized adaptive test system administered by specially trained assistant personnel.
On the basis of fundamental diagnostics it will be decided for which military functions the recruit is definitely unsuitable. The assignment to a military unit will then be effected in accordance with the demand and the individual preferences of the draftee for whom the greatest motivation incentive is very often an assignment near his home town.

Subsequently, the psychological basic test made outside the barrack-gate is to be differentiated in the service by sequentially conducted test procedures, in order to pick out suitable soldiers equipped with the best possible prerequisites for taking over special functions, or those who are likely to complete their training successfully.
In addition to the qualification diagnostics routine, the army psychologists are to make socio-psychological assessments, such as pertaining to satisfaction with the job or organizational development, and to advise military superiors and soldiers with personal difficulties.

ANNEX 2 NUMBER OF DISPOSABLE DRAFTEES 1980 - 1995;
BIRTHRATES 1961 TO 1976;
ESTIMATED RATE OF CONSCIENTIOUS OBJECTORS: 10 PERCENT

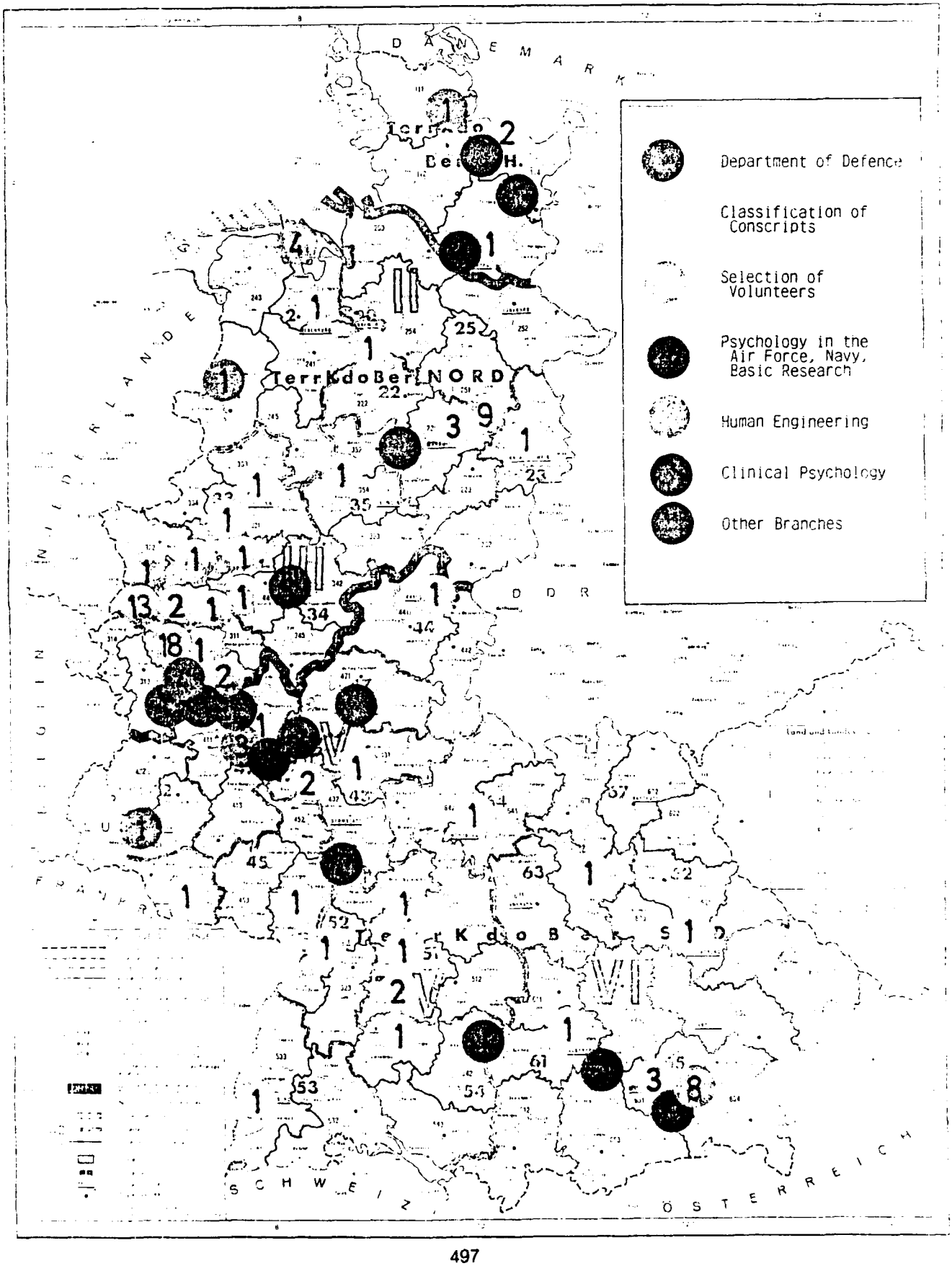


ANNEX 3 FOR THE MILITARY PSYCHOLOGY RELEVANT DETERMINANTS
FOR THE EFFICIENCY OF MILITARY PERSONNEL



ANNEX I

The Psychological Service of the German Armed Forces
- Psychological Branches, Locations, Number of Psychologists -



Does Microcomputer-Based Testing Encourage Truthful Responses?

Paul Rosenfeld, Linda Doherty, Larry Carroll, John Kantor
Navy Personnel Research and Development Center

Marie Thomas
College of Mount St. Vincent

A problem affecting the interpretation of psychological tests, surveys and questionnaires is that respondents' answers may be less than honest. Especially when the information requested is sensitive, embarrassing, or threatening, individuals often exhibit a tendency to "fake good", and misrepresent their answers in socially desirable ways (Schuman & Kalton, 1985). Because military testing frequently requires the disclosure of sensitive material, scores on military assessment instruments used for purposes such as selection and placement may be especially prone to social desirability biases.

An examination of the social psychological, personality, and organizational literatures reveals a number of examples of social desirability distortions. Gordon and Stapleton (1956) reported that high school students scored higher on several dimensions of a personality test when the results were claimed to be for a job application than when for a guidance class. Goldstein (1971) found that more than half the applicants for a nurse's aid job exaggerated their salaries and length of service at previous jobs. Giacalone and Rosenfeld (1986) gave a survey to individuals in a legislative intern program. They found that the interns inflated their self-evaluations and salary aspirations when publicly associated with the survey and the results were to be shown to their supervisor.

In dealing with social desirability response effects it is accepted that though they cannot be totally eliminated, it is possible to lessen their impact on the data (Schuman & Kalton, 1985). Among the techniques that have been offered as a means of increasing truthfulness are: a) physiological measures such as pupil dilation, galvanic skin response and facial muscle contractions, b) the bogus pipeline-- a machine which through an elaborate set of ethically questionable deceptions is portrayed as a powerful lie detector (cf., Tedeschi et al., 1985, chapter 6), and, c) the randomized response technique--a procedure in which the respondent uses randomization to determine whether a sensitive question or an innocuous one is to be answered (Warner, 1965).

While these measurement devices were specifically developed to reduce response distortions, a number of studies have shown that computerized assessment--a technique originally introduced by clinicians for economic and efficiency reasons--also reduces social desirability distortions and increases truthfulness on psychological tests, surveys, questionnaires, and structured interviews (e.g., Carr & Ghosh, 1983; Duffy & Waterston, 1984; Kiesler & Sproull, 1986). Other investigators, however, have failed to obtain increased truthfulness on the computer (Kosin, Kitchen, Kochen & Stodolosky, 1970; Lukin, Dowd, Plake, & Kraft, 1985; Skinner & Allen, 1983).

In the present paper, we consider evidence pertaining to the issue of whether computerized assessment leads to more truthful responses and present relevant data from a study using two versions of the CENSUS system--a microcomputer-based multiuser system developed at the Navy Personnel Research and Development Center in San Diego, CA.

COMPUTERIZED ASSESSMENT AND TRUTHFULNESS

A number of studies have reported greater truthfulness on computerized assessment, especially when the items require sensitive or potentially embarrassing responses. Evan and Miller (1969) had undergraduates at the Massachusetts Institute of Technology complete a questionnaire containing both sensitive, anxiety provoking items from the MMPI Lie and Manifest Anxiety Scales, and neutral, factual items for comparison purposes. As predicted, the group completing the questionnaire on a

computer terminal had higher MMPI Manifest Anxiety Scores and lower MMPI lie-scale scores. There were no differences between computer and paper-and-pencil groups for factual items. Similarly, Carr and Ghosh (1983) obtained higher scores on a computerized Fear Questionnaire than when the same items were presented on a paper-and-pencil inventory or asked in a face-to-face interview.

If, as these studies suggest, sensitivity of material is important in determining when computerized assessment will increase truthfulness, it would be expected that surveys relating to abuse of drugs and alcohol would show significant computer vs. paper and pencil differences. This view is supported by a study in which male patients were surveyed about their alcohol-related problems. As compared to a direct interview, a 30% greater admission of alcohol consumption on a computerized questionnaire was obtained (Lucas, Mullin, Luna, & McInroy, 1977). A study involving a sample in Edinburgh, Scotland found that a computerized survey led to a 33% higher admission of alcohol consumption than a face-to-face interview (Duffy & Waterston, 1984).

Kiesler and Sproull (1986) extended these results obtained in clinical settings to social science surveys and questionnaires. Users of a computer-mail system at Carnegie-Mellon University completed a health and personal characteristics questionnaire which contained five items from the Need for Approval Scale (Crowne & Marlowe, 1964) as a measure of socially desirable responses. The results indicated that there were no response differences between the computer survey and paper and pencil for attitudes toward health. However, computerized assessment did lead to a significantly lower proportion of socially desirable responses on the five items from the Need for Approval Scale.

In organizational settings, supportive evidence has also been obtained. Sproull (1986) reported that while average responses to factual questions on an electronic mail system did not substantially differ from those obtained by paper-and-pencil, those responding by computer were more likely to choose extreme responses, suggesting a tendency to be more forthright. The Wall Street Journal (Feinstein, October 9, 1986, p. 35) recently described a survey conducted at Chevron Corporation's Ortho Consumer Products unit in San Francisco, CA. Salespeople were asked their views on the company's marketing strategy on either paper-and-pencil or computerized surveys. While the paper-and-pencil respondents "had only kind words for their bosses", when the same questions were asked on a computer, "not all the responses were so favorable to management" (Feinstein, 1986, p. 35).

Several studies, however, failed to find increased honesty on computerized surveys, suggesting that additional variables may interact with computerized measurement to determine when socially desirable responding is reduced. These studies typically have demonstrated that the computer obtains equivalent responses to paper-and-pencil on psychological instruments and thus can be used as an alternative to traditional paper-and-pencil means. In a military setting, Moreno, Wetzel, McBride, and Weiss (1984) reported that a computerized version of the Armed Services Vocational Aptitude Battery (ASVAB) obtained comparable results to paper and pencil administration while using only half the number of items. Elwood and Griffen (1972) and Hedl, O'Neil and Hansen (1973) reported correlations of .95 and .75 respectively between paper and pencil and computerized administrations of the WAIS and Slossen Intelligence Tests. White, Clements and Fowler (1985) obtained equivalent mean scores and test-retest reliability on a microcomputerized version of the MMPI as on paper-and pencil. Lukin et al. (1985) found virtually identical scores for computer and paper-and-pencil administration of the Beck Depression Inventory and the State-Trait Anxiety Inventory. And, in a well-designed study at an Addiction Research Foundation, Skinner and Allen (1983) had individuals complete substance abuse questionnaires on a microcomputer, paper-and-pencil or face-to-face interview formats. Surprisingly, no differences were found between the three groups in amount of alcohol, drug and tobacco-related problems, a finding which deviates from the expected increased candor for sensitive questions on computer assessment reported for substance abuse by Lucas et al. (1977) and Duffy and Waterston (1984).

These contradictory findings suggest that a number of intervening variables which previous studies have rarely considered or controlled for may be operating (e.g., anonymity, computer sophistication, gender differences).

In addition to these explanations for the inconsistency of previous research, we presently suggest that an individual difference approach may clarify the effects of computer assessment on socially desirable responses. That is, computer assessment may lead to more honest responses in some people-- those who chronically respond in socially desirable ways in a number of situations (Snyder, 1974). Individuals who tend to be more forthright across situations would show less of an effect of computer assessment because their responses are basically truthful in both cases.

This hypothesis was tested by reanalyzing a previously gathered data set which measured job satisfaction among Navy civilian employees in critical jobs at an airplane rework facility. That study, which used two versions of CENSUS-- a microcomputer-based multiuser survey system (Rosenfeld, Doherty, Carroll, Vicino, Kantor, Thomas & Riordan, 1986), offered an opportunity to assess this individual difference hypothesis on a post-hoc basis.

METHOD

System Description

CENSUS I

CENSUS I uses existing microcomputer technology, modified for survey and questionnaire administration. The host computer is an IBM PC/XT or PC/AT configured with 640K of memory and equipped with a HOSTESS 8-port serial board that supports up to eight computer terminals. A Pinetree Multiuser System board contains the system software on ROM chips that allows communication with the host. CENSUS I uses Northern Telecom Displayphones as the terminals which are linked with the host directly, or remotely through a modem. A survey or questionnaire is first entered into the host microcomputer and can subsequently be completed by eight users simultaneously. The system allows users to be located in the same site as the host or at remote sites using commercial phone lines. Raw data are stored by the host in an ASCII file with the user's ID as the identifier. The system checks the ID number against the user file to determine its validity, and can link background information about the user with his/her responses. Only users entering valid ID numbers are allowed to take the survey.

CENSUS II

The host computer is an IBM PC/AT configured with three megabytes of memory and equipped with two CONTROL SYSTEMS 8-port serial boards. In conjunction with IBM's multiuser XENIX operating system (version 2.0), the system allows up to sixteen terminals to communicate with the host directly, or remotely through a modem. CENSUS II uses Qume QVT 211 GX terminals to link with the host. Each user works at his or her pace, being unaffected by the responses or response times of other users. CENSUS II creates files, allows user access, and links preexisting data in a similar manner to CENSUS I.

SUBJECTS

One hundred and two male civilian employees at the Naval Air Rework Facility (NARF) in San Diego, CA were randomly selected for participation in this study. The participants were individuals working in several "blue collar" job specialties within the NARF, an activity which repairs Naval aircraft.

PROCEDURE

Letters were sent to the selected individuals indicating that civilian headquarters was interested in attitudes toward critical civilian jobs for future manpower planning. Upon arrival at the testing center, participant's names and social security numbers were checked against a master list, after which they were asked to complete a survey containing a modified version of the Job Description Index (Smith, Kendall, & Hullin, 1969) as well as other questions relating to their

situation. The subjects were randomly assigned to either the paper and pencil (N= 33), CENSUS I (N= 26), or CENSUS II (N= 43) groups. Because of logistical considerations, only one of the two CENSUS systems was run on a particular day during the testing week.

RESULTS

Comparison of Assessment Modes

The results showed that nearly equivalent scores on the job satisfaction measures were obtained for the paper-and-pencil and two CENSUS groups. ANOVAs indicated that both CENSUS I and CENSUS II groups did not significantly differ from paper-and-pencil for overall job satisfaction, growth satisfaction, security satisfaction, pay satisfaction, and social satisfaction (all p's $>.20$). When the subscale scores were combined to obtain an overall measure of job satisfaction, the means for CENSUS I (M=48.79), CENSUS II (M= 49.53) and paper-and-pencil (M=48.78) groups were virtually identical. Those who completed the survey on CENSUS II rated the experience as more enjoyable (M=3.42), than those who completed the paper-and-pencil survey (M=2.76), F, (1, 74) = 4.85, $p < .04$. For CENSUS I, there were also higher enjoyment ratings (M=3.38), however, these only approached significance, F, (1, 47)=2.30, $p < .14$.

Post-Hoc Analysis

The two CENSUS systems did not differ on the job satisfaction measures (F's < 1), and were combined. A post-hoc block was created using responses to an item considered a possible index of socially desirable behavior,"For you, how important is it to do what others think you should do" (1=unimportant; 5=very important). A mean-split resulted in the creation of low and high social desirability groups. These groups were crossed with assessment mode (CENSUS vs. P & P) in a 2 X 2 Analysis of Variance with the job satisfaction measures being the dependent variables. The ANOVAs failed to reveal any main effects of social desirability or assessment mode, but significant interactions were obtained for general satisfaction, F, (1, 88) = 3.81, $p < .05$; security satisfaction, F, (1, 88) = 4.94, $p < .03$; and social satisfaction, F, (1, 88) = 5.49, $p < .02$. The interaction for the combined job satisfaction scales approached but did not reach significance, F, (1, 88) = 2.77, $p < .10$. These means are presented in Table 1.

TABLE 1
MEANS FOR SIGNIFICANT ASSESSMENT MODE BY SOCIAL DESIRABILITY INTERACTIONS

SOCIAL DESIRABILITY	Comp.	P & P	Comp.	P & P	Comp.	P & P
low	10.95	9.57	7.74	7.51	11.14	10.52
high	10.88	12.33	7.00	8.67	10.06	11.83
	General Sat		Security Sat		Social Sat	

As can be seen from Table 1, subjects in the high social desirability condition inflated their paper and pencil job satisfaction scores; a tendency reduced by computerized assessment. For individuals in the low social desirability condition, computer and paper and pencil scores did not systematically differ.

DISCUSSION

As stated previously, it is unclear based on past inconsistencies in research findings whether computerized assessment inevitably reduces socially desirable responses. The current results suggest that an individual difference approach provides a more precise determination of the effects of the computer on honest responding and thus should be considered in future military and civilian sponsored computer-based assessment. Job satisfaction scores obtained on the computer were

nearly identical to those on paper and pencil, indicating, as others have, that computerized measurement is at least as accurate as paper and pencil. Additionally, in line with previous findings, subjects reported greater enjoyment of the computer surveys.

Since the job satisfaction questions were scored in a positive direction, one would expect social desirability effects to be indicated by higher job satisfaction scores-- an effect past research suggests should be reduced by computerized assessment. This direct reduction in social desirability responding on the computer did not occur in the present study. However, post hoc ANOVAs supported the notion that response distortion may have been occurring in a subset of our sample: individuals strongly influenced by social desirability motives. The exaggeration of job satisfaction scores in these individuals was reduced by CENSUS.

Certainly, the post hoc flavor of this analysis prohibits firm conclusions, yet these results do indicate that future studies should include individual difference scales assessing levels of socially desirable behavior such as Snyder's Self-Monitoring Scale (Snyder, 1974) and the Need for Approval Scale (Crowne & Marlowe, 1964) as independent variables rather than solely as dependent measures. These further studies, if they confirm the preliminary indications of the present work, will clarify that the issue of increased honesty on computerized assessment is not so much one of if, but rather of when and among whom.

REFERENCES

Carr, A.C., & Ghosh, A. (1983). Accuracy of behavioral assessment by computer. British Journal of Psychiatry, 142, 66-70.

Crowne, D.P., & Marlow, D. (1964). The approval motive:Studies in evaluative dependence. New York:John Wiley.

Duffy, J.C., & Waterton, J.J. (1984). Under-reporting of alcohol consumption in sample surveys:The effect of computer interviewing in fieldwork. British Journal of Addiction, 79, 303-308.

Elwood, D.L., & Griffin, R.H. (1972). Individual intelligence testing without the examiner:Reliability of an automated method. Journal of Consulting and Clinical Psychology, 38, 9-14.

Evan, William, M. & Miller, James, R. (1969). Differential effects on response of computer vs. conventional administration of a social science questionnaire:An exploratory methodological experiment. Behavioral Science, 14, 216-227.

Feinstein, S. (October 9, 1986) Computers replacing interviewers for personnel and marketing tasks. Wall Street Journal, p. 35.

Giacalone, R.A., & Rosenfeld, P. (1986). Self-presentation and self-promotion in an organizational setting. Journal of Social Psychology, 126, 321-326.

Goldstein, I.L. (1971). The application blank: How honest are the responses? Journal of Applied Psychology, 55, 491-492.

Gordon, L.V., & Stapleton, E.S. (1956). Fakability of a forced-choice personality test under realistic high school employment conditions. Journal of Applied Psychology, 40, 258-262.

Hedl, J.J., O'Neil, H.F., & Hansen, D.H. (1973). Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 40, 217-222.

Kiesler, S., & Sproull, L. (1986). Response effects in the electronic survey Public

Opinion Quarterly, 50, 402-413.

Koson, D., Kitchen, C., Kochen, M., & Stodolosky, D. (1970). Psychological testing by computer:Effect on response bias. Educational and Psychological Measurement, 30, 803-810.

Lucas, R.W., Mullin, P.J., Luna, C.B.X., & McInroy, D.C. (1977). Psychiatrists and a computer as interrogators of patients with alcohol-related illnesses:A comparison. British Journal of Psychiatry, 131, 160-167.

Lukin, Mark, E., Dowd, E. Thomas, Plake, Barbara, S., & Kraft, Robert G. (1985). Comparing computerized versus traditional psychological assessment. Computers in Human Behavior, 1, 49-58.

Moreno, K. E., Wetzel, C.D., McBride, J.R., & Weiss, D. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. Applied Psychological Measurement, 8, 155-163.

Rosenfeld, P., Doherty, L., Carroll, L., Vicino, M., Kantor, J., Thomas, M. & Riordan, C. (1986). Assessing job satisfaction on two microcomputer-based multiuser systems. Unpublished Manuscript, Navy Personnel Research and Development Center, San Diego, CA.

Schuman, H., & Kalton, G.. (1985). Survey methods. Chapter 12 in G. Lindzey and E. Aronson (Eds.). Handbook of Social Psychology (Volume 1), New York: Random House.

Skinner, Harvey, A., & Allen, Barbara, A. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug and tobacco use. Journal of Consulting and Clinical Psychology, 51, 267-275.

Smith, P.C., Kendall, L.M., & Hullin, C.L. (1969). The measurement of satisfaction in work and retirement:A strategy for the study of attitudes. Chicago, IL:Rand McNally.

Snyder, M. (1974). Self-monitoring of expressive behavior. Journal of Personality and Social Psychology, 30, 526-537.

Sproull, L.S. (1986). Using electronic mail for data collection in organizational research. Academy of Management Journal, 29, 159-169.

Tedeschi, J.T., Lindskold, S., & Rosenfeld, P. (1985). Introduction to social psychology. St. Paul, MN:West.

Warner, S.L. (1965). Randomized responses:A survey technique for eliminating error answer bias. Journal of the American Statistical Association, 60, 63-69.

White, D. M., Clements, C.B., & Fowler, R.D. (1985). A comparison of computer administration with standard administration of the MMPI. Computers in Human Behavior, 1, 153-162.

The opinions expressed herein are those of the authors, they are not official, and do not necessarily reflect the views of the Navy Department. The authors are grateful to Mitch Vicino, Marlean Free and Catherine Riordan for their contributions to Project CENSUS.

USING A SHIPHANDLING SIMULATOR TO MEASURE SAIL, BOAT, AND SHIP EXPERIENCE

Richard M. Evans
U. S. Naval Training Systems Center
Orlando, Florida

INTRODUCTION

Sail training in the Navy has a long past, but its recent revival is documented mainly in the Naval Institute Proceedings (McWethy, 1980, 1983; Bryce and Evans, 1986). These articles discuss the formation of the U. S. Navy Sailing Association in 1965, the staffing of a Chief of Naval Education and Training Command billet (Director of Navy Sailing) in 1981, the activities of the Naval Academy Off-Shore Sail Training Squadron's coastal and Bermuda cruises the past few years, and two evaluations of the effects of sail cruises on midshipmens' concepts of the sea and seamanship. But other measures of its effects are needed.

The Deputy Assistant Secretary of Defense for Guard/Reserve Readiness Training has called for the exploitation of "arcade technology" to increase readiness (Turley, 1985). The Deputy Secretary of Defense has recently issued a directive which authorizes the Department of Defense to use training simulators and devices that make training systems more effective or help maintain military readiness (Taft, 1986). Low-cost shiphandling technology has recently become available (Hanley and Monaghan, 1986; ECO, 1986). A pilot study was conducted to determine if table-top simulators might be used to test shiphandling performance.

METHOD

The 20 NROTC midshipmen involved in the study were enrolled in one of two major universities located within a few minutes drive of each other. All had completed coursework involving piloting, rules of the road, navigation, and general seamanship training, and were cognizant of the general tasks they were about to be given. Table 1 shows the helm hours of subjects in sailboats, motorboats, ships, and total helm time. Two experience groups were determined by dividing total helm time at the 25 hour median.

An ex post facto study was conducted to assess the effect of two levels of prior helm experience on performance in a simulated officer-of-the-deck (OOD) task. Group differences on the dependent variables of deviations from track, heading, and speed, were taken at three-minute intervals for the first 30 minutes of the problem. Data were assessed by simple and multiple correlations of experience with performance. Contrasts of high and low experience on the foregoing performance measures and a semantic differential measure of perceptions about "the sea" were all completed with a split-plot repeated-measures analysis of variance.

The dependent variables were taken from a task presented by the Computer Aided Portable Training And Instrument Navigation System I (CAPTAINS), a table-top device about the size of a personal computer which simulates the dynamics of a selected ship in a particular port area. Subjects stood a watch as OOD, and were asked to give their orders to the helmsman as if there really were a helmsman present, then carry out those orders themselves using a simulated engine and rudder control device. Three distinct displays were available to the OOD: (1) A view from the bridge containing + or - 45 degrees from heading, with the ability to shift this view to similar degree ranges port, starboard, and aft, (2) Engine and rudder performance indicators, and (3) A radar representation of the maneuver area. Heading and speed information is presented on each of the three screens.

A New York harbor scenario was chosen for the testing. Subjects were given the National Ocean Service chart number 12327, "New York Harbor," dated 29 December 1984, with course and distance information for the task already plotted. All began at the same position, which was about 1.5 nautical miles south of the Verrazano Narrows Bridge on a heading of 300 degrees true and at a speed of 5 knots. Their task was to be OOD on a medium endurance cutter, bringing the speed up to 10 knots, maintain the track which had been plotted, and bring the ship to a complete stop so that it could be anchored at a buoy in Constable Hook Reach. The test was limited to the night scene, and after crossing under the bridge, a 10-minute period of fog was simulated in which subjects were not allowed to use the views from the bridge of the ship. No other ship traffic was in the problem, and there was no tide or current present.

Navigation was accomplished by observing schematic presentations of the maneuver area in the view from the bridge, by observing the ship's position relative to the shoreline and certain buoys on the radar display, and by taking simulated sightings on 24 buoys (marked by heavy numbers 1-24 on the chart). Subjects were reminded that navigating by buoys alone was not consonant with good piloting procedure and that one would normally reduce speed in fog, however, these rules would not be followed in this problem. The device presented true bearing to the buoy and distance in nautical miles when queried.

Performance measures were recorded at three-minute intervals by the evaluator, who observed the same display as the student. A Data Recording Sheet was used to record heading, speed, and range and bearing to the center of the bridge. Each testing session was begun by starting the evaluator's stop watch; then, the foregoing data were recorded at three-minute intervals until the problem ceased at buoy 20, some 36 to 51 minutes later. Approximately 30 seconds after each three-minute fix, the position was reported to the OODs, who could use this information as a check against their own navigation and make necessary course corrections.

Information from the Data Recording Sheets was scored in three ways: (1) A deviation from the charted track was determined by measuring the position of the fix (which was based on bearing and distance from buoy 24, the center of the bridge) with a template with 4 millimeter gradations for each level, 1 to 9 from the plotted track; (2) A deviation from heading score was determined by subtracting heading at the time of fix from the course the ship should have been on for that leg; and (3) A deviation from speed score was determined by subtracting the speed at the time of fix from a 10 knot assigned speed. The scoring was suggested by Keith, Porricelli, Hooft, Paymans, and Witt (1977).

The treatment occurred prior to the study, with the median total helm experience reported as 25 hours (see table 1). Since the distribution of all of the measures of experience were highly skewed, these data were transformed to their natural logarithms after adding a constant of 2 to each value. This transformation had the effect of reducing a 500 hour range to about 6, and drastically reduced the skewness of the helm hours reported. The constant was added to allow log transformation of the zero helm experience subjects. While this lesser range may have lowered the correlation coefficients in some cases, it did satisfy the normality assumptions for use of that procedure.

Table 1. Shiphandling background hours of study participants with log transformations.

S no	Total	log+2	Sail	log+2	Boat	log+2	Ship	log+2
1.	10	2.48	0	0.69	0	0.69	10	2.48
2.	25	3.30	23.75	3.25	1.25	1.18	0	0.69
3.	5	1.95	5	1.95	0	0.69	0	0.69
4.	10	2.48	0	0.69	6	2.08	4	1.79
5.	0	0.69	0	0.69	0	0.69	0	0.69
6.	20	3.09	18	3.00	0	0.69	2	1.39
7.	1000	6.91	500	6.22	490	6.20	10	2.48
8.	0	0.69	0	0.69	0	0.69	0	0.69
9.	200	5.31	160	5.09	0	0.69	40	3.74
10.	50	3.95	10	2.48	37.5	3.68	2.5	1.50
11.	100	4.62	90	4.52	5	1.95	5	1.95
12.	40	3.74	40	3.74	0	0.69	0	0.69
13.	1	2.48	5	1.95	0	0.69	5	1.95
14.	200	5.31	170	5.15	20	3.09	10	2.48
15.	100	4.62	99	4.62	0	0.69	1	1.10
16.	4	1.79	0	0.69	0	0.69	4	1.79
17.	250	5.53	250	5.53	0	0.69	0	0.69
18.	25	3.30	2.5	1.50	20	3.09	2.5	1.50
19.	35	3.61	33.25	3.56	1.75	1.32	0	0.69
20.	1	1.10	1	1.10	0	0.69	0	0.69
Mean	104.25	3.35	70.38	2.86	29.08	1.55	4.80	1.49
Median	25	3.30	14	2.74	0	0.69	2.25	1.45
Std.D.	218.23	1.66	120.33	1.82	106.18	1.42	8.77	0.83

RESULTS

Reliabilities of the absolute values of ten measures of track, heading, and speed were calculated using a single-factor repeated measures analysis of variance procedure and were found to be significant (.78, .69, and .51, respectively). Multiple relationships of total helm hours among these measures was found to be moderate to high, but generally not significant (table 2).

Table 2 Measure Reliabilities and Multiple Correlations of Ten Measures of Deviation from Track, Heading, and Speed with Log Helm Hours in Sail, Boat, and Ship, and Log Total Helm Hours.

Ten Measures	Reliability df = 19/180	Multiple R with log experience, df=10			
		Sail	Boat	Ship	Total
Track	.78*	.57	.76	.80	.65
Heading	.69*	.71	.77	.73	.76
Speed	.51*	.90*	.64	.66	.89*

*p<.05

Figures 1-3 illustrate the differences between high- and low-experience groups for each of the absolute deviation from track, heading, and speed measures. At nearly all three-minute time intervals the deviation scores of the high experience group were smaller. ANOVA contrasts of the foregoing high experience/ low experience groups found no significance.

DISCUSSION

The CAPTAINS I table-top shiphandling simulator can reliably measure the effects of watercraft experience on shiphandling tasks. The study was limited by the time required to test each subject (approximately one hour) and the number of subjects who were available for testing. Tests were conducted during the period between examination week and graduation day at one of the universities visited, and during final examination week at the other. In spite of the obvious demands on student time during this busy period, the device was occupied during almost every hour it was available for student testing. Students kept their appointments. Several indicated that the experience they were having was meaningful for their future Navy careers, while also being fun.

This interest suggests using a simulator for teaching navigation, piloting, and watch skills in NROTC training. The training would be consonant with the recommendations of Heidt, Braby, Peeples, and Roberts (1983) for similar training with the Naval Reserve. In this regard, it is in particularly agreement with the recommendation by Turley (1985) that modern low-cost technology be adapted for military readiness training.

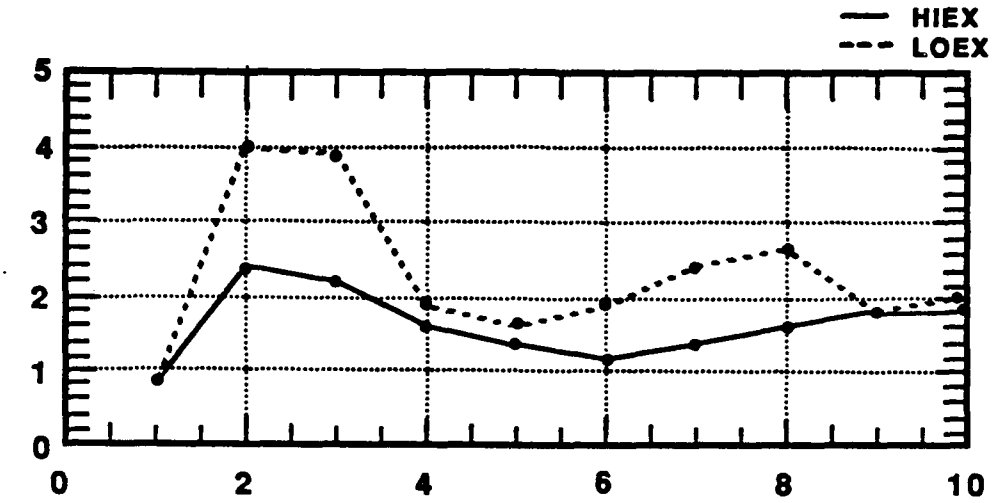


FIGURE 1. DEVIATION FROM TRACK BY HIGH AND LOW HELM HOURS GROUPS ACROSS TEN TIME INTERVALS.

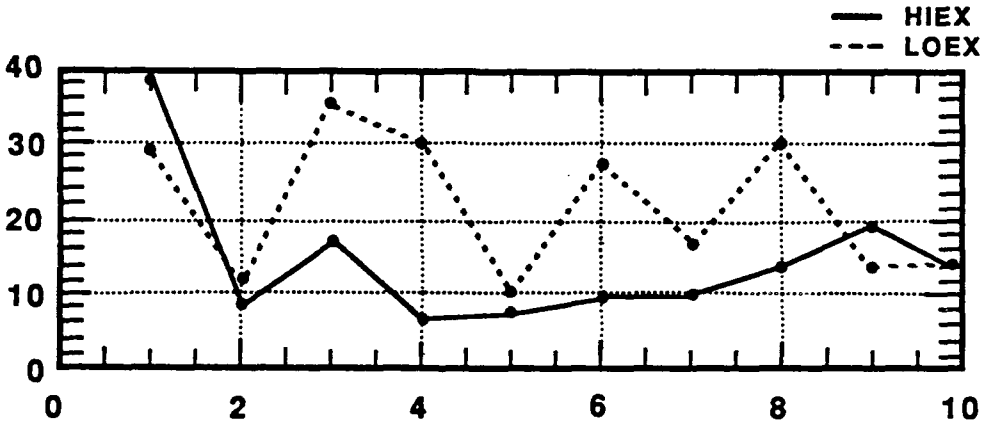


FIGURE 2. DEVIATION FROM HEADING BY HIGH AND LOW HELM HOURS GROUPS ACROSS TEN TIME INTERVALS.

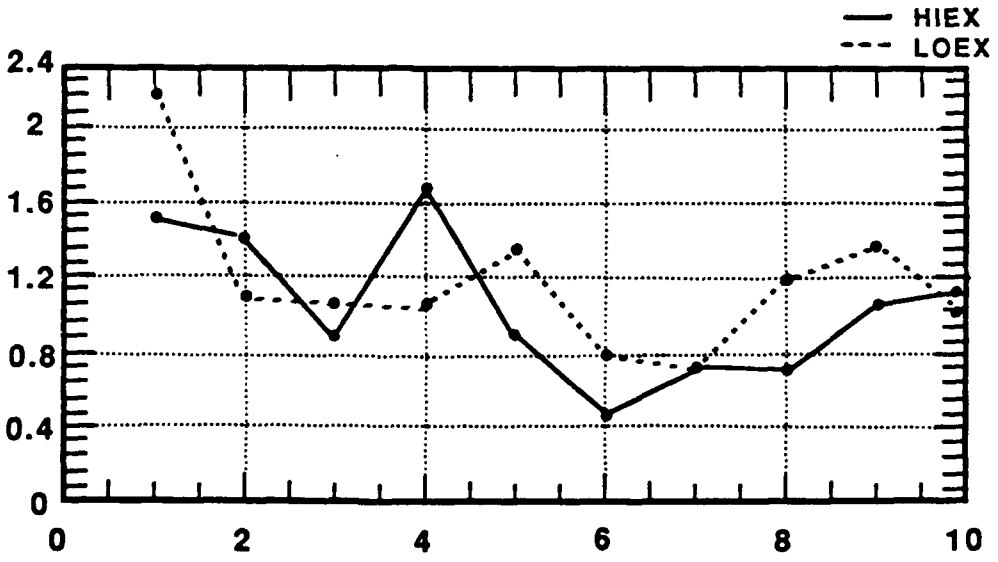


FIGURE 3. DEVIATION FROM SPEED BY HIGH AND LOW HELM HOURS GROUPS ACROSS TEN TIME INTERVALS.

REFERENCES

ECO. (1986). The CAPTAINS System: Operating Guide. Annapolis: Engineering Computer Optecnomics.

Bryce, D. S., and Evans, R. M. (1986, September). Evaluating Sail Training. U. S. Naval Institute Proceedings, pp. 44-45.

Hanley, M. J., and Monaghan, J. (1986, In press). Results of the Shiphandling Table-top Trainer (CAPTAINS) Training Effectiveness Evaluation. U. S. Naval Training Systems Center, Code 71. Orlando, Florida.

Heidt, E A., Braby, R, Peeples, T. O., and Roberts, E. G. (1983, July). A Technique for Training in the Naval Reserve. TAEG Technical Report 148. Training Analysis and Evaluation Group. Orlando, Florida.

Keith, V. F., Porricelli, J. D., Hooft, J. P., Paymans, P. J., and Witt, F. (1977). Real-Time Simulation of Tanker Operations for the Trans-Alaska Pipeline System. Transactions of Society of Naval Architects & Marine Engineers, 85, 419-458.

McWethy, R. D. (1980, December). U. S. Navy Sail Training 1980. U. S. Naval Institute Proceedings. pp.97-100.

McWethy, R. D. (1983, February). U. S. Navy Sail Training Update. U. S. Naval Institute Proceedings. pp. 115-118.

Taft, W. H. III (1986, 22 August). Training Simulators and Devices. DOD Directive Number 1430.13 ASD(FM&P).

Turley, G. H. (1985, November/December). Exploiting "Arcade Technology" for Increased Readiness. Government Executive. pp. 37-38.

ARMY CIVILIAN PERSONNEL RESEARCH PROGRAM

Paul van Rijn[1]

U.S. Army Research Institute
Alexandria, VA 22333-5600

The purpose of this paper is to describe the beginnings of a research program concerned with Army civilian employees. It includes a brief summary description of results from three separate surveys and a discussion of future civilian research by the U.S. Army Research Institute (ARI). The surveys to be described are initial efforts by ARI to become more actively involved in research to support the Army's civilian personnel management function. The results are designed to provide Army managers and policymakers with more accurate information about the beliefs and attitudes of the civilian workforce, which currently constitutes nearly 40% of the Army's personnel strength.

Values Survey

In support of the Army's 1986 theme of "Values," ARI developed a Values Survey and administered it to 5,011 soldiers and 734 Army civilians at 10 CONUS posts (5 FORSCOM and 5 TRADOC). Respondents assembled at each location were asked to indicate "how important" each of 50 values was to them "personally." On the 7-point scale, most values were rated "very" or extremely important." Table 1 shows the top 20 civilian values and the percentage of Army civilians and military who rated each value "very important" or "extremely important." Of the top 20 values, civilians share 17 with the military, and they share 9 of the top 10.

In general, the percentages of civilians rating the values very important tend to be higher than the military. There are at least three factors that must be carefully considered before interpreting these data. First, for any comparison of subgroups it is essential to recognize that the survey solicited attitudes or beliefs about values and that possible differences in attitudes do not necessarily imply corresponding differences in behavior. Second, the civilian subsample differs markedly from the total military sample in terms of age, sex, marital status, education level, and a variety of other demographic variables that are likely to impact on stated values. Third, the brevity of the values statements themselves suggests the possibility that subgroup differences in values ratings may be due, in part, to differential interpretation of the values statements.

---

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

Table 1

Similarity of Civilian and Military Values Ratings

Value	Percentage of Respondents Rating the Value Very Important	
	Civilian	Military
Freedom	97	96
Family Security	97	91
Doing Your Job Well	94	84
Self-Respect	94	92
Being Honest and Truthful	94	85
Taking Responsibility	94	88
Loyalty to the U.S.	94	86
National Security	92	88
Standing Up for What is Right	92	91
Sense of Accomplishment	90	87
A World at Peace	90	83
Freedom of Speech	88	85
Drive to Succeed	87	85
Happiness	84	84
Voting	83	75
Care of Wounded	83	88
The Constitution	82	77
Freedom of Religion	81	78
Teamwork	79	70
Treating Everyone Fairly	79	85

Additional research is ongoing to determine more precisely the correlates of high or low values ratings and the behavioral significance of the ratings. Data from a second sample in USAREUR have already indicated that the differences between CONUS and OCONUS units are minimal and analysis of the USAREUR civilian data is designed to determine more specifically which demographic variables impact most importantly on stated values.

Important Job Characteristics

A second survey was developed in conjunction with the Baltimore Field Office of the U.S. Army Civilian Personnel Center to determine the extent to which various job characteristics are important to Army civilians and the extent to which these characteristics are satisfactorily met. Twenty-five job characteristics were rated by 1,352 civilians in an Army-wide sample that was stratified on the basis of race (white,

nonwhite), sex (male, female), and job type (blue-collar, white-collar). The response rate was 64% for those surveys that were delivered at the local distribution site. Table 2 shows the percentages of respondents who rated each characteristic "important" or "very important" on a 5-point scale as well as the percentages of respondents who indicated they were "satisfied" or "very satisfied" with the characteristic.

In general, the most important job characteristics tended to be those associated with quality of working life and personal growth. This is consistent with the findings of Yankelowich and Immerwahr (1984) who argue that the American work ethic is changing from an emphasis on material well-being to greater interest in self-development. Directly contrary to this finding, however, is the importance rating received by "pay." It received the highest importance rating and at first glance suggests that Army civilian employees place greatest value on monetary compensation. An

Table 2

Importance and Satisfaction Ratings for Job Characteristics

Job Characteristics	Percent		Rank	
	Important	Satisfied	Important	Satisfied
Pay	97	66	1	5
Kept informed	96	53	2	15
Recognition	95	54	3	14
Advancement	92	45	4	16
Clean workplace	90	75	5	2
Training	90	44	6	17
Challenge	90	67	7	4
Fair assignment	90	60	8	8
Interested supervisor	89	68	9	3
Accomplishments	88	55	10	13
Creativity	85	65	11	6
Fringe benefits	83	59	12	10
EEO	83	58	13	11
Convenient location	80	80	14	1
Money for performance	72	34	15	19
Medical services	63	38	16	18
No stress	60	56	17	12
Flexible hours	57	60	18	9
Financial services	52	61	19	7
Counseling	40	28	20	21
Fitness program	30	19	21	24
Rec. facilities	25	29	22	20
Carpool program	21	27	23	22
Social activities	19	25	24	23
Day care	18	16	25	25

alternative interpretation, more consistent with the ratings on the other job characteristics, is that "pay" received its high ratings because in the federal personnel system it is so intrinsically associated with grade level, responsibility, and influence. Followup research is necessary to identify more precisely the exact origins of these high ratings for "pay."

Satisfaction ratings generally tended to be lower than importance ratings. Only the satisfaction ratings for a "clean workplace" (75%) and "convenient location" (80%) received high ("satisfied" or "very satisfied") ratings from more than 70 percent of the respondents. "Being kept informed," "recognition," "and "advancement" received high ratings from 53, 54, and 45 percent of the respondents.

Although "pay" received the highest rating, the findings suggest that job characteristics associated with quality of working life and personal growth play a significant role in work values of Army civilian employees. The "social service" programs received the lower importance ratings. However, this is not to imply that these programs should not be an integral part of the civilian personnel management function. Rather, it may be that for relatively few resources, significant improvements in the quality of worklife can be achieved for substantial portions of the Army's civilian workforce. To the extent that these programs help increase the productivity or increase the tenure of the more productive civilian workers, the potential payoffs can be even greater.

Performance Appraisal Survey

In a followup of the work begun by Steinberg and Burke (1986), a portion of the preceding survey was devoted to the performance appraisal process, with particular focus on the performance appraisal interview. The survey was completed by the same 1,352 employees described in the preceding section, plus a random sample of 625 supervisors (with a response rate of 72%). Nearly 80% of the employees indicated that they had a written performance appraisal within the last 18 months. Another 13% were not yet on the job long enough to receive a rating and only about 8% reported that they had no written rating or one that was more than 6 months overdue. Of those who received a written rating, 52% indicated that their ratings "accurately" or "very accurately" reflected their performance, and 32% believed their ratings to be "somewhat accurate." The remaining 16% felt their ratings had "very little" or "no" relationship to their performance.

Sixty (60) percent of the employees in the sample indicated that they had a formal (prescheduled) performance appraisal interview in accordance with performance appraisal regulations. Another 15% were ineligible for an interview, and about 25% had had no formal interview.

Table 3 focuses on the topics discussed during the performance appraisal interview and compares employee perceptions with those of supervisors (not necessarily the supervisors of the employees). The results show a large disparity between employee and supervisor perceptions of the extent to which various topics were discussed in the interviews. This disparity and the relatively small (41%) percentage of employees who reported that their job performance was discussed during the interview require further investigation.

Table 3

Comparison of Employee and Supervisor Perceptions of Topics Discussed in the Appraisal Interview

Survey Question(s):

In the last year, TO WHAT EXTENT DID YOU (YOUR SUPERVISOR) DISCUSS the following DURING formal (prescheduled) performance appraisal interview(s) with your employees (with you)?

Topic	Discussed	Mentioned	Not Discussed	No Interview
	SUP/EMP	SUP/EMP	SUP/EMP	SUP/EMP
Job Performance	80/41	8/21	2/23	10/16
Goals and priorities	75/37	10/17	4/24	11/21
Comparison of performance with written plan	60/31	17/15	13/16	10/39
Job-related problems	68/30	13/16	9/38	10/16
Ways to improve performance	69/24	15/13	6/46	10/16
Need for training and development	63/21	21/16	7/47	9/15
Supervisor's performance	24/15	20/8	49/60	5/17

Future Directions

In addition to the more in-depth analysis and interpretation of the results presented here, the ARI Civilian Personnel Research Program is involved in five initiatives. First, it is currently monitoring a contract to identify the job requirements that are important for selecting effective first-line supervisors. Second, it will assist the civilian personnel management community in the design of a new attitude survey system that will provide Army policymakers and managers with answers to important management questions. Third, an assessment will be made of the state-of-the-art in career development practices. This assessment will result in recommendations for improving the Army's new civilian leadership development program.

A fourth initiative is the development of a long-range civilian personnel research program based, in part, on a recently completed "Roadmap for Civilian Personnel Research," which outlines the civilian personnel information needs as articulated by over 60 Army managers, both military and civilian. ARI is also involved in a project to identify programs that have been successful in increasing the number of women and minorities in scientific and engineering fields; and, pending resources, the development of organizational productivity measures for possible use in an Army demonstration project.

References

Steinberg, A. G., & Burke, W. P. (1986). Recommendations for improving performance appraisal in the Federal sector. (Research Report No. 1418). Alexandria, VA: US Army Research Institute. ADA 168777.

Yankelovich, D., & Immerwahr, J. (1984). Putting the work ethic to work. Society, 21, 58-76.

TRADEOFF CONSIDERATIONS IN PROVIDING IMMEDIATE FEEDBACK TO ORGANIZATIONS

Capt Michael S. Williams, USAF
Capt Jeffrey S. Austin, USAF
LtCol Frank R. Wood, USAF

United States Air Force Academy, Colorado

ABSTRACT

It is generally thought that organizational feedback, like individual feedback, is most effective when provided as close to the time of data gathering as possible. Delayed feedback can be problematic because issues change, interest diminishes, and personnel turn over. Truly immediate feedback has only become possible with recent changes in data processing technology -- development of portable scanning devices, computing equipment and more powerful statistical software. Even so, recent experience providing immediate feedback to organizations in the field suggests there are still numerous limiting factors which must be considered. For example, environmental conditions may affect the operation of the equipment just as the social dynamics of the organization may affect the efficiency of the data effort and the effectiveness of the feedback process. This paper will raise these issues, identify pitfalls to be avoided and discuss trade-offs which must be considered as the "cost" of providing organizations immediate feedback.

INTRODUCTION

This research paper began as part of a larger effort of investigating the impact of shift work on nuclear security worker effectiveness. Our problem at the time was to obtain attitudinal and personal attribute data from large numbers of security personnel at numerous sites on various shift schedules to look for consistent trends. Our ultimate goal for the project is to provide pertinent trade-offs that appear to exist across specific types of shift arrangements. Our need to obtain data, and our felt obligation to provide feedback to the units from whom we received the data, provide the data for this paper. Our difficulty in finding information on the benefits and disadvantages of providing immediate feedback to organizations in the field substantiate our purpose for this paper.

There are several reasons we considered the benefits of more immediate feedback. The most basic of these considerations is that of cost effectiveness. If the data can be collected, analyzed, and explained to the client on the same trip, the costs can conceivably be reduced by half. Other more academically driven ideas seemed to support the same conclusion. From basic learning theory, it is generally held that feedback is reinforcing in and of itself (Pritchard, et al, 1981). It has also been long accepted (Skinner,1938), that increased time between response and the subsequent delivery of reinforcement (feedback) inhibits learning.

From the organizational literature, evidence suggests that employees and supervisors depend on and even generally seek feedback (Robbins,1986). Feedback about results of one's behavior can serve two distinct functions. The first is a directional function of providing the key information, at the right time, regarding behaviors concerning the successful performance of their jobs. A second feature is that it provides information about the outcomes of behaviors that are associated with potential rewards.

Ilgen et al. (1979) list numerous influences on the effectiveness of feedback. Included in these are the credibility of the person who is the source of the feedback, the timing of the feedback, and the relevance of the feedback. For example, how much relevance is there to data (at least at the perceived level) if there have been changes in personnel from the time of survey administration to return with feedback data? What happens to the credibility of the person returning with the feedback if a supervisor discounts all the data because of its (perceived) loss of relevance due to the time delay? What happens when specific comments which tend to provide "life" to the data are lost because of the delay of feedback?

Delayed feedback becomes problematic for a variety of reasons. Most return feedback sessions appear to involve a period of four to six weeks delay. This time is necessary for analysis of data, preparation of feedback materials and graphics, and the selection of strategies for interventions. In most military organizations, this type delay results in an 8.6% turnover to as high as 33% turnover if the timing of the process occurs during the high summer turnover months. Rationalizing away the feedback data is easier when a supervisor can blame identified issues as problems that have already been corrected during the interim. While good consultants can confront that type behavior, getting the individual to "buy in" is considerably more difficult if any "loop holes" can be found in the data set.

In summary, it appears that there are many reasons for immediate feedback to the client. Indeed, no literature sources caution about immediate feedback with trained consultants. From the literature, one would conclude that if the equipment was available, and the skills of the consulting team were diverse enough, then immediate feedback would appear to be the obvious choice.

METHODS

During this effort we surveyed 1311 U.S. Air Force personnel at three European bases and one in the United States. The survey we used was specifically designed to measure personal attributes and the impact of work schedules on attitudes toward the job. The subjects were primarily first term enlisted security policemen who work various schedules around the clock. The surveys were administered during the subjects normal duty time at their duty location.

The equipment we used included a Kaypro II portable computer (CPM), a National Computer System Sentry 3000 scanner, and a Zenith dot matrix printer. The software used was the commercially available "StatPac" statistical analysis package by Walonick

Associates.

Our plan was to spend one week at each base surveyed. The first two and a half days would be dedicated to data gathering around the clock so as to catch all workers during their normal duty hours at their duty location. The next day and a half was spent doing data analysis and preparing graphics for presentation. On the last day we provided feedback presentations. Those who received feedback were the top three levels of supervision (Squadron Commander and staff, Flight Commanders, and the Shift Leaders). Only aggregated data were shown to each level so we would not violate the confidentiality of the data. For the overall commander, data we provided were broken down by work schedule and marital status. The Commander also saw comparitive data for his squadron versus all other units previously measured. Generally, this was a typical data gathering and feedback process with the primary exception being use of new technology that enabled the team to analyze and feedback the data in a far more rapid manner.

RESULTS AND DISCUSSION

When we consider the possible advantages of portable computing equipment we tend to make the assumption that portablilty equates to immediacy and immediacy equates to effectiveness in the feedback process. While this may be true, these relationships might not always be as predicted.

Portability does not always mean the feedback will be immediate. In fact, portable systems may be more prone to unplanned delays. Along with enhanced flexibility, portability may also introduce too much or the wrong kind of change into the total data analysis system. For example, portable systems by their very nature must operate in changing physical environments and unknown conditions. Stable power conditions cannot be assumed, especially in foreign countries, unless the computer system also has its own power supply. Variable voltage, brown outs, power failures and the like, typical of many geographic regions can cost much in terms of lost output, data files, and software and render the researcher no better off with the equipment than they would be without it. In other words, the stability of laboratory environments cannot be assumed.

The transportability of equipment does not guarantee that it will reach its destination either. Visiting overseas locations, we learned the hard way that some countries prohibit the entry of certain electronic equipment, specifically computers into their country. In short, such equipment may be of little use to you if it is confiscated by customs officials.

Even the portability of some systems must be questioned. Some portable units are bigger and less portable than others. They may be so heavy that they are only "luggable" and more suited for short trips. Others may be light but not be small enough to fit under an airline seat so they must be checked as baggage. If the system must be checked or shipped, the probability of damage goes up and the need for additional protective packing may render the system less portable.

Finally, in regard to portability, such systems seem to be

highly specialized and idiosyncratic by design. At present, the systems are few, relatively new and rapidly changing. Since there seems to be less standardization among these specialized systems, compatibility becomes and issue. In fact, compatability problems with the equipment we chose for an upgrade prevented us from taking the new equipment on one of our scheduled visits and required the expertise of several computer specialists to insure it was ready for follow-on data collection efforts.

So, portability has a cost. Changing operational environments introduce unplanned and often dysfunctional changes into the total data analysis system. The end result may be no differnt than not using the equipment in the first place -- except for for the frustration of not accomplishing what is expected.

Our presumption that immediate feedback is advantageous is just as questionable as our presumption that portability allows us immediate feedback. While immediate feedback can be very helpful, it can also be very threatening, depending on the social dynamics of the organization being studied. Further, these social dynamics may alter the data collected so that the findings are questionable. A critical example of this is in the potential misinterpretation of the study's purpose.

Organizations which are frequently evaluated tend to view any measurement of its operation in that manner. The researcher's assertion that the "feedback is not an evaluation" is difficult to believe; after all, it is an objective measure of their unit's performance which might be available to others. Also, immediate feedback, more than any other, is especially threatening because it is difficult to discount, whereas data gathered some time earlier can be discredited with the reply that "things have changed."

In our experience, misinterpretation of the study's purpose is most likely to occur during the initial contact with the unit and when the purpose is recommunicated downward to subordinate commanders in the unit. Therefore, interaction between the researchers and the unit's point of contact (POC) is critical because miscommunication of purpose at that early point may not be correctable by the time the researchers arrive on the scene. Generally, it is best to insure the senior commander clearly understands the nature and purpose of the study whether he/she is the POC or not. In other words, even if you deal with someone else, always talk personally with the senior client. His articulation of what you are doing is the perception of the unit no matter what you have told everyone else. Another problem is the understanding and perception of subordinate commanders and supervisors. Be attentive to misunderstanding (e.g., labeling the researchers as "the evaluators") and deal with that perception immediately, at the level it surfaces and with the smaller unit commander. During any of our visits it was not uncommon for one of the senior researchers to spend one or two hours with a subordinate commander explaining what the study was and how the results would be used in order to get their support -- after such understanding and were already assumed by the POC and the senior commander.

The key to setting straight the perceptions about the study is the careful articulation of how the results will be used. In

our case, we explain: "Scores on the squadron as a whole will be provided to the squadron commander and scores of subordinate units will be given only to the individual units with comparisons to the squadron as a whole. In no case will the scores of one unit be provided to another in the study -- even if they are outstanding." Only after some discussion of the extent to which we have insured anonymity in the past and some testing to see if we will reveal another unit's identity, are perceptions that the study is "evaluative" reconsidered.

Another aspect of the unit's social dynamics which can be problematic is the subject's need for recognition. The nature of our data collection method -- surveying subjects on their duty time, at the researcher's expense and interacting personally with them -- meets the respondent's needs but incurs both advantages and disadvantages for the research effort. On the one hand, subjects are more willing to "tell it like it is" and feel "someone cares." On the other, there is a chance that the researchers may be interfering with the data being collected. For example, we know that the word passes from shift to shift that we are "OK" or that "we will listen and may be able to make a difference." In fact, after we had out-briefed at one base and were leaving the main gate enroute to our next destination, we we stopped by the gate guard and asked if he "could take our survey." He had heard that it was "a good one" and since he was on temporary assignment with another unit, he was afraid he might not have the chance to express his opinions. Such enthusiasm may be a problem if the subjects respond differently than they would if the researchers were not meeting the subject's need for recognition.

The hidden costs of truly immediate feedback are subtle but important. Such feedback can be seen as more threatening (more evaluative) and the subjects enthusiasm may alter their responses. Researchers, then, must be attentive to the social dynamics of the organizations they are studying. Perceptions about the purpose of the study must be set correctly at the beginning. Differing perceptions must be countered as soon as they surface. Assurances that the study's purpose is to provide "evaluation-free feedback" must be taken to the extreme. Researchers must also constantly check each response against other measures of the organization's climate to insure extreme cases are understood and properly weighed. Thus, immediate feedback has its own hazards which require special senitivity.

In sum, the relationship between portability, immediacy and effectiveness cannot be assumed. Portability introduces change into the data analysis system which in some cases may render the system useless and the feedback may be delayed as much as if the portable equipment had not been used. The social dynamics of the organization may be such that the intent of the study is misunderstood and the data collected may be unreliable.

In this paper, we have intentionally emphasized the costs more than the benefits of portable systems and immediate feedback because we want others to be aware of the pitfalls and to be more effective in their own use of this process. However, we would be remiss if we overlooked the benefits. Clearly, in the cases where the systems worked as planned and the social dynamics of the organization were supportive, the process was very effective. For

example, dollar and time savings were very apparant when visiting overseas locations because the entire data collection and feedback process was accomplished in a single visit. Also, the credibility of the feedback was enhanced by the real-time nature of the process. Essentially, we were immersed in the organization for the entire data collection and feedback process. We shared the same organizational experience as they and were able to discuss current issues with them, in their language and using examples they had given us. Finally, turnover in the organization was not an issue for us. We reported our findings to those who were on the scene when we collected the data. If they were supervisors, the issues raised by us were generated by their own policy and decisions. Before we had this capability, we found ourselves returning more than four weeks later to "different" units with different problems. While they always listened politely to our feedback, it was not very important to them.

Truly immediate feedback, allowed by recent portable computing technology, can be effective. When the systems work, supervisors listen because we tell them what 95% of their people have told us about their oraganization in the last 48 hours. In many cases they begin working the problems we surface before we walk out the doors of their organization!

REFERENCES

Ilgen, D.R., Fisher, C.D. & Taylor, M.S. (1979). Motivational consequences of individual feedback on behavior in organizations. Journal of Applied Psychology, 64,349-371.

Pritchard, R.D., Bigly, D.G., Beiting, M., Coverdale, S., and Morgan, C. (1981). Engancing productivity through feedback and goal setting (AFHRL technical report 81-7). Brooks AFB, Tx.

Robbins, S.P. (1986). Organizational Behavior: Concepts, Controversies, and Applications. Englewood Cliffs, NJ: Printice-Hall.

Skinner, B.F. (1938). The Behavior Of Organisms. Englewood Cliffs, NJ: Prentice-Hall

Mentoring, Chapter Three: Perceptions of Potential Proteges in the USAF

Capt Jeffry A. Gouge
Air Force Institute of Technology, Wright-Patterson AFB, OH

Captain Benjamin L. Dilla
Air Force Military Personnel Center, Randolph AFB, TX

Previous research by the Air Force Institute of Technology has examined perceptions of the mentoring process from the perspectives of young field grade officers who have been proteges (Uecker & Dilla, 1984) and more senior officers who have served as mentors (Lewandowski & Dilla, 1985). This latest study examined perceptions of the process from young officers who are potential proteges. A sample of 106 officers in technical training at the Aircraft Maintenance Officers Course responded to our survey. Items concerned their expectations for gaining an Air Force mentor, perceived roles or functions of the mentor, expected outcomes of the process, and various background factors relevant to the process. Analysis indicated substantial interest in, and positive expectations of, mentoring; however, having a mentor was not seen as essential to a successful military career. Perceptions of the potential proteges were compared to those of more experienced Air Force proteges and mentors.

Introduction

The topic of mentoring has gained considerable attention in the private sector and, more recently, in the military services. A survey of business executives by Roche (1979), published in Harvard Business Review, established the prevalence and beneficial effects of mentoring among top-level executives. Roche found that "nearly two-thirds of the respondents reported having had a mentor or sponsor "; furthermore, data showed that "executives who have had a mentor earn more money at a younger age, are better educated, are more likely to follow a career plan, and, in turn, sponsor more proteges than executives who have not had a mentor" (Roche, 1979, pp. 14-15). Also, executives who had been mentored reported greater job satisfaction and satisfaction with their career progress than their unmentored counterparts.

Two recent studies conducted by the Air Force Institute of Technology used an expanded and modified version of Roche's survey to examine mentoring in the U.S. Air Force and compare the process and its effects to those in the private sector. At a previous Military Testing Association (MTA) conference, Uecker and Dilla (1984) reported the first and most directly parallel study to Roche's original work. Two hundred fifty-two Air Force officers attending Air Command and Staff College (ACSC) and Air War College (AWC) responded from the perspectives of having been proteges within the Air Force. The prevalence of mentoring was not as great even in this select sample (42% mentored officers versus 64% mentored executives in Roche's survey); however, the effects were similar. Mentored officers had a greater incidence of advanced degrees and were more likely to have formulated a career plan. As a parallel to Roche's finding that mentored executives earned more money at an earlier age, mentored officers were more likely to have been promoted early. Also, they reported greater job and career progress satisfaction (Uecker & Dilla, 1984).

At last year's MTA conference, Lewandowski and Dilla (1985) reported an extension of this research to examine the perspective of officers who had served as mentors. A modified survey was sent to 112 officers selected for AWC attendance and occupying position such as squadron commanders, directors at air division level, and system program directors. Of 95 respondents, 61% reported having had a mentor (not significantly different from the 64% reported by Roche, 1979) and 48% reported having served as a mentor to another individual. Because of small sample size, this study failed to replicate the previously discovered effects of having a mentor except for significantly greater career progress satisfaction. Officers who had served as mentors reported significantly higher job satisfaction than those who had not mentioned others; no differences were found for early promotions or career progress satisfaction for this subgroup.

Another perspective not yet examined in our research, or by any in the private sector, concerned the expectations for mentoring among young people just entering their professional careers as Air Force officers. Thus, the target population of our third study was labeled as "potential proteges". A number of questions regarding this population have never been addressed. Do young people plan on seeking a mentor? What types of people seek or accept a mentor? What are their expectations of the relationship and its effects? Answers to these questions will help give a more complete picture of mentoring in the Air Force.

A primary focus of our research has been identifying what a mentor is and does in terms of functions or roles. Although many have offered their opinion of these functions, we have found the list of ten roles of the mentor by Lea and Leibowitz (1983) to be the most complete and comprehensive. Their ten roles include teaching, guiding, advising, counseling, sponsoring, role modeling, validating, motivating, protecting, and communicating. Comparing perceptions of these roles should reveal if the expectations of potential proteges correspond with the experience of proteges and mentors in the U. S. Air Force.

Other areas of interest in our research include background factors which may predispose a person to seek a mentor, expected effects of a mentoring relationship on protege, and possible contributions to the mentor. It is of interest to find out if the potential proteges' expectations correspond with the reality of mentoring as it is today in the officer corps.

This research was performed by Captain Jeffry A. Gouge for his master's thesis requirement at the Air Force Institute of Technology. The primary advisor for the thesis was Lt Col Paul Reid, with technical advice from Captain Benjamin L. Dilla to help relate the research to previous work.

Method

Sample

Because of the specific interests of Captain Gouge (an aircraft maintenance officer) and the focus of our institution on Air Force logistics, 115 officers attending the Aircraft Maintenance Officers' Course (AMOC) at Chanute AFB, IL, formed the pool for the survey. All of these individuals were active duty Air Force officers attending the six month technical training course prior to their first assignment as aircraft maintenance officers. Since the maintenance career field has no unique selection requirements for officers, it was assumed that the sample would be representative of new Air Force officers throughout the non-rated line.

Procedure

Surveys were distributed, administered, and collected on-site by Captain Gouge. Participation was voluntary, and respondents were assured of anonymity.

Measures

The survey was based on Uecker's (1984) adaptation of Roche's (1979) instrument. Changes in response options reflected the perspective of potential future involvement in a mentoring relationship rather than past experience. In some cases, response options were changed to produce an interval response scale rather than dichonomous or categorical responses. From the previous survey, ten items asked for background information, five concerned previous mentoring experience, three items focused on expectations of mentoring effects, sixteen asked for ratings of importance on characteristics of the (potential) mentor, ten focused on roles of the mentor (should/should not be performed), and fifteen items listing characteristics of a successful career were rated for their importance to self and others.

Four items were adapted from Lewandowski's (1985) survey regarding the importance of the protege to the mentor in terms of the mentor's job satisfaction, success, technical currency, and staying informed.

A key item added to this survey asked respondents if they would seek or accept a mentor. This item differentiated the sample in terms of self-stated desire to participate in mentoring, just as previous studies had used past mentoring experience to split their samples. Finally, an item was added to gauge the extent to which respondents saw mentoring and sponsoring as the same phenomenon in the Air Force.

The revised survey was reviewed for content validity and consistency with previous surveys by several experts in survey development. Furthermore, the survey was reviewed and approved by the Survey Branch of the Air Force Military Personnel Center.

Results

Respondent Profile

Of the 115 officers enrolled in AMOC, 108 responded to the survey. Two incomplete questionnaires were deleted from the sample, resulting in a sample of 106 and a response rate of 92%. In terms of demographics, 10% of the sample were females. The predominant rank in the sample was second lieutenant (97%), with two captains (2%) and one major (1%) among the respondents. The age distribution showed a surprising spread, with the median age at commissioning being 25 and 23% of the respondents reporting age 28 or more. Most respondents were commissioned through OTS (70%) vice ROTC (29%) or USAFA (1%); 39% of the sample reported prior military service. Furthermore, a large share (82%) reported at least one full-time non-military employer in the past. Thus, although the respondents were new to the Air Force officer corps, they did not appear to be naive or inexperienced.

Expectations of Mentoring

In terms of desire to obtain a mentor, 42 respondents (40%) stated they would seek a mentor while 54 (51%) said they would not seek but would a mentor while 54% (51%) said they would not seek but would accept a mentor. Only 10 (9%) said they would not seek nor accept a mentor or were undecided. The sample was thus divided into mentor seekers and non-seekers (a 40%/60% split).

The nine demographic items were used in a discriminant analysis of these groups. Of the nine variables, four were found to be discriminators at the seekers tended to have higher undergraduate GPAs and more nonmilitary full-time employers and were more likely to have a father who was a military officer and to have had a prior mentor. Factors which did not discriminate included age at commissioning, extracurricular activity, prior service, gender, and source of commission.

In examining the expected effects of mentoring, respondents were first divided into two groups based on the rank they expected to attain--those who expected to attain general officer rank and those who did not. Using a Kruskal-Wallis test, there was no significant difference in the extent to which individuals in either group would seek a mentor. Furthermore, mentored officers were not seen as more likely to be promoted early, and a mentor was not seen as an essential factor in a successful military career in the ratings of fifteen factors for self and others. When asked if they saw mentoring as the same as sponsoring in the Air Force, 26 respondents (25%) stated they were unfamiliar with the term "sponsoring". Only 24% agreed with the assertion; however, 20% were neutral. Grouping the neutral responses with the "agree" responses and testing the distribution using a binomial test failed to reject the null hypothesis that mentoring and sponsoring are perceived as the same.

In examining the potential protege's perceptions of their importance to the mentor, four items were tested using binomial tests. Only in the area of the mentor's job satisfaction was a significant effect found.

Roles of the Mentor

The binomial test using an approximation to the normal distribution was used to test each of the ten roles identified by Lea and Leibowitz (1983). Grouping undecided responses with positive responses toward the role and using a test proportion of .50, all of the roles were supported except for sponsor and protector.

Using the actual ratings and an overall rank ordering of the roles for this sample allowed for comparison of the perceptions of the roles by the potential proteges of this study and the actual proteges and mentors of previous research. Data are presented in Table 1.

The highest rated roles were similar across studies and the bottom two roles were identical in all three cases. The roles of guide and teacher seemed to be given more relative importance by potential proteges than by the previous groups of proteges and mentors.

Discussion

This study examined mentoring from a completely new perspective, considering the perspective of potential proteges who are relatively new to the organization. Expectations among this group of Air Force officers seemed to be positive, with only three percent saying they would not seek nor accept a mentor. In fact, mentoring was not a new experience for a substantial

Table 1
Roles of the Mentor

	Average Rating (& Rank within Group)		
	Potential Proteges[1] (n = 106)	Study 1 Proteges[2] (n = 106)	Study 2 Mentors[3] (n = 46)
Role Model	3.547 (1)	1.924 (1)	1.711 (4)
Advisor	3.536 (2)	1.853 (2)	2.158 (1)
Guide	3.453 (3)	1.500 (7)	1.675 (5)
Motivator	3.415 (4)	1.800 (3)	1.798 (3)
Teacher	3.368 (5)	1.441 (8)	1.611 (6)
Counselor	3.245 (6)	1.598 (5)	1.932 (2)
Communicator	3.240 (7)	1.505 (6)	1.561 (7)
Supporter	2.886 (8)	1.613 (4)	1.500 (8)
Sponsor	2.443 (9)	1.426 (9)	1.343 (9)
Protector	1.765 (10)	0.964 (10)	1.095 (10)

1
Ratings and assigned scale values were:
Definitely should assume this role = 4; Probably should = 3;
Undecided = 2; Probably should not = 1; Definitely should
not = 0.

2
Data adapted from Uecker, 1984. Ratings and assigned scale
values were: Most important role = 3; Primary = 2 Secondary = 1;
Not Played = 0.

3
Data from Lewandowski, 1985. Ratings and assigned scale
values were: Most important role = 3; Primary = 2; Secondary = 1;
Not Played = 0.

portion of the sample; 45% reported an average of almost two mentors prior to entering the Air Force officers corps.

Analysis revealed four background factors which were discriminators between mentor seekers and non-seekers -- undergraduate GPA, previous nonmilitary employers, military officer fathers, and prior mentoring relationships. Although the discriminant analysis function was significant at the .01 level, it correctly predicted group membership only 63% overall (with some bias due to the function being tested on the same data used to derive it). Obviously, the results show only trends, not hard-and-fast discriminators which will always apply.

There was insufficient evidence to conclude that the respondents felt having a mentor was required to attain general officer rank, be promoted early, or have a successful career. Neither did they see themselves as proteges being essential to their mentors' success; however, respondents did feel that mentoring produced greater job satisfaction for the mentor. This result is consistent with the finding from our earlier study (Lewandowski and

Dilla, 1985) that officers who had served as mentors reported significantly greater job satisfaction than those who had not served as mentors.

There was also insufficient evidence to reject the null hypothesis that respondents would view mentoring and sponsoring as the same phenomenon; however, responses to this item were marked by a high proportion (20%) of neutral responses and many (25%) who stated they were not familiar with the term "sponsoring." In responding to the roles of the mentor, sponsor was one of the lowest rated roles. Apparently the distinction between mentoring and sponsoring is somewhat fuzzy but does exist among those new to the Air Force.

Ratings of "sponsor" and other roles of the mentor were fairly consistent with past results. Like actual proteges in previous research (Uecker and Dilla, 1984), these potential proteges saw the mentor primarily as a role model and advisor. They were also consistent with both past proteges and mentors (Lewandowski & Dilla, 1985) in the lowest ratings assigned to the roles of sponsor and protector. In fact, rankings of the four lowest roles were identic l between this group and the mentors of the latter study.

In general, mentoring was not viewed as an essential characteristic of a successful military career, ranking 11th out of 16 characteristics. However, mentoring was rated substantially higher by mentor seekers than by non-seekers. Among the mentor seekers, 60% assigned extreme or moderate importance to a mentor, while only 35% of non-seekers gave it a similar rating. This is consistent with the findings of Uecker and Dilla (1984) for mentored versus unmentored officers.

This finding must be balanced by the fact that, for both groups, factors such as the ability to complete assignments, leadership, and decision-making were rated as the key requirements for a successful career. Even among those with a strong desire to gain a mentor, mantoring does not appear to be viewed as a free ride to the top.

An indirect conclusion from this research is that the Air Force may not need to publicize the concept of mentoring, as recommended in previous studies. It would appear that most young officers are at least somewaht aware of the concept and favorable towards gaining a mentor or professional role model. Further publicity as an Air Force initiative could lead to attempts to implement mentoring programs in a rigid, bureacratic manner. Mentoring is a natural phenomenon which probably does best without such structure. Any attempts to further encourage the process should be kept low-key and informal. Those with favorable attitudes toward the process appear to be present at all levels.

References

Lea, D., & Leibowitz, Z.B. (1983). A mentor: Would you know one if you saw one? Supervisory Management, 28, 32-35.

Lewandowski, F., & Dilla, B.L. (1985). Mentoring in the United States Air Force: The mentor's perspective. Proceedings of the 27th Annual Conference of the Military Testing Association, pp. 216-221.

Roche, G.R. (1979). Much ado about mentors. Harvard Business Review, 57(1), 14-28.

Uecker, M.E., & Dilla, B.L. (1984). Mentoring as a leadership development tool in the United States Air Force. Proceedings of the 26th Annual Conference of the Military Testing Association, pp. 423-428.

PREDICTING ACADEMIC SUCCESS OF OFFICER CANDIDATES [1]

Albert H. Melter, Ph.D.

Central Personnel Office of the
German Federal Armed Forces
Cologne, FRG

All applicants for the army, the air force, and the navy take part in the aptitude test of the examination center for officer candidates at Central Personnel Office. Klassmann (1984) reported to the 26th Annual Conference of the MTA in Munich about the selection procedure. Now, there is a survey of the courses of studies at the universities in Hamburg and in Munich, of classifying the applicants with academic plans, and a report about the empirical test of some classification data.

Courses of Studies at GFAF Universities

Studies have become an integral part of training for future regular officers and for most officers with at least twelve-year signing up. Studies are to enable the officers to use scientific knowledge and methods. They make available better prerequisites for taking the responsibilities as military leader, trainer, and educator, and for proving as an officer.

Students have to complete their studies within a three-year period of time. The academic year subdivides into three 1/4-year terms in which lectures are held and one 1/4-year term without lectures taken up by annual leave, practical courses, and examinations. Three academic years are followed by a trimester to prepare and to hold the finals. If a student has failed, studies have to be completed within the maximum period of four years.

Candidates with university entrance qualification can choose out of pedagogics (PAD), economics & organizational science (WOW), civil engineering (BI), electrical engineering (ET), mechanical engineering (MB), aerospace technology (LRT), surveying (VM), computer science (INF), some new courses projected for the following years, and out of practice-orientated courses of studies. Candidates with the special (FHS) entrance qualification can only choose out of the practice-orientated courses of studies: management (BWL FHS), civil engineering (BI FHS), electrical engineering (ET FHS), mechanical engineering (MBK FHS), and aerospace engineering (MBL FHS).

Army officer cadets with pilot or unit training are allowed to begin their studies 3 1/4 years after entrance. Army officer

---

[1] All views expressed in this paper are those of the author and do not necessarily reflect the official policies or positions of the Central Personnel Office or the Federal Ministry of Defense FMOD - P II 4.

cadets with training in supplies, maintenance, and aircraft technique begin their studies already 2 1/4 years after entrance. Air force and navy officer cadets begin their studies 15 months after entrance.

Mathematics are considered as the critical criterion for nearly all courses of studies. But good school results in mathematics give no guarantee for academic success. Motivation for studies, special interests in the subjects, and ability to take stress are also important. It is a fact, that studying at GFAF university involves, irrespective of the course of studies, higher demands on the student, and there is no doubt that the state of knowledge and readiness to learn require more attention. For example, in the first academic year, engineering calls for learning basic mathematical, technical, and natural sciences which are lectured with different weight. Computer science has high standards, particularly for the ability to abstract. Pedagogics do not give the teaching profession. The main subjects are education, sociology, and psychology. Prerequisites are good school results in german language, interests in associating with groups, and in human development and change within the social reality.

Psychological Analysis and Decision

The military advisers of the course guidance service and the psychologists of the aptitude test department are focussing different fields: the former base on statements about the extent of knowledge in specific subjects, on interests, inclinations as well as informations about qualifications and capabilities that are important for the successful completion of studies, the second refer to problems encountered in school, test results, and personality ratings.

Some criteria of academic success are achievement in university seminars and examinations, duration, change or failing of one's course of studies, and individual satisfaction and integration. Psychological determinants are cognitions of situational conditions, abilities, skills, motivations, and emotions which, multiplied by each other, decide about achievement in studies (Florin & v. Rosenstiel, 1976).

The psychological analysis aims at appropriate placement in a field or course of studies and at classification of the chances of succeeding. It is favourable to assess at least field- and, if it's possible, course-specific informations. Placement and classification have to accept shortcomings from the length of training before university entrance, from the duration of the studies, from the instability of young adults' plans, and from the psychometric quality of methods and their use.

Psychologists use data from ability and knowledge tests, school reports, biographical data, self- and expert-ratings in order to survey aptitude and cognitive interests for studies. Longitudinally validated university entrance tests are missed for that

purpose which may facilitate a more sophisticated diagnosis (Trost, 1985).

Test informations and school report data are combined (not by statistical means) with biographical data of the applicants, their verbal and reasoning behavior in the examination stations, with self-ratings of aptitudes, cognitive interests and social activities, learning and motivation for studies. The psychologist classifies the expected success in the applicant's course of studies. Such a study recommendation can be: no objections, in part there may be difficulties but success is expected, or success doubtful. Confidence or doubts and the expected difficulties in parts have to be well-founded with objective, reliable, and valid diagnostic informations.

Empirical Test of Classification Data: Problem

There is the hypothesis that diagnostic informations so far before the university entrance of the candidates can be obtained in a objective and reliable manner (f. e. Otte, 1985, reports reliability coefficients of .90 for the ability tests), but that they are not sufficient for predicting university examination results. In addition, it is still uncertain, whether test informations or school results allow better predictions, and which combination of them is the best for each course of studies. This problem area was investigated to get a general idea from easily available and processing test, school, and examination informations.

Subjects

The sample consists of two officer student crews. These 2214

Crew	1982	1983
University Course of Studies	894	974
Practice-Orientated Course of Studies (FHS)	155	191

students had succeeded in the test procedure of the Central Personnel Office and in the military training before university entrance. They entered the basic course of four trimester and had an intermediate university examinantion with written and oral parts.

Variables

Each officer student scored on seven variables. The mean of the results in the last school report was calculated in the aptitude test procedure correct to two decimal places on a six-point scale (SAB). The second school result was the single achievement score in mathematics, converted into a six-point scale (MAT). The mathematical test variable was a total standard score out of algebra, geometry, and functions items (MTAS), designed for stu-

dents before and after university entrance about 15 years ago. A new test is in progress.

The ability test variables were the sum raw score out of the analytic-constructive part of the intelligence-structure test consisting of the five subtests arithmetical problems, sequences of numbers, figure completion, cube rotation, and verbal memory (IT59), the total raw score out of the verbal part with four subtests sentence completion, verbal comprehension, analogies, and common concepts (IT14), and the raw score out of the Raven Advanced Progressive Matrices (RAM). The general knowledge test variable was the total raw score out of 13 subtests, for example,history, chemistry, economins, and other fields of cognitive interests (WT72).

The intermediate examination mark after the first (basic) academic year was the dependent variable (six-point scale) in a stepwise analysis of multiple regression to determinate the most powerful test or school report predictors for each crew and course of studies. Interpretations depended on significance level (.05) for correlation coefficients and variance comparisons, on the squared multiple correlations, and on regression weights and standard errors of estimate which were analyzed to come to a statistical combining and control of base-line informations for study recommendations in the future.

Results

SAB and MAT of the last school report before aptitude testing and the MTAS score out of the aptitude test battery proved to be the best predictors (see table 1). In 14 out of 24 regression analyses, the mean of the last school results (SAB) was at the first place in explaining the criterion variance. The school result in mathematics was in four analyses the first significant predictor. In three analyses , the mathematical test score (MTAS) received the first place, and in five analyses, this test predictor held the second place in significant criterion variance determination.

The results of the regression analysis are not stable both within the fields of studies (f. e. engineering) and between the crews. The criterion variance determination in the pedagogics subsample (PAD) due to the result in the analytic-constructive part (IT59) of the intelligence-structure test is not yet understandable. Above all, this test information is considered important for engineering, but the results are only positive for mechanical engineering (MB) and for aerospace technology (LRT). There are inconsistent findings for electrical engineering (ET), civil engineering (BI), and surveying (VM), too.

The mean of the school results (SAB) had significant correlations with the intermediate examination result in nearly all course of studies and crew subsamples. This finding applies also to the mathematical mark (MAT). There are obviously important symmetri-

Table 1: Results of the Multiple Regression and Correlation Analysis.

Course of Studies & Crew		Significant Predictors: 1.	2.	3.	R^2	N	Significant r with Intermediate Examination Mark: SAB	MAT	MTAS	IT59	IT14	RAM	WT72
WOW	82	SAB	MTAS	-	.16	300	.31	.25	.31	.18	-	.11	.10
	83	MTAS	SAB	-	.12	338	.23	.21	.28	-	-	-	-
BWL	82	SAB	-	-	.11	60	.34	.30	-	-	-	-	.21
FHS	83	SAB	-	-	.10	62	.31	-	-	-	-	-	-
PAD	82	SAB	-	-	.10	205	.32	-	-	.15	.13	-	-
	83	SAB	IT59	-	.17	208	.38	.15	.20	.16	-	-	-
ET	82	MTAS	MAT	-	.26	145	.31	.36	.42	.17	.18	-	.16
	83	SAB	RAM	MTAS	.23	166	.34	.31	.34	.24	.20	.29	-
ET	82	-	-	-	-	23	.38	-	-	-	-	-	-
FHS	83	SAB	-	-	.27	31	.52	.32	-	-	-	-	-
MB	82	IT59	-	-	.11	61	-	-	.31	.33	-	-	-
	83	MAT	-	-	.08	89	.25	.28	.19	-	-	-	-
MBK	82	MAT	-	-	.15	33	-	.39	.34	-	.32	-	-
FHS	83	SAB	-	-	.22	47	.47	.36	-	-	-	-	-
LRT	82	SAB	MTAS	-	.21	82	.40	.32	.31	-	-	-	-
	83	SAB	IT59	-	.27	68	.47	.34	.27	.25	-	-	.31
MBL	82	SAB	-	-	.18	30	.42	.42	-	-	-	-	-
FHS	83	SAB	-	-	.34	33	.58	-	-	-	-	-	.35
BI	82	WT72	-	-	.13	32	-	-	.34	-	-	-	.36
	83	MAT	RAM	-	.34	26	.40	.43	.34	-	.37	-	-
VM	82	SAB	MTAS	-	.44	26	.56	.36	.36	-	-	-	-
	83	MAT	-	-	.43	26	.61	.66	.34	.46	-	.40	-
INF	82	SAB	MTAS	-	.31	43	.49	.39	.28	-	-	-	-
	83	MTAS	SAB	-	.48	53	.51	.45	.56	.30	.28	.21	.32

cal characteristics - such as range of the scale, mode of awarding marks, symmetrical mistakes - between the two predictors and the criterion variable.

The result of the mathematical test (MTAS) correlates significantly with the criterion in nearly all university course of studies and crew subsamples. This doesn't apply to practice-orientated courses of studies. The other test predictors correlate only in few cases and in an inconsistent manner. In the subsamples electrical engineering (ET) and computer science (INF), there are significant correlations of the analytic-constructive (IT59), verbal (IT14), matrices (RAM), and general knowledge (WT72) predictors with the examination criterion, actually in the 1983 crew. These are cues that the combination of the diagnostic informations is subject to great fluctuations.

Consequences

Correlation coefficients are surely reduced due to unreliability of the criterion and school report measures, and due to restriction of the predictor variances, missing scores of unsuited officer candidates. Some subsamples are too small to receive meaningful results. Nevertheless, it should come to appropriate steps. Test informations and school results should be statistically combined as base-line information in a future sequentially designed network of diagnosis, guidance, and coaching to prepare cadets for their university studies and to establish a regular use of these informations for recommendations.

The dimensional and methodological aspects of psychological base-line and enlarged analysis and decision should be changed towards a more detailed information about studies-related requirements, demands, and standards, which subgroups of officer students have to cope, and towards longitudinallyvalidated instruments for entrance testing and questioning candidates and cadets. Psychological testing and course guidance service by military and university experts should be linked with increased individualized coaching in suitable training periods before the university entrance of the officer cadets.

References

Florin, Irmela & Rosenstiel, L. von (1976). Leistungsstörung und Prüfungsangst, Ursachen und Behandlung. München: Goldmann, 25-34.

Klassmann, Helga (1984). Fundamental Principles and Diagnostic Methods Used for the Selection of Officer Candidates. Proceedings 26th Annual Conference of the Military Testing Association. Volume II, 871-876.

Otte, R. (1985). Das Modell der Eignungsreihenfolge für Offizierbewerber der Bundeswehr. Arbeitsbericht aus dem Psychologischen Dienst der Bundeswehr. Bonn: Bundesministerium der Verteidigung P II 4.

Trost, G. (1985). Pädagogische Diagnostik beim Hochschulzugang, dargestellt am Beispiel der Zulassung zu den medizinischen Studiengängen. In: R.S. Jäger, R. Horn, K. Ingenkamp (Eds.). Tests und Trends Nr. 4, Jahrbuch der Pädagogischen Diagnostik. Weinheim: Beltz, 41-78.

Acknowledgements

The author likes to express thanks to LRD Klaus Puzicha Ph.D., head of department at GFAF Administration Office, who created concrete freedom for research ideas and change conceptions, to RD Friedrich W. Steege Ph.D., psychologist at FMOD - P II 4, who supported working and publication, and to ORR Rolf Otte Ph.D., research psychologist at CPO, who arranged for SPSS data processing and who reviewed the draft of this paper.

Short Versus Long Term Tenure as a Criterion for Validating Biodata

Elizabeth P. Smith and Clinton B. Walker[1]
U. S. Army Research Institute for the Behavioral and Social Sciences

This research tests the hypothesis that the traditional criterion for validating biodata in military research, viz. attrition during the first six months of service versus successful completion of that period, has produced less effective scoring keys and lower validities than a longer criterion period would. This hypothesis is based on two findings. First, at least half of attritions in previous research have occured after the first six months of service (Goodstadt & Yedlin, 1980; Hicks, 1981). Second, only half as many items in a 60-item biodata instrument were keyable at the six-month point as were at tenures of one to three years in data from 5,941 applicants to the Army in FY1981 and 1982 (Walker, 1985). If these findings are generally true, then keying on tenures longer than six months will move many first term attritions from the successful criterion group to the unsuccessful one, where they belong, and will produce a larger pool of keyable items. Both of those results should improve validity. In the present paper, items from the Army's Military Applicant Profile (MAP) are keyed on status at the 6-month and then at the 39 - 45 month point, depending on date of entry, and the validities are compared for those two criterion periods.

Method

Instrument

A 240-question research version of the MAP, which is a multiple choice biodata questionnaire, provided the items. Two forms of the instrument, with different sequences of the items, were used. In content, the questions deal with self-esteem, motives for enlisting, experiences in school, work experience, expectations of military life, social habits, experiences in the family, athletic activity, and miscellaneous other experiences.

Sample

The sample was 9,416 receptees at all seven Army Reception Stations who took the instrument in January-June 1982. This number included 7,653 males, of which 6,403 were high school graduates and 1,250 were non-graduates or GED holders. Also in the sample, but examined only for cross-validity, were 1,763 females, all high school graduates.

Criteria

All cases were divided into "stayers" and "leavers" as follows. Stayers were either on Active Duty at the end of the period being examined or had

---

[1]The opinions in this paper are the authors' and do not necessarily reflect views or policy of the Army Research Institute or the Department of the Army. Richardson, Bellows, Henry, and Company, Inc., under contract to Army Research Institute, developed the items for this work and collected the raw predictor data and six-month criterion data. Joseph Stephenson created the dataset with the longer tenures. We gratefully acknowledge his support.

been discharged for positive reasons (e.g., end of enlistment, transfer into an officer candidate program) or "no fault" reasons (e.g. medical, hardship). Leavers were cases who had been discharged for any causes other than those above. These latter cases were presumed to have been discharged early for any of various failures to adapt to Army life. For comparing short and longer tenures as criteria, the status of the cases was examined first at the end of the initial six months of service and then as of 1 October 1985, which was from 39 to 45 months after accession. Leavers after the first six months were in the successful group for the first analysis and in the unsuccessful group for the longer tenure.

Procedure

Empirically derived scoring keys were developed on a 60% sample of all of the males. To select items for keying, we ran item-level chi square tests on the frequencies with which the separate response choices were picked by the criterion groups (stay vs. leave). Items giving $p < .05$ were retained for keying. These items were keyed using a horizontal percentage method (Cascio, 1982; Riegelhaupt & Bonczar, 1985), weighted for differences in sizes of the criterion groups. These weighted percentages of stayers were then rounded and converted to single digit weights ranging from -1 to +3. The conversion rule was as follows: up to 24% stayers = -1; 25 to 34% = 0; 35 to 44% = 1; 45 to 54% = 2; >54% = 3. Under this rule for assigning weights, some items were weighted more heavily than others by having a wider range of possible scores.

Item scores for each case were summed and tested for differences between criterion groups. Then, point biserials were calculated on the relation between total scores and the dichotomous stay-leave criterion. After finding validities on the development sample, we computed validities on the independent holdout sample of all males, on two random samples of the females 60% and 40%), and on similar splits of the two male groups (graduates and non-graduates) which were subsets of the larger development and holdout groups. These procedures were followed first for the short criterion period (maximum service of six months) and then, on the same cases, for the longer criterion period.

As a check on whether the same items would be effective for predicting success over both short and long criterion periods, we divided items into those which were unique to each key (i.e., two sets) and those that were common to both keys. Total keyed scores for each set were then validated. We also ran a second kind of cross-validation to find how well each key works in predicting the the length of service on which it was not developed. That is, we calculated validities for the long-tenure key on the short criterion period and for the short-tenure key on the long criterion period.

Results

Table 1 shows how many items were keyable at both tenures and how many were uniquely keyable at only one. Validities for these sets of items and for the total set that was keyable for each condition (unique plus common) are

Table 1
Validities for males of sets of items that were keyable at only the short tenure, only the long tenure, and at both

Criterion	Tenure at Which Items Were Keyed: Items (n) — Short: Total (145)	Short: Unique (23)	Short: Common (122)	Long: Total (181)	Long: Unique (59)	Long: Common (122)
Short						
Development sample	.25	.17	.24	.18	.10	.21
Holdout sample	.19	.14	.19	.18	.11	.19
Long						
Development sample	.22	.09	.23	.31	.27	.30
Holdout sample	.18	.11	.18	.26	.25	.24

Note. The critical value for a difference between two independent correlation coefficients, one for the development sample (n = 4,594) and one for the holdout sample (n = 3,059), is .046 ($p < .05$, two-tailed).

Table 2
Validities by sample and by tenures for keying and for validating; rates of success

Group	N	Tenure on Which the Items Were Keyed: Short (145 Items) — Criterion length: Short	Criterion length: Long	% Stay	Long (181 Items) — Criterion length: Short	Criterion length: Long	% Stay
All Males							
Development	4,594	.25	.22	.87	.18	.31	.75
Holdout	3,059	.19	.18	.86	.18	.26	.75
Females							
Sample 1	1,077	.14	.11	.80	.14	.15	.77
Sample 2	686	.19	.15	.79	.16	.16	.76
Non-graduate males							
Sample 1	743	.20	.12	.79	.13	.19	.56
Sample 2	507	.20	.11	.80	.12	.19	.58
Graduate males							
Sample 1	3,888	.22	.20	.88	.17	.27	.79
Sample 2	2,515	.21	.20	.88	.16	.25	.79

also given. Validities and cross-validities at both the tenure for keying and the other tenure are given in Tables 1 and 2. Table 2 gives validities and success rates for various groups of cases: all males, females, graduate males, and non-graduate males.

Table 3
Descriptive statistics on development and holdout samples as a function of the tenure for keying items and the criterion for validating total scores

	Tenure on Which the Items Were Keyed							
	Short				Long			
Criterion	N	m	sd	t[a]	N	m	sd	t[a]
Short								
Development sample								
Stayers	3,993	252.5	15.6	14.01	3,993	257.3	20.3	11.27
Leavers	601	240.2	20.6		601	245.9	23.6	
Holdout sample								
Stayers	2,629	252.3	15.5	9.43	2,629	256.9	19.9	9.77
Leavers	430	243.1	19.2		430	246.7	21.3	
Long								
Development sample								
Stayers	3,457	253.0	15.5	14.02	3,457	259.6	19.4	20.64
Leavers	1,137	244.3	18.9		1,137	244.5	21.9	
Holdout sample								
Stayers	2,306	252.7	15.4	9.55	2,306	258.5	19.5	14.88
Leavers	753	245.7	18.1		753	246.2	20.1	

[a] $p = .0001$

Table 3 gives mean total scores and standard deviations for stayers and leavers in the development and cross-validation samples and results of t-tests on their means. These results are given for the cases where items were keyed and validated on the same and on different time periods.

Discussion

In five different respects, these data support the hypothesis that tenures longer than the traditional six months are better for keying and validating biodata. First, a full 46% of attrition in this sample occurred after the six-month point. Thus, a key developed at that point is degraded by the presence of almost half of the leavers in the successful criterion group. Second, while over half of the valid items are keyable at both the short and long tenures, more than twice as many are uniquely keyable at the longer one (59 vs 23). Thus a longer instrument results from extending the period for keying.

Third, validity and cross-validity are higher when items are keyed and validated on the longer period. It is true that congruence in the tenures for keying and validating (i.e., either Short key with Short criterion or Long key with Long Criterion) produce the highest sets of validities here; but still the original validity in the Short-Short condition (.25) does not exceed the cross-validity in the Long-Long condition (.26). Similarly, the Long key for the common items has as high a cross-validity for the Short criterion as does the Short key for any set of items, while it has a higher validity at the Long criterion than any set of items with the Short key does.

Fourth, shrinkage of cross-validities is less for item sets that are keyed at the long tenure. In Table 1 the median shrinkage for Short keys is .045 while for Long keys it is .02. Finally, the largest mean differences in total score, both in terms of keyed points and in t-value are for keying and validating at the longer tenure (Table 3).

The data in Table 1 support one other optimistic conclusion. Although the sets of unique items have fairly low validities for the criterion on which they were not keyable, the 59 items which were significant at only the long tenure have a good validity and cross-validity for the longer criterion. Among the highest validities in that table are those that come from this set of about one-third of the items that are useful over that longer period. This finding implies that there may be enough valid items to produce several test forms of satisfactory validity. Among other things, the issue of how to assign items to forms needs to be addressed.

A second topic for further research is that of possible differences in early and late leavers. If found, any such differences might help to explain differences between leavers and stayers. A comparison of the content of the two unique sets of items may yield some hypotheses on this issue.

Although these results confirm the statistical superiority of keying and validating on longer tenures, that practice has a cost: that of delaying implementation of the instrument while the criterion matures. One question for further research is how to balance the benefits of high validity with those of early implementability so as to maximize the net benefit.

The results for females and for non-graduate males are not as positive as for men overall. Whether a good unisex scoring key could be developed remains to be seen. From the the percents of stayers in Table 2, attrition seems to be a somewhat different process in males and females: unlike males' attrition, almost all of females' occurs in the first six months.

Even though the samples of females and non-graduate males are large in absolute numbers, they may not be large enough in these data to produce stable performance in a biodata instrument. Two aspects of the military research setting make results from validations of non-cognitive predictors relatively unstable. First, attrition is managed, and policy on acceptable levels thereof varies over the years. Thus the criterion is driven by at least one force that is not tightly connected with the characteristics of the examinees. Second, the characteristics of the applicant and accession pools also change over the years. For example, a decade ago about half of accessions were non-graduate males; now the rate is around 10%. These facts make

it important to use large, stable samples for developing keys.

Previous attempts to evaluate the validity of MAP in the operational setting have found validities to be much lower than in the research setting (Walker, 1984). Unlike the present research, past work in developing scoring keys has not cross-validated them. The robust cross-validities for the long-long condition here give reason to believe that the keys developed here would retain a good level of validity if put into operation. Even with that assumption, further research on the rates of accurate and inaccurate selection decisions to be expected should be carried out to see whether the instrument is likely to be cost-effective.

References

Cascio, W. F. (1982). Applied psychology in personnel management. Reston, VA: Reston.

Goodstadt, B. E. & Yedlin, N. C. (1980). First tour attrition: implications for policy and research (Research Report 1246). Fort 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.

Riegelhaupt, B. J. & Bonczar, T. P. (1985, October). The utility of educational and biographical information for predicting military attrition. Proceedings of the 27th Annual Meeting of the Military Testing Association. San Diego, CA

Walker, C. B. (1984, November). Validating the Army's Military Applicant Profile against an expanded criterion space. Proceedings of the 26th Annual Meeting of the Military Testing Association. Munich, FRG.

Walker, C. B. (1985, October). Three variables that may influence the validity of biodata. Proceedings of the 27th Annual Meeting of the Military Testing Association. San Diego, CA.

THE IMPACT OF INCREASED TRAINING TIME ON NATIONAL GUARD RETENTION

Glenda Y. Nogami
US Army Research Institute for the Behavioral
and Social Sciences

David W. Grissmer
Rand Corporation

Background

In 1983, the first Army National Guard round-out unit attended the National Training Center (NTC) with its Active Affiliate. This first unit, a Georgia Guard armored battalion, experienced a 15% loss in strength within six months of returning from NTC. Concern was expressed by the Commander, Fort Stewart (the home of the Active Affiliate), the Vice Chief of Staff of the Army, and the Deputy Chief of Staff for Personnel, that this loss was in some way related to the NTC experience. Since that first Georgia Guard unit, six other Guard round-out battalions have attended NTC. Although unit strength figures in these Guard units decline after NTC, none appear to have experienced the magnitude of loss that affected the first Guard unit. The differences in loss may potentially be attributed to differential organizational policies, unit personnel policies, local economic conditions, employer problems, and family issues. Many of the issues and problems are specific to the individual units and their surrounding locale. This paper summarizes some of the issues surrounding National Guard unit participation in the NTC.

Methodology

Focus group interviews were conducted with each of the seven National Guard round-out units that attended NTC. These interviews were conducted within a year of each unit's participation at NTC. Small group (4 to 9 persons) interviews were conducted separately for Unit Officers, NCOs, and in the later units (Alabama, North Carolina, and Lousiana), Junior Enlisted. Each interview lasted approximately two hours and was conducted during the unit's scheduled weekend drill. The following table lists the seven National Guard units, their Active Affiliates, and their NTC rotation dates.

---

[1]The views, opinions, and/or findings contained in this paper are those of the authors and should not be construed as official Department of Army position, policy or decision. An earlier version of this paper was presented at the Reserve Manpower, Personnel, and Training Research Workshop held at Monterey, California, 25-27 June 1986.

TABLE 1

National Guard Units Included in the Study

Guard (State)	Active Affiliate	NTC Rotation
Georgia	Stewart	9-22 Sept 83
Minnesota	Riley	19 Apr - 8 May 84
Georgia	Stewart	3-22 Oct 84
Georgia	Stewart	18 Mar - 6 Apr 85
Alabama	Polk	1-22 June 85
North Carolina	Carson	26 Jun - 15 Jul 85
Louisiana	Polk	11-31 Aug 85

Topics covered in the interviews included: events leading from notification through train-up through NTC to the time of the interview, compensation and pay issues, family issues, readiness and retention, and general training problems. As can be anticipated, some issues were more salient to certain groups, e.g., pay issues to unemployed guardsmen, or training issues to NCOs and Officers, etc.

Training Characteristics of a Guard NTC Rotation

Preparation for and participation at NTC required some changes to the "normal" Guard environment. The characteristics unique to NTC participation can be classified into three categories: training environment, increased training time, and specific personnel policies. Following NTC, the units went back to a more "normal" environment, so that the following were not in evidence at the time of the interviews.

Training environment.

Train-up for NTC was characterized by intensive weekend drilling with their Active Affiliate. In many cases, this training occurred on the active installation (FT Stewart, FT Polk, etc) using equipment with which the Guard had been unfamiliar. Training on the active installation sometimes meant an additional transportation time to and from training. For example, for the first Georgia unit, drills at FT Stewart required a 10 hour drive in each direction. Normally, weekend drilling would be accomplished at the local Armory or local training site with little or no Active Affiliate participation.

Increased training time.

The train-up was characterized by additional training/drill requirements. In addition to the required one weekend drill per month, there were additional weekend drills, and longer drills (MUTA 5's and 6's)for all Guard. For NCOs and Officers, there were also supplemental planning and leadership training during the week for no pay. On top of all of this preparatory training, NTC itself requires an extra week of annual training time - three weeks as opposed to the usual two-week AT. With all of these training requirements, it is no wonder that many NCOs and Officers reported not seeing their families on weekends for months at a time.

Personnel Policies.
In order to maximize train-up experiences for NTC, Guard units have found it necessary to implement certain personnel policies for the duration of train-up and NTC. The intent of these policies is to stabilize personnel in leadership and job positions for most effective training. In the first Georgia unit, this was translated into not allowing any Guard to leave the unit (either transferring to another unit or leaving the Guard) until after NTC. Some of the later units, learning from the experiences of the first unit, had a more flexible personnel policy. They allowed reasonable attrition (for cause) to occur and replaced these separatees with fillers from other in-state or out-of-state Guard units for the NTC exercises.

NTC Can be a Catalyst For Permanently Increased Readiness.
The train-up for NTC and the NTC experience were seen by all participants as the "best training", the "most realistic training" , the "most challenging training" around. Units reported being in their most ready posture after NTC even after sustaining strength losses. There was a certain pride about surviving NTC and winning and losing battles together with their Active Affiliate. The training for NTC and the experience of NTC itself may have the effect of increasing pressure from both the Guard and the Active Affiliate for continued quality training. This would be especially true for those Guard units with close bonding with their Active Affiliate. Close working relationships had increased both the Guard's and Active Affiliate's respect for each other. The total NTC experience can also be seen as a catalyst for increased readiness through improved personnel retention and selection. During the train-up, less productive and/or physically deficient NCOs and Officers were selectively "pruned" from the units. This was partly self-selection out of the unit by Guardsmen, themselves, and partly the commander's decision for selectively retaining high performing and committed personnel. One unit, using information gleaned from their NTC experience, has developed criteria for more selectively screening applicants for motivation and commitment. In addition to serving as a screen for highly motivated personnel, the NTC training itself increases Guard readiness. The knowledge of skills required and the experience of combat and combat training will stay with the unit personnel. There is an unanticipated readiness benefit from NTC: vicarious learning. Through increased communication, units in- and out-of-state are learning the lessons of NTC without going through a rotation. Fillers have played an important part in communicating NTC lessons learned. Fillers have taken the NTC experience into their home units and disseminated the information as "war stories". The lessons of the modern battlefield are effectively brought back anecdotally. These units have improved their drills to more accurately reflect the battlefields of today.

Potential Issues Associated with Guard NTC Participation.
Issues associated with the Guard NTC participation we discuss here center around three areas: (1) recruiting, (2) retention, and (3) training and readiness. Some of the recruiting issues have been addressed above. NTC may help establish criteria for recruit selection, and recruiting advertising. In the short term, the "macho" image of NTC training may help recruiting efforts by offering adventure, travel (to California), patriotism (against Soviet strategy), and escape from the mundane. Retention issues center around four factors: employer relations, family time, local economic conditions, and NTC scheduling. Most employer problems are at the first line supervisor level. Any

additional training that has a negative impact on the 40 hour work week or work team performance will worsen Guard-employer relations. Additional training that takes weekend, or worse, vacation time from the family will aggravate any family problems. Unfortunately, military leave does not cover a three-week annual training period; so consequently, many Guardsmen must use their annual leave to cover the additional training and NTC. Local economic conditions present a two-edged sword. If the local economy is stagnant (i.e., high unemployment), the additional drills and training times provide an excellent source of alternate income to unemployed Guardsmen. If, however, one is employed on a full time basis, the additional Guard requirements are more likely to cause conflicts with employers conflicts and possible dismissal. Additional training is seen as disruptive to the other workers and to productivity. Finally, the timing of NTC may be problematic. The Minnesota unit went to NTC in April. This was probably the worst time for them to attend NTC because many of the Guard were farmers and April was planting season. Finally, training and readiness pose potential issues. It is not clear how long readiness in a unit can be maintained following NTC. How often, then, should units be recycled through NTC to optimize retention of skills? For whom is NTC training most effective? There seems to be a consensus among the units that yearly NTC is too much; every 3 - 4 years is optimal. Yet, historical attrition data would indicate that there would probably be a greater than 50% turnover rate in that time. In early deployable units, this level of readiness may not be acceptable. NTC training seems to be more of a learning experience for NCOs and Officers than for the Junior Enlisted. NCOs and Officers reported seeing the "big picture", learning the importance of continuous field maintenance, use of sleep-wake cycles, delegation of responsibilities, and training in logistics and navigation as positive aspects of NTC training. The Junior Enlisted seemed to be less involved and to gain less from the experience.

Analytical Limitations

Caution should be exercised before drawing conclusions from this case study. (1) These units are not representative of Guard units: All were Mechanized Infantry or Armor units; and all but one are from the Southeast. (2) This has been a retrospective case study, which relies exclusively on the individual and collective memories of the unit. (3) There are unique sets of factors connected to each unit. The composition of the units varies from primarily textile workers to primarily farmers and students, etc. However, this case study does indicate avenues of future research.

Where Do We Go From Here?

The next logical step would be to develop a comprehensive case study of matched Guard units either undergoing NTC train-up or new equipment training. Both would entail additional training and drills, which would facilitate teasing out the unique problems of NTC. This comprehensive case study should include not just Officers, NCOs, and Junior Enlisted, but also employers, families, the Active Affiliate, Guard who have attrited. By surveying or interviewing all groups, one could get a more complete and accurate picture of the impact of train-up and NTC. This case study should start at the point the Guard unit is notified about their NTC rotation. This would provide a longitudinal, prospective case study which would offer more opportunities for unbiased perceptions. The opportunities for this research will grow as the number of Guard units scheduled to attend NTC grow, and Congress perceives the Reserves and Guard as a less costly alternative to the Active Component.

The Project A Concurrent Validation Data Collection[1],[2]

James H. Harris
Human Resources Research Organization

John P. Campbell
University of Minnesota

Charlotte H. Campbell
Human Resources Research Organization

Introduction

The purpose of this paper is to describe the Project A concurrent validation data collection and relate some "lessons learned" about the administration of large scale data collections. During this data collection, predictor and criterion measures were administered to approximately 9,500 entry-level soldiers and rating scales were administered to approximately 7,000 supervisors of these soldiers. The original Project A Research Plan specified a concurrent validation target sample size of 600-700 skill level (SL1) job incumbents for each of 19 MOS, using procedures that had been tried out and refined during the predictor and criterion field tests. The Research Plan specified 13 data collection sites in the United States (CONUS) and two in Europe (USAREUR). The number of sites was the maximum that could be visited within the Project's budget constraints, which dictated that sites be chosen to maximize the probability of obtaining the required sample sizes. The data collection schedule, by site, is shown in Figure 1.

The basic sampling plan, data collection team training, data collection procedures, and lessons learned are presented in the following sections.

Sampling Plan

The general sampling plan was to use the Army's World-Wide Locator System to identify all the first-term enlisted personnel in the 19 MOS at each chosen site who entered the Army between 1 July 1983 and 30 July 1984. If possible,

---

[1]This research was funded by the U. S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U. S. Army Research Institute or the Department of the Army.

[2]The material in this paper is from two sources: Campbell, C.H., & Hoffman R.G. (in press). Concurrent validation hands-on data collection: Lessons learned. Alexandria, VA: Human Resources Research Organization (HumRRO).

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute and Army Research Institute (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report. ARI Technical Report ____. Alexandria, VA: Army Research Institute.

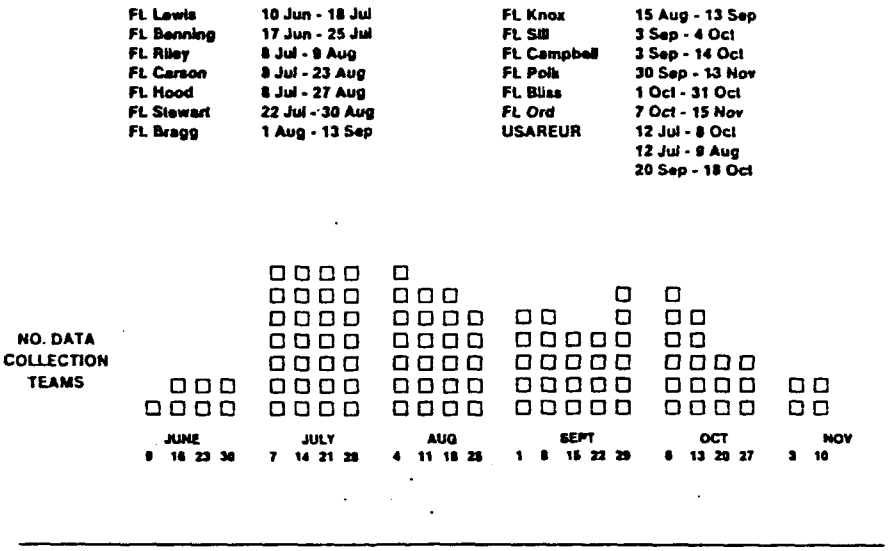


Figure 1. Concurrent validation schedule.

the individual's unit identification was also to be retained. The steps described below were then followed. The intent was to be as representative as possible while preserving enough cases within units to provide a "within rater" variance estimate for the supervisor and peer ratings.

A. Preliminary Steps

1. Identify the subset of MOS (within the sample of 19) for which it would be possible to actually sample people within units at specific posts. That is, given the entry date "window" and given that only 50-75 percent of the people on <u>any</u> list of potential subjects could actually be found and tested, what MOS are large enough to permit sampling to actually occur? List them.

2. For each MOS in the subset of MOS for which sampling is possible, identify the smallest "unit" from which 6-10 people can be drawn. Ideally, we would like to sample 4-6 units from each post and 6-12 people from each unit. For the total concurrent sample this would provide enough units to average out or account for differential training effects and leadership climates, while still providing sufficient degrees of freedom for investigating within-group effects such as rater differences in performance appraisal.

3. For the four MOS in the Preliminary Battery (PB) sample, identify the members of the PB sample who are on each post.

B. The ideal implementation would be to obtain the Alpha Roster list of the total population of people at each post who are in the 19 MOS and who fit our "window." The lists would be sent to the data collection manager where the following steps would be carried out.

1. For each MOS, randomize units and randomize names within units.

2. Select a sample of units at random. The number would be large enough to allow for some units being truly unobtainable at the time of testing.

3. Instruct the Point-of-Contact (POC) at the post to obtain the required number of people by starting at the top of the list and working down (as in the Batch A field test) within each of the designated units. If an entire unit is unavailable, go on to the next one on the list.

4. In those MOS for which unit sampling is not possible, create a randomized list of everyone on the post who fits the window. Instruct the POC to obtain the required number by going down the list from top to bottom (as in the Batch A field tests).

C. If it is not possible to bring the Alpha Roster to the data collection manager, provide project staff at the post to assist the POC in carrying out the above steps.

1. If it is not possible to randomize names at the post, first use the World-Wide Locator to obtain a randomized list, carry the list to the post and use it to sample names from units drawn from a randomized list of units. If there are only 6-8 units on the post, then no sampling of units is possible. Use them all.

D. If it is not possible for project personnel to visit the post, then provide the randomized World-Wide Locator list to the POC and ask him or her to follow the sampling plan described above with written and telephone assistance. That is, the POC would identify a sample of units (for those MOS for which this is possible), match the unit roster with the randomized World-Wide Locator list, and proceed down each unit until the required number of people was obtained. If the POC can generate their own randomized list from the Alpha Roster, so much the better. The World-Wide Locator serves only to specify an a priori randomized list for the POC.

E. If none of the above options is possible, then present the POC with the sampling plan and instruct him or her to obtain the required number of people in the most representative way possible (the Batch B procedure).

The final sample sizes are shown by post and by MOS in Figure 2. Note that it was not always possible in all MOS to find as many as 600 incumbents with the appropriate accession dates at the 15 sites. Some MOS simply aren't that big.

Data Collection Team Training

Each data collection team was composed of a Test Site Manager (TSM) and six or seven project staff members who were responsible for test and rating scale administration. The teams were made up of a combination of regular

	BATCH A									BATCH Z											
MOS / Location	11B	13B	19E	31C	63B	64C	71L	91A	95B	12B	16S	27E	51B	54E	55B	67N	76W	76Y	94B	Total	% Total
Ft Benning	46	23	41	7	13	39	16	9	13	13	16	3	0	12	10	9	13	15	12	316	3.35
Ft Bliss	0	26	30	15	61	45	17	9	44	15	6	2	0	14	8	12	6	31	30	347	3.68
Ft Bragg	65	46	0	0	37	25	41	10	72	82	75	13	19	72	20	7	42	30	62	730	7.74
Ft Campbell	96	25	0	25	60	45	54	44	43	90	23	10	6	32	10	42	51	61	46	767	8.83
Ft Carson	60	60	77	30	49	53	36	25	46	49	57	13	0	25	7	0	23	40	47	689	7.31
Ft Hood	25	56	0	30	40	25	38	30	80	81	60	4	12	52	36	44	72	41	57	767	8.13
Ft Knox	29	32	111	16	38	48	22	46	31	43	10	6	0	0	12	0	10	23	34	524	5.56
Ft Lewis	75	46	13	11	43	46	23	27	56	37	25	1	11	51	31	20	48	41	36	631	6.69
Ft Ord	30	0	0	14	30	42	31	43	51	51	7	0	1	4	7	15	23	40	28	425	4.51
Ft Polk	73	47	19	29	47	47	18	46	44	60	45	9	0	10	7	23	26	51	35	644	6.87
Ft Riley	30	43	65	27	28	46	35	30	40	31	20	0	2	25	52	0	20	39	45	579	6.14
Ft Sill	0	100	0	20	43	61	44	0	20	42	11	0	0	0	0	15	7	35	32	407	4.43
Ft Stewart	44	46	30	17	28	51	31	45	45	30	30	0	0	17	29	26	44	34	35	617	6.64
USAREUR	132	122	120	130	122	121	114	119	118	120	78	61	41	96	84	63	105	134	113	1963	20.8
Total	782	667	603	366	637	696	514	561	692	704	478	147	106	434	291	276	490	630	612	9430	
% Total	7.44	7.07	6.33	3.88	6.76	7.27	5.45	5.31	7.34	7.47	4.96	1.58	1.15	4.59	3.09	2.93	5.20	6.68	6.49		

Figure 2. Concurrent validation sample soldiers by MOS by location.

project staff and individuals (e.g., graduate students) specifically recruited for the data collection effort. The test site manager was an "old hand" who had participated heavily in the field tests. This team was assisted by eight NCO scorers (for the hands-on tests), one company-grade officer POC, and up to five NCO support personnel, all recruited from the post.

The project data collection teams were given three days of training at a central location. During this period, Project A was explained in detail, including its operational and scientific objectives. After the logistics of how the team would operate (transportation, meals, etc.) were discussed, the procedures for data entry from the field to the computer file were explained in some detail. Every effort was made to reduce data entry errors at the outset via correct recording of responses and correct identification of answer sheets and disketttes.

Next, each predictor and criterion measure was examined and explained. The trainees took each predictor test, worked through samples of the knowledge tests, and role played the part of a rater. Considerable time was spent on the nature of the rating scales, rating errors, rater training, and the procedures to be used for administering the ratings. All administrative manuals, which had been prepared in advance, were studied and pilot tested, role playing exercises were conducted, and hands-on instruction for maintenance of the computerized test equipment was given.

The intent was that by the end of the three-day session each team member would (a) be thoroughly familiar with all predictor tests and performance measures, (b) understand the goals of the data collection and the procedure

for avoiding negative critical incidents, (c) have had an opportunity to practice administering the instruments and to receive feedback, and (d) be committed to making the data collection as error-free as possible.

As noted above, eight NCO scorers were required for Hands-On test scoring. They were recruited and trained using procedures very similar to those used at each post in the criterion field tests. Training took place over one full day and consisted of (a) a thorough briefing on Project A, (b) an opportunity to take the tests themselves, (c) a check-out of the specified equipment, and (d) multiple practice trials in scoring each task, with feedback from the project staff. The intent was to develop high agreement for the precise responses that would be scored as GO or NO-GO on each step.

Data Collection Procedure

The data collection proceeded as follows: The first day was devoted to equipment and classroom set-up, general orientation to the data collection environment, and a training and orientation session for the post POC and the NCO support personnel.

On the first day of actual data collection the soldiers who arrived at the test site were divided randomly into two equal groups, identified as Group 1 or 2. Each group was directed to the appropriate area to begin the administration for that group. They rotated under the direction of the test site manager through the appropriate block according to the schedule.

For soldiers in a Batch Z MOS, like 12B, the procedure took one day. For soldiers in a Batch A MOS, like MOS 91A, the procedure was similar but took two days to rotate the soldiers through the appropriate blocks. The measures administered in each block are shown in Figure 3.

BATCH A MOS 4 Blocks 4 Hrs. Each		BATCH Z MOS 2 Blocks 4 Hrs Each	
Block 1	Predictor Tests	Block 1	Predictor Tests
Block 2	School and Job Knowledge Tests Army-Wide Ratings	Block 2	School and Job Knowledge Tests Army-Wide Ratings
Block 3	MOS Specific Hands-On Tests		
Block 4	MOS Ratings MOS Specific Written Tests		

Figure 3. Concurrent validation test outline.

Lessons Learned

Collecting data from 16,000 soldiers in 15 locations over six months is a difficult task, one that requires careful planning, attention to detail, an ability to adapt, a fondness for crisis management, and a special relationship with the telephone. For anyone planning an effort of like grandeur (or even grander), a few lessons learned from some of the survivors seems appropriate. We divide the lessons into three categories: planning, coordinating, and operating Each category is briefly discussed below.

Planning. Start as early as possible (18 months before collecting data) to identify the support you will need, to include personnel, equipment, facilities, and time requirements. Once you know what you need and when you need it, schedule a series of briefings with the Commanders. Start at the top with the CG of FORSCOM, TRADOC, and USAREUR and work your way through a series of briefings until you reach the local POC responsible for seeing that you get what you need when you need it. Be prepared to change your plans at each step to meet local concerns. Once you meet and brief your POC, you can begin coordinating.

Coordinating. The closer the time to begin data collecting, the more frequently you will speak to the POC. Expect to speak daily when you get within 30 days of data collection. In some instances, you may have to make a trip to the installation for a final coordination meeting. Be prepared to be very flexible with regard to the installation's internal schedule.

Operating. Most of the lessons learned in this category have to do with hands-on testing.

1. Many instances of equipment variation can be (and were) anticipated. Test developers and site coordinators must find out what major pieces of equipment are not likely to be available at the selected sites in advance of actual testing if high quality tracked tests are to be prepared.

2. Printed scoresheets must be proofed carefully to ensure that for every step which should be scored, a score can be recorded.

3. Scorers must be thoroughly trained, not only on how to set up and administer the tests, but also on how to record data on the scoresheets. They must be given practice in using the scoresheets (not just talked through it) before testing, and monitored closely during testing, especially with the first few soldiers tested. Continual monitoring must also occur throughout the testing.

4. Scorers and hands-on managers must document meticulously who was tested on what, and also who wasn't tested on what, and why.

5. Experienced hands-on managers are often able to implement procedures to deal with equipment malfunctions or variations, but these too must be documented.

6. Completed scoresheets must be checked as soon as possible after testing so that careless or incorrect scoring can be detected, and the errant scorer can be retrained.

THE DEVELOPMENT OF A MODEL OF THE PROJECT A CRITERION SPACE[1]

John P. Campbell
University of Minnesota

Lawrence M. Hanser
Army Research Institute

Lauress Wise
American Institutes for Research

Conceptual Background

The goals of performance measurement in Project A are to define, or model, the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor. The performance measures are to serve as criteria for validating selection/classification tests, and not, at this point, as operational appraisals.

Some additional specific goals are to: a) make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency, b) compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multi-trait, multi-method approach), c) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and d) evaluate existing archival and administrative records as possible indicators of job performance.

Given these intentions, the criterion development effort focused on three major methods: hands-on job sample tests, multiple choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in the development of the rating methods.

Modeling Performance

The development efforts to be described were guided by a particular "theory" of performance. The basic outline is as follows.

First, job performance really is multi-dimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization (Army).

Two General Factors

For the population of entry level enlisted positions we postulated that there are two major types of job performance components. The first are specific to a particular job. That is, measures of such components would reflect specific technical competence or specific job behaviors that are not

[1]This research was funded by the U. S. Army Research Institute for the Behavioral and Social Sciences, Contract No. MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions or policies of the U. S. Army Research Institute or the Department of the Army.

required for other jobs. We anticipated that there would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills.

The second kind of performance factors include components that are defined and measured in the same way for every job. These are referred to as Army-wide criterion factors and incorporate the basic notion that total performance is much more than task or technical proficiency. It might include such things as contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity.

Factors vs. a Composite

Saying that performance is multi-dimensional does not preclude using just one index of an individual's contributions to make a specific personnel decision (e.g., select/not select, promote/not promote). As argued by Schmidt and Kaplan (1971) some years ago, it seems quite reasonable for the organization to scale the importance of each major performance factor relative to a particular personnel decision that must be made and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision.

A Structural Model

If performance is characterized in the above manner, then a more formal way to model performance is to think in terms of its latent structure, postulate what that might be, and then resort to a confirmatory analysis. Within limits, this is what we tried to do. Unfortunately, it is true that we simply know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter.

Unit vs. Individual Performance

Finally, people do not usually work alone. Individuals are members of work groups or units and it is the unit's performance that frequently is the most central concern. Project A has not incorporated unit effectiveness in its model of performance. The project is focused on the development of a new selection/classification system for entry level personnel and is concerned with improving personnel decisions about individuals and not units. The task is to maximize the average payoff per individual selected.

What we have chosen to do is to try to identify the factors, or means, by which individuals contribute to unit performance and to assess individual performance on those factors via rating methods.

Criterion Development

Actual criterion development proceeded from two basic types of information. First, all available task descriptions were used to generate a population of job tasks for each MOS. The principal sources of task

description are the Army's periodic job description surveys and the Soldier's Manual for each MOS which is a specification by management of what the task content of the job is supposed to be. After much editing, revising to insure non redundancy and a uniform level of generality, and a formal review by a panel of subject matter experts, a population of 130-180 tasks was enumerated for each MOS.

An additional series of expert judgments was then used to scale the relative difficulty and importance of each task and to cluster tasks on the basis of content similarity. Sampling tasks for measurement was accomplished via a kind of Delphi procedure. That is, each member of a team of task selectors was asked to select 30 tasks from the population of tasks such that those selected were representative of task content, were important, and represented a range of difficulty. The individual judge's choices were then regressed on the task characteristics and both the choices and the captured "policy" of each person were fed back to the group members, who each revised their choices as they saw fit. The consensus of the task selection panel was then thoroughly reviewed by the Army command responsible for that particular job. This last review was the "final" word on the representativeness of task samples and produced a sample of 30 tasks for each job.

Standardized job samples, the paper-and-pencil job knowledge tests, and numerical ratings scales were then constructed to assess knowledge and proficiency on these tasks. Each measure went through multiple rounds of pilot testing and revision. The job sample tests were fairly elaborate and were composed of multiple test stations sometimes spread over a football field size area. Because of time limitations (4 hours), only 15 of the tasks could be tested hands-on.

The second procedure used to describe job content was the critical incident method. Panels of NCO's and officers generated thousands of critical incidents of effective and ineffective performance. There were two basic formats for the critical incident workshops. One asked participants to generate incidents that potentially could occur in any job. The second type focused on incidents that were specific to the content of the particular job under consideration. The behaviorally anchored rating scale procedure was used to construct rating scales for performance factors specific to a particular job (MOS-specific BARS) and performance factors that were defined in the same way and relevant for all jobs (Army-wide BARS). The critical incident procedure was also used with workshops of combat veterans to develop rating scales of "expected" combat effectiveness.

Since one major objective was to determine the relationships between training performance and job performance and their differential predictability, if any, a comprehensive training achievement test was constructed for each MOS by carefully matching the content of the program of instruction (POI) with the content of the population of job tasks, and writing items to represent each segment of the match.

The final entry in the array of criterion measures was produced by a concerted effort to get what we could from the files or archival records. We began by enumerating all possibilities from three major sources of such records: the enlisted master file, the enlisted military personnel file, and the military personnel records jacket (the 201 File).

We systematically compared these three sources using a sample of 750 people and a standardized information recording form. The 201 file looked the most promising in terms of recency and completeness, but of course, it is by far the most expensive to search. As a consequence, we collected eight archival performance indicators via a self report questionnaire. That is, people were asked what was in their personnel file as regards letters of commendation, disciplinary actions, etc. Field tests on a sample of 500 people showed considerable agreement between self report and archival records, for both positive and negative things. Further follow-up questionnaires and interviews suggested that self report may be the more accurate. The self report items were combined into five indicators that were actually used as criterion measures.

Determining Actual Criterion Scores

The first step in our analyses was to identify the basic criterion scores whose structure we would analyze. If all the rating scales are used separately and the MOS-specific measures are aggregated at the task or instructional module level, there are approximately 200 criterion scores on each individual. Some aggregation was needed.

Reduction of the Hands-On and Written Variables

The 30 tasks sampled for each job were clustered via expert judgment into 8 to 15 functional categories on the basis of similarity of task content. Each of the school knowledge items was similarly mapped into a specific functional category.

Ten of the functional categories were common to some or all of the jobs (e.g., first aid, basic weapons, field techniques). Each job also had two to five functional performance categories that were unique.

After category scores were computed, separate factor analyses were executed for each type of measure within each job. There were several common features in the results. First, the unique functional categories for each job tended to load on different factors than the common functional categories. Second, the factors that emerged from the common functional categories tended to be fairly similar across the nine different jobs and across the three methods.

Using the empirical factor analysis to guide us, we adopted a set of content categories which became the performance test scores used in subseuqent analyses.

Reduction of the Rating Variables

The individual ratiny scales were, for the most part, highly reliable. Empirical factor analyses of the Army-wide rating scales suggested three factors. These were:

1. Effort/Leadership, including effort and competence in performing job tasks, leadership, and self-development.
2. Maintaining Personal Discipline, including self-control, integrity, and following regulations.

3. Physical Fitness and Military Bearing, including physical fitness and maintaining proper military bearing and appearance.

Similar factor analyses were reviewed for the job-specific scales for each job. Two factors were identified based on these results. The first consisted of those aspects of job performance that were central to the specific technical content of each job. The second factor included the remaining, less central job performance components.

The individual items in the combat performance prediction battery also were subjected to an empirical factor analysis. Two factors emerged. The first factor consisted of items depicting exemplary effort, skill, or courage under stressful conditions. The second factor consisted of negatively worded items portraying failure to follow instructions and lack of discipline under stressful conditions.

Building the Target Model

The next step was to build a target model of job performance that could be tested for goodness of fit within each of our nine jobs. The project began with an initial model of performance (Borman, Motowidlo, Rose, & Hanser, in press) which had been modified on the basis of field test data (Campbell & Harris, 1985). Principal components factor analyses within MOS were used to suggest further modifications.

Several consistent results were observed. First, the expected "method" factors appeared, specifically one factor for the ratings and one for the written tests. The evidence for a "hands-on" method factor was less compelling. Second, the nature of the substantive factors tended to be similar across MOS.

Based on the empirical analyses, a revised model was constructed to account for the correlations among our performance measures. This model included five job performance constructs and two measurement method factors.

Confirming the Model Within Each Job

The next step in the analysis was to conduct separate tests of goodness of fit of this target model within each of the nine jobs. This was done using the LISREL confirmatory factor analysis program (Joreskog & Sorbom, 1981).

In conducting a confirmatory factor analysis with LISREL, it is necessary to specify the structure of three different parameters matrices: the hypothesized factor structure matrix (a matrix of regression coefficients for predicting the observed variables from the underlying latent constructs); the matrix of uniqueness of error components (and intercorrelations); and a matrix of covariance among the factors. In these analyses, we set the diagonal elements of the covariance matrix to one, forcing a "standardized" solution. This meant that the off-diagonal elements would represent the correlations among and between our performance constructs and method factors. We further specified that the correlation among the two method factors and each performance construct should be zero. This effectively defined the method factor as that portion of the common variance among measures from the same method that was not predictable from (i.e., correlated with) any of the other related factor or performance construct scores.

To be perfectly clear, the approach we used was obviously not purely confirmatory. The hypothesized target model was based in part on analyses of these same data.

Confirmation of the Overall Model

Given the certain amount of prior examination of the data described above, the results of the confirmatory procedures applied to each job seemed to support a common structure of job performance. The procedures also yielded reasonably similar estimates of the intercorrelations among the constructs and of the loadings of the observed variables on these constructs across the nine jobs.

The final step in our analyses was to determine whether the variation in some of these parameters across jobs could be attributed to sampling variation. The specific model that we explored stated that: (1) the correlation among factors was invariant across jobs and (2) the loadings of all of the Army-wide measures on the performance constructs and on the rating method factor were also constant across jobs.

The overall model fit extremely well. The root mean square residual was .047, and the chi-square was 2508.1. There were 2403 degrees of freedom after adjusting for missing variables and the use of the data in estimating uniqueness. This yields a significance level of .07, not enough to reject the model.

Summary and Discussion

Some aspects of the final structure are noteworthy. First, in spite of some confounding with measurement method, the latent performance structure appears to be composed of very distinct components. It is reasonable to expect that the different performance constructs would be predicted by different things, so that validity generalization may not exist across the performance constructs within a job. If this is so, there is a genuine question of how the performance constructs should be weighted in forming an overall appraisal of performance for use in personnel decisions. Using regression techniques to partial the methods factors from the substantive factors should also tell us more about what does or does not predict the residual variance.

Finally, since (a) the five-factor solution is stable across jobs sampled from this population, (b) the performance constructs seem to make sense, and (c) the constructs are based on measures carefully developed to be content valid, it seems safe to ascribe some degree of construct validity to them.

References

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (in press). Development of a model of soldier effectiveness (Technical Report ____). Alexandria, VA: U. S. Army Research Institute.

Campbell, J. P., & Harris, J. H. 1985). Criterion reduction and combination via a participation decision-making panel. Paper presented at the 93rd Annual Meeting of the American Psychological Association, Los Angeles.

Joreskog, K. C., & Sorbom, D. (1981). LISREL VI: Analysis of Linear Squares methods. Uppsala, Sweden: University of Uppsala.

Schmidt, F. L., & Kaplan, L. B. (1977). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.

NEW PREDICTORS OF SOLDIER PERFORMANCE

Norman Peterson, Leaetta Hough, Steve Ashworth, and Jody Toquam
Personnel Decisions Research Institute

Introduction

New predictors of soldier performance have been developed as part of Project A. Previous papers presented to this association have described the theoretical approach, development, and pilot and field testing of those predictors (Hough, McGue, Kamp, Houston, & Barge, 1985; McHenry & Toquam, 1985; Peterson, 1985; Rosse & Peterson, 1985; Toquam, Dunnette, Corpe, & Houston, 1985). Very briefly, those papers showed that a construct-oriented approach was utilized to identify and develop new measures that would complement the Armed Services Vocational Aptitude Battery (ASVAB) in terms of abilities measured and likelihood of increasing the prediction of training and job performance. Both paper-and-pencil and computer-administered measures were developed to tap constructs in cognitive (primarily spatial) ability, perceptual/psychomotor, temperament, biographical, and vocational interest domains. Pilot and field testing results showed the new measures were psychometrically sound and were measuring constructs relatively unique from the ASVAB.

This paper describes some of the results of analyzing the properties of the new measures, collectively called the Trial Battery, as exhibited in the concurrent validity sample of Project A. This sample consisted of over 9,000 active duty soldiers in their first three years of service, from 19 different military occupational specialties. Other papers in this symposium provide more detailed descriptions of the data collection procedures and job performance criteria also collected from that sample (Harris, 1986; Campbell, Hanser, & Wise, 1986).

New Predictor Factor Scores

The Trial Battery consisted of three major types of instruments: 1) six timed paper-and-pencil tests of cognitive spatial ability, 2) ten computer-administered tests of perceptual/psychomotor ability, and 3) three untimed paper-and-pencil inventories measuring temperament/biographical data (the Assessment of Background and Life Experiences or ABLE), vocational interests (the Army Vocational Interest Inventory or AVOICE), and job reward preferences (the Job Orientation Blank or JOB); collectively referred to as non-cognitive inventories.

Over 60 separate scores are obtained from the full Trial Battery. Space does not allow presentation here of statistics for all these scores. We used principal components factor analysis (varimax rotation) to identify a smaller number of factor scores for use in validity analyses. Examination of these solutions led us to choose 19 factor scores; these were formed by simply summing the scores that defined each factor, not by using a multiple-regression, factor-scoring method. Therefore, we are here using the term factor to denote simply a higher-order organization of Trial Battery test scores, and do not intend these factors as representations of underlying pyschological constructs. These 19 factors are simply a parsimonious method of combining the larger number of individual scale scores for purposes of validity analyses in a way that is faithful to their covariances. Table 1 shows the names of these factors, the number of scores making up the factor, the median reliability coefficients of the scores entering each factor, and the median uniqueness estimate of the factor. Figure 1 shows the names of the scale scores that made up each factor, organized by type of instrument.

The medians of the internal consistency reliability coefficients range

from .46 to .93; mean = .78. All but four are greater than .70. One of these, General Reaction Accuracy, is the sum of percent correct scores on very simple, computerized perceptual tasks. These scores have, by design, severely restricted variance--we were concerned primarily with General Reaction Speed which does have high reliability. The other three factors with relatively low internal consistency reliability are from the Job Orientation Blank, especially the Routine Work and Job Autonomy factors. These are really just single scale scores, with only three or four items on each scale, which probably accounts for the low values.

The test-retest reliabilities range from .13 to .85; mean = .67. The paper-and-pencil measures all have reliabilities of .70 or greater, with the exception of Food Service Interests which is .66. The reliabilities of the computer-administered measures, however, are between .46 and .62, except for the .13 value for General Reaction Accuracy which we discussed above. Although these values are not as high as we would like, keep in mind that these computerized tests are all relatively short (all ten tests are administered in about one hour). Measures that prove most valid could be lengthened to increase reliability. Also, we point out that these are retest intervals of two to four weeks; test-retest coefficients reported for computerized tests are often same-day or next-day intervals which, of course, would yield much higher coefficients.

The uniqueness coefficients in Table 1 are indexes of the amount of reliable variance that does not overlap with, or is unique from, other measures--in this case, the ASVAB. The higher this index, the greater the opportunity for incremental validity (over ASVAB). These values range from .40 to .90; mean = .71. The Trial Battery measures, as a whole, do appear to have high potential for incremental validity, especially for the non-cognitive measures.

In sum, with a few exceptions, the Trial Battery factors appear reliable and relatively unique based on analyses of this large, concurrent validity sample. We add that these results are highly similar to those reported a year ago on a much smaller sample (about 200).

Prediction of Job Performance

Table 2 shows results of initial analyses of the validity of new predictors for predicting job performance and Table 3 shows results of initial analyses of the Trial Battery's incremental validity (over ASVAB) for predicting job performance.

There are five criterion factors shown in both tables. The first two represent "can do" factors and are made up largely of hands-on and written job knowledge test scores (labeled Core Technical Proficiency and General Soldiering Proficiency). The last three represent "will do" factors and are made up largely of peer and supervisor ratings on behaviorally-anchored rating scales and self-reported administrative actions, such as awards and Articles 15 (labeled Effort and Leadership; Personal Discipline; and, Physical Fitness and Military Bearing). As earlier stated, Campbell, et al. (1986) report in more detail the development of these criteria.

Six predictor composites are shown in Table 2, one made up of four factors derived from the ASVAB; the other five made up from the Trial Battery factor scores, combined within instrument type. The composites were formed via multiple regression.

Several things are noteworthy about Table 2. First, it shows the ASVAB does an excellent job of predicting the "can do" criteria, a moderately good job for one of the "will do" factors (Effort), and not very well for two of the "will do" factors. Second, it shows that the Spatial and

Perceptual/Psychomotor composites from the Trial Battery follow a pattern similar to the ASVAB, but do not outpredict the ASVAB. We point out that the perceptual/psychomotor, computer-administered battery requires about 60-75 minutes to administer, but yields validities of .49 and .56 for the "can do" criteria. Also, the six spatial tests require about 90 minutes to administer, and do nearly as well as the ASVAB. Finally, the non-cognitive portions of the Trial Battery do only moderately well at predicting the "can do" criteria, but the ABLE equals or outperforms the ASVAB and the cognitive/perceptual/psychomotor portions of the Trial Battery for predicting the "will do" criteria. Indeed, the ABLE is 13 and 16 points higher than the ASVAB for the Discipline and Fitness/Bearing criteria. All in all, the overall pattern of the findings in Table 2 is about what we expected.

Table 1

Trial Battery Factors, Number of Scores in Each Factor, Median Reliability Coefficients and Uniqueness Estimates of Scores in Each Factor

Composite	Number of Scores	Median Reliability[1] Coefficients Internal Consistency	Test-Retest	Median Uniqueness[2]
Overall Spatial	6	.83[3]	.70	.55
Psychomotor	6	.80	.62	.71
Perceptual Speed/Accuracy	6	.80	.57	.72
Number Speed/Accuracy	4	.91	.58	.67
General Reaction Speed	2	.93	.46	.90
General Reaction Accuracy	2	.52	.13	.45
Achievement	3	.82	.78	.81
Dependability	2	.77	.77	.74
Adjustment	1	.81	.74	.79
Physical Condition	1	.84	.85	.83
Skilled Technician Interests	7	.89	.75	.82
Structure/Machines Interests	4	.92	.81	.75
Combat-Related Interests	3	.90	.80	.75
Audiovisual Arts Interests	3	.83	.74	.81
Food Service Interests	2	.81	.66	.78
Protective Service Interests	2	.83	.76	.81
Organization/Co-Worker Support	4	.67	N/A	.65
Routine Work	1	.46	N/A	.40
Job Autonomy	1	.50	N/A	.47

Note: N varies, but all > 7,000

[1] These are odd-even coefficients, corrected with Spearman Brown procedure, or coefficient Alpha for internal consistency and correlations over a two-four week interval, N=470, for test-retest.

[2] Uniqueness = $R - R^2$, where R = internal consistency reliability estimate and R^2 = squared multiple correlation of all ASVAB tests with each new predictor.

[3] This is based on a separately-timed, split-half coefficient collected during pilot testing, N = 118, because some of these tests are speeded, making odd-even coefficients inappropriate.

FROM PAPER-AND-PENCIL TESTS

Overall Spatial
- Assembling Objects Test
- Map Test
- Maze Test
- Object Rotation Test
- Orientation Test
- Figural Reasoning Test

FROM COMPUTERIZED MEASURES

Psychomotor
- Cannon Shoot Test (Time Score)
- Target Shoot Test (Time To Fire)
- Target Shoot Test (Log Distance)
- Target Tracking 1 (Log Distance)
- Target Tracking 2 (Log Distance)
- Pooled Mean Movement Time

Perceptual Speed and Accuracy
- Short Term Memory Test (Percent Correct)
- Perceptual Speed & Accuracy Test (Decision Time)
- Perceptual Speed & Accuracy Test (Percent Correct)
- Target Identification Test (Decision Time)
- Target Identification Test (Percent Correct)

Number Speed and Accuracy
- Number Memory Test (Percent Correct)
- Number Memory Test (Initial Decision Time)
- Number Memory Test (Mean Operations Decision Time)
- Number Memory Test (Final Decision Time)

General Reaction Speed
- Choice Reaction Time
- Simple Reaction Time

General Reaction Accuracy
- Choice Reaction Percent Correct
- Simple Reaction Percent Correct

FROM NON-COGNITIVE INVENTORIES

Organizational and Co-Worker Support (JOB)
- Job Pride
- Job Security Comfort
- Serving Others
- Ambition

Routine Work (JOB)
- Routine

FROM NON-COGNITIVE (CONTINUED):

Job Autonomy (JOB)
- Autonomy

Achievement (ABLE)
- Self-Esteem Scale
- Work Orientation Scale
- Energy Level Scale

Dependability (ABLE)
- Conscientiousness Scale
- Non-Delinquency Scale

Adjustment (ABLE)
- Emotional Stability Scale

Physical Condition (ABLE)
- Physical Condition Scale

Skilled Technician Interest (AVOICE)
- Clerical/Administrative
- Medical Services
- Leadership/Guidance
- Science/Chemical
- Data Processing
- Mathematics
- Electronic Communications

Structural/Machines Interest (AVOICE)
- Mechanics
- Heavy Construction
- Electronics
- Vehicle/Equipment Operator

Combat Related Interest (AVOICE)
- Combat
- Rugged Individualism
- Firearms Enthusiast

Audiovisual Arts Interest (AVOICE)
- Drafting
- Audiographics
- Aesthetics

Food Service Interest (AVOICE)
- Food Service Professional
- Food Service Employee

Protective Services Interest (AVOICE)
- Law Enforcement
- Fire Protection

Figure 1. Test and inventory scale scores making up Trial Battery Predictor Factors.

Table 2

Multiple Correlation[1] of Six Independent Predictor Composites with Each of Five Job Performance Criterion Factors.

	PREDICTORS					
CRITERION FACTORS	ASVAB[2] Composite K = 4	Spatial Abilities Composite K = 1	Perceptual/ Pyschomotor Abilities Composite (Computerized) K = 5	JOB Composite (Preferences) K = 3	ABLE Composite (Temperament/ Biodata) K = 4	AVOICE Composite (Interests) K = 6
1. Core Technical Proficiency	.60	.54	.49	.26	.24	.33
2. General Soldiering Proficiency	.66	.64	.56	.29	.25	.37
3. Effort and Leadership	.35	.28	.27	.19	.34	.26
4. Personal Discipline	.19	.16	.14	.11	.32	.15
5. Physical Fitness & Military Bearing	.21	.11	.11	.12	.37	.12

Note: Entries in the table are averaged across 9 Army MOS with complete sets of Job Performance Criterion measures. Total sample size is 3902. Sample sizes range from 281 to 570; median = 432.

[1] Multiple Rs are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2] K = the number of predictor scores in the composite.

Table 3

Increments in Multiple Correlations[1] (Over R Using ASVAB Composite) as A Function of Adding Trial Battery Factor Scores for Each of Five Job Performance Criterion Factors.

	CRITERION FACTORS				
PREDICTOR	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Personal Discipline	Fitness & Bearing
ASVAB[2] Composite Alone (K = 4)	.60	.66	.35	.19	.20
ASVAB Plus Trial Battery Factors (K = 23)	.64	.70	.45	.37	.42
Increment	.04	.04	.10	.18	.22

Note: Entries in the table are averaged over 9 Army MOS with complete sets of criterion measures. Total sample size is 3902. Sample sizes within MOS range from 281 to 570; median = 432.

[1] Multiple Rs are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2] K = the number of predictor scores in the composite.

While the AVOICE does not show higher prediction than the ASVAB for the "can do" criteria, it is interesting that it correlates .33 and .37 with those criteria. The AVOICE was intended primarily to assist in classification rather than prediction per se, so it is encouraging to see these correlations with "can do" criteria. Finally, with respect to Table 2, we note that the JOB, ABLE, and AVOICE are expected to add most to the prediction of attrition; those analyses have not been done yet.

Table 3 shows a first, very crude look at the incremental validity of the Trial Battery. In these analyses, we simply added all 19 Trial Battery Factor scores to the ASVAB factor scores and looked at the increase in the mulitiple correlation. The third row in Table 3 shows that 1) the prediction of all five criteria is increased, 2) little increase occurs for the "can do" criteria, and 3) sizeable increases occur for the "will do" criteria.

Efforts are underway now to make more refined Trial Battery composites and to estimate the classification efficiency increments obtained via use of the Trial Battery. These initial results, however, show that the new predictors do 1) predict soldiers' job performance at meaningful levels in the way that was expected and 2) make meaningful increments over the ASVAB to validity for important aspects of soldiers' job performance.

References

Campbell, J., Hanser, L., & Wise, L. (1986). *The development of a model of Project A criterion space.* Paper presented at the 28th Annual Military Testing Association Conference Mystic, Connecticut.

Harris, J. (1986). *The Project A concurrent validation data collection.* Paper presented at the 28th Annual Military Testing Association Conference, Mystic, Connecticut.

Hough, L. M., McGue, M. K., Kamp, J. D., Houston, J. S., & Barge, B. N. (1985). *Measuring personal attributes: Temperament, biodata, and interests.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

McHenry, J., & Toquam, J. L. (1985). *Computerized assessment of perceptual and psychomotor abilities.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Peterson, N. G. (1985). *Mapping predictors to criterion space: Overview.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Rosse, R. L., & Peterson, N. G. (1985). *Using microcomputers for assessment: Practical problems and solutions.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., & Houston, J. S. (1985). *Adding to the ASVAB: Cognitive paper-and-pencil measures.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

Note: This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences. Contract Number MDA903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

ASVAB VALIDITIES USING IMPROVED JOB PERFORMANCE MEASURES

Lauress L. Wise, Jeffrey J. McHenry - American Institutes for Research
Paul G. Rossmeissl - U.S. Army Research Institute
Scott H. Oppler - American Institutes for Research

Project A job performance measures are unique in their combination of depth (work samples, ratings, knowledge tests, and administrative measures) and breadth (19 very diverse jobs). This paper examines the validity of the Army's ASVAB Aptitude Area (AA) Composites for predicting job performance as assessed by these new measures. Project A performance measures have been organized into five constructs (Wise, Campbell, McHenry, Hanser, 1986). Four of these constructs (General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing) are the same for each Military Occupational Specialty (MOS). Armed Forces Qualifying Test (AFQT) scores and other selection criteria (e.g. high school graduation, moral and physical requirements) are designed to predict performance on these common constructs. The fifth construct, Core Technical Proficiency (CTP), covers aspects of job performance unique to each MOS. AA scores, used as job specific selection criteria, are appropriately validated against this construct.

In addition to evaluating current AA composites, we identified specific alternative composites. We did not identify alternative composites for every MOS, since we had data for only 19 of the more than 250 entry-level MOS. Instead, we identified alternative composites for each cluster of jobs that currently use the same AA composite. In this paper, we only considered redefining the existing composites. We did not consider changing the assignment of MOS to specific composites.

Methods

Current forms of the ASVAB generate nine subtest scores: General Science (GS), Arithmetic Reasoning (AR), Verbal (VE combining Work Knowledge and Paragraph Comprehension), Coding Speed (CS), Numerical Operations (NO), Auto/Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). AA composites are defined as unweighted sums of four or fewer of the standardized subtest scores. There are 255 such possible composites (126 using four subtests, 84 using three, 36 using two, and 9 using a single subtest). We evaluated all of them.

Project A Concurrent Validation (CV) data were used in evaluating the current composites. The CV data included the new job performance measures applied to over 9,000 soldiers in 19 different MOS. Table 1 shows CV sample sizes by MOS and race and gender and also the ASVAB subtest and the CTP criterion means and standard deviations.

This research was funded by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract Number MDA903-82-C-0531. Statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

* Dr. Rossmeissl is now with Hay Systems, Inc. in Washington, D.C.
** Mr. Oppler has returned to graduate work at the University of Minnesota.

Four separate criteria were used in evaluating current and alternative composites: (1) predictive validity, (2) fairness to Blacks and females, (3) classification efficiency, and (4) face validity. Each is described briefly before proceeding to a discussion of the results.

Predictive Validity. The correlation of each composite with the CTP score was adjusted for restriction of range due to explicit selection. A multivariate correction due to Lawley (Lord & Novick, 1968, p. 146) was used with each of the ASVAB subtests treated as a separate selection variable. The result was used as the measure of predictive validity. No adjustment was made for "shrinkage" in cross-validation since separate regression coefficients were not estimated. For evaluation of the current composites, this is entirely appropriate. Because we did pick among a large number of alternative composites on the basis of the data at hand, some shrinkage should be expected for the alternatives that appear most extreme. Conventional shrinkage formulas do not handle this situation, so our best approach is to be somewhat conservative in adopting new alternatives to the existing composites.

Fairness to Blacks and Females. Separate regression equations were computed by race and gender where there were at least 50 examinees. Both slope and intercept differences were identified. A single overall measure of the difference in the separate equations was defined in terms of the expected criterion difference for an AA score of 100 (the estimated 1980 norm population mean.) Since selection cutoffs varied between 85 and 110 for the MOS in question, a score of 100 was selected as being in the heart of the critical region for evaluating the selection fairness of alternative composites. Differences in the prediction equations at points significantly below or above this value would have little impact on determination of applicant qualification. The difference in predicted values was converted to a t score by dividing by the standard error of the estimate of the difference (Pothoff, 1964).

Classification Efficiency. The Brogden index, defined as the square root of the average validity times the square root of one minus the average of the intercorrelations among the composites was used as a measure of classification efficiency. This statistic is an indicator of the accuracy of predictions of differences in an individual's expected performance across jobs.

Face Validity. The final evaluation factor was face validity. Face validity is not easily quantifiable, but is more appropriately used as a check of the "reasonableness" of the results. It is our attempt to check purely empirical results against some conception of theory. We would be uncomfortable, for example, with results indicating that AS is an important predictor for clerical jobs, but quite comfortable with AS as an important predictor for vehicle mechanics.

Results

Table 2 shows validities, Brogden indices (Clss. Eff.), and, where appropriate, race and gender t statistics for each contending AA composites. Separate statistics are shown for each applicable MOS and unweighted averages of the validities and t statistics are shown for the cluster as a whole. Each row of statistics corresponds to a different composites. The first row gives statistics for the current composite. Rows with data on alternative composites are labelled A1 through A9. Data are also shown for the CL and SC composites replaced in 1984 after our prior analyses (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984) with

the previous composites labelled PR. Where some other of the current composites has a higher average validity than the operational composite the cluster, data are shown in rows that are labelled according to the other composite. The results presented in Table 2 are discussed separately for each of the current AA composites.

Clerical (CL). The current CL composite has a higher average validity than any alternative. It does, however, underpredict female performance in the two clerical specialties where separate predictions were generated. The addition of either NO or CS significantly reduces the underprediction for females without significantly reducing validity. Adding NO reduces underprediction the most, while adding CS has the greatest face validity and results in slightly greater classification efficiency. A slightly different pattern was found for 76W. The addition of AS increases validity for predicting 76W performance, while decreasing validity for predicting 71L and 76Y performances. Notwithstanding these differences, the current and primary alternative CL composites predict performance in all three clerical MOS quite well.

Combat (CO). The current CO has high validity each of the MOS examined. Some gain in validity would be realized by substituting GS for CS and, perhaps, also swapping MK for AR. The inclusion of GS would improve prediction in all three MOS. The greater contribution of GS also is rational in light of increasing technical sophistication in the systems used in combat specialties. Adding GS would also reduce the small degree of overprediction of the performance of Blacks.

Electronic (EL). The current EL composite does quite well for the one EL specialty examined. Substitution of NO for one or both of the quantitative subtests would increase both predictive validity and classification efficiency, but not to any practical extent.

Field Artillery (FA). Neither the current FA nor any alternative appears to have a very high validity for predicting 13B performance. Consideration of alternative composites is motivated by the fact that several other current composites have higher validities for predicting 13B performance than the current FA composite. Substitution of NO and AS for CS and MK would yield the most significant gains. Such substitution also significantly reduces overprediction for Blacks.

General Maintenance (GM). Very high validities were found for the current GM composite for both 51B and 55B. Very slight gains might result from substituting VE for EI or from simply dropping EI, but these gains would be offset by small increases in overprediction of Blacks' performance and slightly lower classification efficiency estimates.

Mechanical Maintenance (MM). High validities were found for the current MM composite in predicting both 63B and 67N performance. Small gains in the prediction of 63B performance and increased classification efficiency would result from dropping the NO subtest.

Operators/Food (OF). The OF results closely parallel the CL results. Female performance is significantly underpredicted for 94B. Another specialty, 64C, shows a somewhat different pattern of validities, with AS again (and not surprisingly) adding significantly to the predictive validity of this one specialty. In fact, the same composites appear optimal for both the CL and OF MOS -- AR+VE+MK+NO for 16S and 94B (as for 71L and 76Y) and AR+VE+MK+AS for 64C (as for 76W). Substituting AR and MK for AS and MC would significantly reduce underprediction of

female performance for 94B while increasing overall validity.

Surveillance and Communication (SC). A high predictive validity was found for the current SC composite. Some gain in validity, along with a slight increase in classification efficiency, would result if MC were replaced by NO. This would lead to a small increase in the underprediction of performance for Blacks. If MK were also substituted for AR, the same gains in validity and classification efficiency could be obtained along with a decrease in underprediction of Blacks' performance.

Skilled Technical (ST). The current ST is a true Army composite -- it is all that it can be. It has a higher average validity than any possible alternative, and it shows no significant differences in the prediction of performance for Blacks and females.

Summary

The Army's existing AA composites were found to have very high validity for predicting job-specific performance as assessed with the Project A measures. A few changes to the existing AA composites to improve validity or reduce gender differences were identified for further consideration. Specific recommendations are:

CL: Add NO to reduce gender differences.
CO: Replace GS with CS to increase validity/reduce race differences.
FA: Replace CS and MK with NO and AS to increase validity.
MM: Drop NO to increase validity.
OF: Replace NO and MC with AR and MK to increase validity.
Reassign 94B (and similar MOS) to CL to reduce gender differences.
SC: Replace AR and MC with MK and NO to increase validitiy and reduce race differences.

Recommendations for further analyses include: (1) investigation of criterion factors associated with low ASVAB correlations for the 13B measures and significant gender differences for 71L and 94B and (2) evaluation of alternative assignment of MOS to composites, particularly for the CL and OF composites.

References

Lord, F. & Novick, M. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McLaughlin, D., Rossmeissl, P., Wise, L., Brandt, D., & Wang, M. (1984). Validation of current and alternative ASAB Area Composites, based on training and SQT information on FY 1981 and FY 1982 Enlisted Accessions (Technical Report 651). Alexandria. VA: U. S. Army Research Institute.

Potthoff, R. (1964). On the Johnson-Newman technique and some extensions thereof. Psychometrika, 29, 241-245.

Wise, L., Campbell, J., McHenry, J., & Hanser, L. (1986, August). A latent structure model of job performance factors. Paper presented at the annual meeting of the American Psychological Association.

Table 1. Descriptive Statistics

MOS	COMP	GROUP	N	MEAN										STANDARD DEVIATION									
				CTP	GS	AR	VE	CS	NO	AS	MK	MC	EI	CTP	GS	AR	VE	CS	NO	AS	MK	MC	EI
11B: INFANTRY	CO	ALL	491	514	529	539	519	525	515	557	515	551	533	80	80	73	63	64	65	76	78	75	73
12B: COMBAT ENG	CO	ALL	544	509	506	527	496	510	499	555	502	539	522	96	86	70	71	66	60	81	77	83	76
		BLACK	108	453	433	482	440	495	489	479	460	478	473	78	66	47	58	65	54	62	47	58	57
		WHITE	385	529	533	542	519	514	501	584	515	559	539	94	77	70	62	66	61	65	80	81	73
13B: CANNON CREW	FA	ALL	464	510	487	519	488	516	497	514	495	509	502	85	87	69	70	61	65	91	66	83	78
		BLACK	168	485	438	491	456	516	493	458	478	466	467	84	76	60	68	59	68	71	55	70	67
		WHITE	250	528	528	544	518	518	501	563	507	546	533	82	73	65	57	61	63	74	71	74	69
16S: MANPAD CREW	OF	ALL	338	516	509	519	505	527	498	548	495	531	527	94	81	79	66	64	76	81	77	84	76
		BLACK	89	494	449	469	460	540	489	481	464	477	484	78	77	57	52	60	75	67	55	67	60
		WHITE	232	524	534	541	524	522	500	578	510	553	546	99	71	77	62	65	76	69	81	79	74
19E: ARMOR CREW	CO	ALL	394	514	527	536	513	515	506	567	515	549	535	75	84	73	69	66	67	79	77	78	80
		BLACK	71	469	459	497	465	499	483	497	477	488	474	69	77	56	64	65	67	66	63	63	63
		WHITE	297	524	548	547	530	517	511	588	525	568	553	75	75	74	60	65	66	70	76	73	76
27E: TOW/DRG REP	EL	ALL	123	505	540	552	524	518	504	561	532	548	560	101	66	62	58	69	68	75	69	72	70
31C: RADIO/TTY	SC	ALL	289	508	518	540	521	554	547	547	521	527	514	85	76	72	59	54	60	80	79	86	79
		BLACK	74	488	461	494	494	564	557	498	493	479	482	69	68	71	60	44	66	70	63	77	64
		WHITE	204	513	538	555	532	550	542	565	529	543	525	89	68	66	56	56	56	77	82	84	82
51B: CRPNT/MSNRY	CM	ALL	69	513	508	510	497	505	481	555	491	536	533	101	72	72	60	70	66	70	67	76	64
54E: NPC SPEC	ST	ALL	340	507	540	543	529	517	503	543	533	543	531	99	71	73	57	70	69	82	74	72	76
		BLACK	84	466	505	516	515	508	482	485	516	500	493	98	66	69	55	70	72	63	64	55	68
		WHITE	223	522	558	554	541	520	511	571	538	562	549	95	64	74	52	71	67	74	76	70	71
55B: AMMO SPEC	CM	ALL	203	507	497	495	475	491	476	526	481	490	523	97	64	65	62	64	68	69	60	76	57
		BLACK	75	472	477	469	458	492	475	486	470	451	516	99	48	56	53	61	69	52	43	56	44
		WHITE	112	531	513	513	486	491	475	556	486	519	527	89	69	63	65	68	70	66	70	78	64
63B: VEHCLE MECH	MM	ALL	478	513	506	528	496	520	509	579	501	543	536	76	78	71	62	63	59	78	69	79	65
		BLACK	78	464	445	478	456	520	491	510	476	479	503	72	64	59	64	61	63	70	54	57	52
		WHITE	374	526	522	541	507	519	513	598	508	559	546	70	72	69	57	63	59	69	71	75	66
64C: MOTOR TRANS	OF	ALL	507	510	486	498	481	513	499	548	483	522	509	72	75	76	63	65	67	75	68	76	71
		BLACK	121	487	444	456	450	523	492	498	456	471	471	73	65	60	54	61	69	70	54	65	74
		WHITE	358	520	502	513	493	508	501	568	493	541	523	66	71	77	62	66	65	68	69	72	66
		FEMALE	52	495	485	503	520	554	559	464	490	480	454	71	73	78	55	65	67	65	61	72	54
		MALE	455	512	486	498	477	509	492	558	483	526	515	72	75	76	63	63	64	70	69	75	70
67N: HELCPTR REP	MM	ALL	238	510	567	567	546	550	531	613	550	601	582	93	60	59	47	53	63	54	67	54	57
71L: ADMIN CLERK	CL	ALL	427	506	493	528	514	562	552	476	515	484	481	87	82	72	59	49	61	79	75	79	69
		BLACK	159	491	464	499	495	563	535	444	498	454	464	81	74	65	59	45	63	61	69	71	55
		WHITE	235	516	518	548	531	560	560	502	528	505	494	89	79	70	51	52	58	84	75	79	74
		FEMALE	237	524	486	519	522	566	561	447	508	461	465	72	73	67	49	50	63	64	66	68	52
		MALE	190	483	502	539	505	558	540	514	524	512	501	98	91	76	68	48	57	82	84	83	82
76W: PETRO SPPLY	CL	ALL	339	519	479	511	494	536	512	508	491	500	498	95	90	74	69	54	65	99	72	91	81
		BLACK	139	476	430	472	463	539	500	447	461	444	461	88	73	63	65	52	64	73	60	66	67
		WHITE	174	551	521	539	522	535	518	560	514	548	530	88	85	69	60	55	64	90	73	84	78
76Y: UNIT SUPPLY	CL	ALL	444	516	489	518	500	550	531	496	507	496	496	93	85	74	67	51	58	86	75	84	78
		BLACK	169	487	442	479	473	553	518	455	473	453	463	90	69	62	60	46	54	71	60	65	63
		WHITE	231	536	528	547	524	547	538	532	530	530	524	93	76	71	60	56	61	83	78	83	78
		FEMALE	75	519	463	501	492	569	551	429	494	448	453	84	73	62	59	48	61	57	71	72	62
		MALE	369	516	494	522	501	546	527	510	509	506	504	95	87	76	68	51	57	84	76	83	78
91A: MEDIC SPEC	ST	ALL	392	514	547	544	540	525	520	528	530	543	524	79	62	64	46	69	70	82	71	70	68
		BLACK	91	486	519	512	521	519	508	486	511	495	496	72	50	58	42	74	62	70	65	54	58
		WHITE	260	525	562	555	550	527	524	548	538	560	536	80	61	64	42	68	71	80	72	68	70
		FEMALE	116	513	532	545	542	550	549	465	543	504	475	81	59	59	48	58	65	66	66	64	52
		MALE	276	515	554	544	539	514	508	555	525	559	544	79	63	66	46	71	68	73	72	67	63
94B: FOOD SERVCE	OF	ALL	368	526	496	515	503	533	510	516	495	510	503	90	80	77	63	63	69	82	72	76	75
		BLACK	124	493	449	466	471	534	501	469	463	464	471	77	70	58	56	60	72	63	51	56	66
		WHITE	222	546	524	543	524	532	517	546	515	536	524	94	74	73	58	66	66	79	73	75	74
		FEMALE	78	553	474	499	513	562	546	448	489	467	446	79	80	65	53	57	82	64	64	64	59
		MALE	290	519	502	519	501	526	501	534	497	522	518	92	79	80	65	63	62	77	74	75	72
95B: MIL POLICE	ST	ALL	597	504	562	554	542	530	519	573	537	571	550	74	53	60	42	62	62	68	61	58	62

Table 2. Validity, Cultural Fairness, and Classification Efficiency Indicators for Current and Other ASVAB Composites

Current/Other Composites	Avg. Val	Avg t Race	Avg t Sex	Class Eff*	Val	t by Race	t by Sex	Val	t by Race	t by Sex	Val	t by Race	t by Sex
CL: CLERICAL					71L: ADMIN SPEC			76W: PETRO SPPLY			76Y: UNIT SUPPLY		
CL: AR+VE+MK	.661	-2.2	16.1	.231	.64	.6	20.4	.67	-5.8		.67	-1.4	11.8
PR: VE+NO+CS	.578	-5.7	3.1	.248	.59	-.2	5.6	.55	-12.8		.60	-4.3	.5
A1: AR+VE+NO+MK	.656	-3.1	6.7	.232	.65	.4	10.6	.65	-7.8		.67	-2.0	2.9
A2: AR+VE+CS+MK	.656	-2.2	8.1	.233	.65	1.6	11.4	.65	-7.0		.67	-1.1	4.9
A3: AR+VE+AS+MK	.655	-.5	22.2	.222	.60	1.0	32.2	.70	-2.0		.67	-.4	12.3
CO: COMBAT					11B: INFANTRYMAN			12B: COMBAT ENG			19E: ARMOR CREW		
CO: AR+CS+AS+MC	.617	-3.2	-	.231	.66	-	-	.64	-3.5	-	.55	-3.0	-
A1: GS+AS+MK+MC	.648	-1.9	-	.229	.67	-	-	.67	-2.9	-	.60	-1.0	-
GM: GS+AS+MK+EI	.641	-2.5	-	.230	.67	-	-	.67	-3.5	-	.58	-1.5	-
A2: GS+MK+AS	.643	-2.4	-	.230	.67	-	-	.67	-3.3	-	.59	-1.4	-
EL: ELECTRORONIC					27E: TOW/DRGN REP								
EL: GS+AR+MK+EI	.779	-	-	.231	.78	-	-						
A1: GS+NO+EI	.791	-	-	.235	.79	-	-						
A2: GS+NO+MK+EI	.791	-	-	.232	.79	-	-						
FA: FIELD ARTILLERY					13B: CANNON CREW								
FA: AR+CS+MK+MC	.341	-8.4	-	.231	.34	-	-						
A1: GS+NO+AS+MC	.383	-3.1	-	.227	.38	-	-						
A2: AR+NO+AS+MC	.381	-3.8	-	.227	.38	-	-						
GM: GENERAL MAINTENANCE					51B: CRPNT/MSNRY			55B: AMMO SPEC					
GM: GS+AS+MK+EI	.785	-5.0	-	.231	.81	-	-	.76	-5.0	-			
A1: GS+VE+AS+MK	.798	-6.3	-	.229	.84	-	-	.75	-6.3	-			
A2: GS+AS+MK	.791	-6.4	-	.230	.84	-	-	.74	-6.4	-			
A3: GS+AR+VE+AS	.789	-4.5	-	.228	.82	-	-	.76	-4.5	-			
A4: GS+CS+AS+MK	.789	-10.0	-	.229	.86	-	-	.72	-10.0	-			
MM: MECHANICAL MAINTENANCE					63B: VEHCLE MECH			67N: HELCPTR REP					
MM: NO+AS+MC+EI	.729	-4.7	-	.231	.66	-4.7	-	.80	-	-			
A1: AS+MC+EI	.745	-4.5	-	.240	.69	-4.5	-	.80	-	-			
A2: GS+AS+MC+EI	.742	-4.4	-	.233	.68	-4.4	-	.81	-	-			
A3: AS+MK+MC+EI	.739	-5.6	-	.229	.67	-5.6	-	.81	-	-			
A4: AR+MC+AS+EI	.739	-4.3	-	.230	.67	-4.3	-	.81	-	-			
A5: GS+AS+MC	.738	-3.9	-	.234	.67	-3.9	-	.81	-	-			
A6: AS+MC	.733	-3.5	-	.244	.68	-3.5	-	.79	-	-			
OF: OPERATORS/FOOD					16S: MANPAD CREW			64C: MOTOR TRANS			94B: FOOD SERVCE		
OF: VE+NO+AS+MC	.538	-1.0	8.4	.231	.44	.9	-	.52	-1.4	-4.6	.65	-2.5	21.3
A1: AR+VE+AS+MK	.571	.8	9.0	.228	.51	3.0	-	.53	-.5	-14.1	.68	-.2	32.1
A2: GS+AR+AS+MK	.568	.5	10.7	.228	.50	2.9	-	.54	-.1	-4.8	.67	-1.5	26.2
A3: AR+AS+MK	.567	-.2	12.3	.230	.49	2.1	-	.54	-1.1	-2.3	.66	-1.4	26.9
A4: GS+AR+MK	.561	-1.1	10.0	.232	.52	2.2	-	.49	-3.6	-15.5	.68	-1.9	35.5
A5: GS+AR+VE+MK	.561	-.8	13.3	.231	.52	2.7	-	.48	-3.7	-17.6	.69	-1.5	44.1
A6: AR+VE+MK	.558	-1.4	13.2	.228	.52	2.0	-	.46	-5.2	-19.0	.69	-1.2	45.4
A7: AR+VE+MK+MC	.566	-.4	6.4	.234	.50	1.7	-	.51	-2.4	-24.7	.69	-.6	37.6
A8: AR+VE+NO+MK	.548	-4.8	-1.8	.236	.51	-.1	-	.44	-10.8	-16.5	.70	-3.4	13.0
A9: AR+VE+CS+MK	.546	-3.2	2.3	.236	.51	.2	-	.44	-6.9	-14.7	.70	-2.9	19.3
EL: GS+AR+MK+EI	.558	-.8	9.4	.228	.50	2.1	-	.51	-2.0	-7.1	.66	-2.6	25.8
ST: GS+VE+MK+MC	.557	.6	7.1	.228	.50	-1.4	-	.51	1.9	-16.9	.66	-3.0	31.1
FA: AR+CS+MK+EI	.555	-2.9	6.3	.230	.49	-.7	-	.49	-5.1	-22.6	.69	-3.0	35.3
SC: SURVEILLANCE & COMMUNICATION					31C: RADIO/TTY OP								
SC: AR+VE+AS+MC	.693	1.9	-	.231	.69	1.9	-						
PR: VE+NO+CS+AS	.701	.5	-	.232	.70	.5	-						
A1: AR+VE+NO+AS	.729	2.4	-	.233	.73	2.4	-						
A2: VE+NO+AS+MK	.729	.9	-	.233	.73	.9	-						
A3: AR+VE+NO+EI	.728	1.2	-	.234	.73	1.2	-						
A4: GS+AR+NO+EI	.727	2.0	-	.232	.73	2.0	-						
ST: SKILLED TECHNICAL					54E: NBC SPEC			91A: MEDIC SPEC			95B: MIL POLICE		
ST: GS+VE+MK+MC	.683	-1.5	.1	.231	.69	-1.6	-	.73	-1.3	.1	.63	-	-
A1: GS+CS+AS+MK	.679	-1.1	.5	.231	.67	-1.5	-	.75	-1.5	.5	.62	-	-

DEVELOPMENT OF BEHAVIORAL ASSESSMENT PROTOCOLS FOR VARIED REPEATED-MEASURES TESTING PARADIGMS

R. S. Kennedy[1], N. E. Lane[1], R. L. Wilkes[1], and L. E. Banderet[2]

Essex Corporation[1]
1040 Woodcock Rd., Orlando, FL 32813

U.S. Army Research Institute[2] of Environmental Medicine
Natick, MA 01760-5007

ABSTRACT

Recent developments in test methods, data analysis/reduction capabilities, and computer and display technologies raise serious issues and concerns about the appropriateness of designing and assembling a single, general-purpose test battery for a variety of testing requirements. These trends suggest that modular assessment protocols for specific testing paradigms, selected from a larger menu of proven tests, offer several advantages. This paper will describe methods we have developed for selecting efficient and appropriate testing instruments for a variety of testing requirements. Selection of tests for a particular testing requirement is guided predominately by the demonstrated psychometric properties of each test in a repeated-measures paradigm, (e.g., constructs or factors evaluated, stability of means and standard deviations, amount of practice required to achieve reliability and stability, and intertrial correlations).

Many types of unusual and often dangerous stressors are encountered in military, space, and hazardous civilian work settings. Although the effects of these agents are frequently issues of speculation, the extent of actual performance degradation has largely remained unquantified. Human performance testing has been recommended as a potentially valuable tool for the accurate assessment of the various environmental agents on performance (Hannien, 1979; Kennedy & Bittner, 1977; Baker, Letz, & Fidler, 1985; Thorne, Genser, Sing, & Hegge, 1983; and Foree, Eckerman, & Elliot, 1984). The Army (Thorne, Genser, Sing, & Hegge, 1983; Banderet & Burse, 1984), Navy (Kennedy & Bittner, 1977), Air Force (O'Donnell, 1981; Reid, Shingledecker, Nygren, & Eggemeier, 1981; Payne, 1982), and private sector (Foree, Eckerman, & Elliot, 1984) have responded by initiating developmental programs. These human performance testing systems are usually designed for use in atypical work conditions with limited numbers of critical personnel. The demands associated with these environments necessitate the use of

Paper presented at the 28th Annual Military Testing Association Conference, Mystic, CT, November 3-7, 1986.

repeated-measures employing the subject as his own control. Furthermore, measurement must occur quickly and conveniently. The general need for adequate evaluation of assessment tools has been extensively discussed in the literature (Thorndike & Hagen, 1977; Cronbach & Snow, 1977). Jones (1980), places even greater emphasis on the importance of evaluation when the assessment tools are to be applied in highly unusual or exotic research settings. Our researchers (Kennedy & Bittner, 1977) have noted that performance test batteries are often assembled largely for practical reasons by persons whose major interest is not performance testing, and others (Wilkes, Kennedy, Dunlap, & Lane, 1986) have indicated that lack of attention to test metric properties may be the single most important barrier to adequate performance assessment.

We use an engineering approach to performance test selection that was established through the Performance Evaluation Tests for Environmental Research (PETER) program (Kennedy & Bittner, 1977). The PETER approach requires that comparable forms of a task be administered through a series of 10 to 15 trials over a period of successive days. Test performance scores are then subjected to rigorous analyses to surface metric characteristics. Over 150 performance tests were examined with the PETER model and the critical nature of the evaluation is underscroed by the finding that 80% of the evaluated tasks did not meet minimum standards (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984). Excellent reviews of the essential metric characteristics, selection criteria, and evaluation methodologies may be found in the literature (Jones, 1980; Kennedy & Bittner, 1977; Bittner, Carter, Kennedy, Harbeson, & Krause, 1984). These evaluation criteria are briefly summarized below.

1. Stability. Jones, Kennedy, and Bittner (1981) make the point that when repeatedly tested, most subjects demonstrate improvement with practice. An obvious consequence of such a pattern is that the obtained point measures for a subject may differ significantly over time. A second consequence of particular concern is the fact that different subjects may respond differently rather than uniformly to repeated exposures of the task. Therefore, the relative standings of subjects on the first measures may not resemble the relative standings on the final measure. Only after relative standings are clearly and consistently established between subjects (i.e., asymptotic performance with parallel curves for subjects) can the investigator place confidence in the adequacy of his measures. Generally, a test is defined as stable when: (a) the group means for successive trials become constant (i.e., are level, asymptotic, or exhibit constant slope); (b) the between-subject variances for successive trials become constant (i.e., homogeneity of variance); (c) the correlation between a trial and subsequent trials becomes constant. This latter criterion of stability has been labeled "differential stability" by Jones (1969, 1972). If a task has not been stabilized, the correlations among successive trials will very likely show "superdiagonal form" (Jones, 1969). That is, the correlations are greatest between two immediately adjacent trials, with greater separation between trials resulting in progressively smaller correlations. Examination of an intertrial matrix of an unstabilized task makes the pattern readily apparent. Correlations within rows decrease from left to right and correlations within columns decrease from bottom to top.

Therefore, the smallest intertrial correlation would be found in the upper right-hand corner of the matrix. When these correlations cease to change within a row and column and subsequent rows and columns of the matrix, differential stability has been achieved. Theoretically, correlations among stabilized trials are equal.

2. Stabilization Time. Good performance measures should quickly stabilize. Stabilization time must be determined for the group means, standard deviations, and intertrial correlations (differential stability).

3. Task Definition. Once differential stability has been achieved, the average reliability of the task must be determined. Task Definition is obtained by averaging stable intertrial correlations. The minimum acceptable task definition has been operationally defined as $\underline{r} \geq 0.707$.

4. Reliability Efficiency. The Reliability Efficiency of a test is the Task Definition, corrected with the Spearman-Brown prophecy formula, to a 3-minute administration base. Reliability comparisons between tests can only be made on the basis of this standardized metric criteria.

5. Task Ceiling. If all subjects asymptote at the maximum level of performance, then the task is said to have a ceiling (Jones, 1980). Ceilings are undesirable since differences between subjects become impossible to discriminate and/or overlearning could make performance unresponsive to environmental agents. Ceilings are represented by decreasing group standard deviations over trials and by between-trial correlations that fall to zero.

Concerns regarding the use of innovative testing methods without prior evaluation have also been voiced (Smith, Krause, Kennedy, Bittner, & Harbeson, 1983). The advantages of microbased testing in human performance testing have been clearly identified (Wilkes, Kennedy, Dunlap, & Lane, 1986) and use of microprocessors for test administration and data collection is commonplace. Although it has been demonstrated that tests found to be metrically sound in the paper-and-pencil mode may not retain their metric characteristics in the microbased mode (Smith, Krause, Kennedy, Bittner, & Harbeson, 1983), few researchers expend the time and effort necessary for critical comparisons. When microbased testing is employed, we believe comparative examination of paper-and-pencil and microbased versions of a test should be a standard part of the performance test evaluation process (Kennedy, Wilkes, & Kuntz, 1986). That is, the criteria identified above must be established and compared for both modes of testing.

EXAMPLE STUDY

METHOD

The following study has been provided as an example of the engineering analysis procedures advertised above. The example was selected for presentation due to its simplicity and served as a pilot study for NASA-sponsored research. Subsequent research employing larger numbers of subjects, tests, and trials confirmed the findings of the abbreviated

evaluation. In appropriate instances, data from the more comprehensive study have been substituted to clarify evaluation criteria.

Subjects. Twenty subjects from a summer school introductory psychology class were recruited for participation. Men and women volunteers ranged in age from 18 to 47, were in good physical and mental health, and varied from freshman to senior standing.

Materials. Five paper-and-pencil and six comparable performance tasks were selected for evaluation (Table 1). An extensive review of each task may be found in Wilkes, Kennedy, Dunlap, and Lane (1986). Two of the paper-and-pencil tasks (Aiming and Spoke Control) were not directly adaptable to the microbased testing mode and tapping tasks using key-press operations were substituted. All tasks were "speed" type and different but comparable forms were employed on each trial.

TABLE 1. SUMMARY OF PAPER-AND-PENCIL AND COMPARABLE MICROBASED TEST METRIC EVALUATION CRITERIA

	Metric Evaluation Criteria[1]				
Test	Trial Mean Stabilizes	Trial Std. Deviation Stabilizes	Trial Task Definition Established	Task Defin.	Reliability Efficiency
Paper-and-Pencil					
Aiming	5	3	3	0.86	0.99
Spoke Control	5	3	3	0.89	0.96
Pattern Comparison	4	3	4	0.90	0.96
Grammatical Reason.	6	4	3	0.83	0.88
Code Substitution	4	3	4	0.71	0.79
Microbased					
Tapping[2]	3	2	2	0.84	0.99
Pattern Comparison	3	2	5	0.83	0.88
Grammatical Reason.	3	3	5	0.83	0.89
Code Substitution	4	3	3	0.65	0.74

1 Data reported were taken from research performed under NASA Contract No. NAS9-17326.

2 Tapping data have been averaged across the three tapping tests.

Apparatus. Microbased testing was accomplished with a NEC PC8201A microprocessor. More detailed information may be found in Wilkes, Kennedy, Dunlap, and Lane (1986).

Procedure. Subjects were examined over two consecutive days with a modified PETER approach. On each day a subject was first tested with paper-and-pencil tests followed by the microbased versions. A short rest break was provided and then the procedure was repeated. Overall, four trials were obtained for each subject with all tests in both testing modes.

RESULTS

The group means, standard deviations, and intertrial correlations were determined for each test in both testing modes over the four trials. Examination of the evaluation criteria established through these analyses (Table 1) indicates that all tests (in both forms) conformed with the specified criteria, although the microbased versions appear to stabilize more quickly than their paper-and-pencil counterparts. In general, paper-and-pencil test group means stabilized by trial 4 to 5, standard deviations by trial 3, and differential stability was established by trial 3 to 4. In the microbased mode, group means stabilized by trial 3, standard deviations by trial 2 to 3, and differential stability was established by trial 4. Comparison of the Task Definition and Reliability Efficiencies associated with the two modes of testing indicates that, in most cases, the paper-and-pencil testing mode fared slightly better than the microbased. However, the differences are very slight and the evaluation criteria for the microbased tests are far above minimum requirements.

DISCUSSION

All the tests in both modes gave strong indications of meeting established selection criteria. Furthermore, cross-modal comparisons indicate that adaptation from paper-and-pencil to microbased testing does not radically alter metric characteristics for the evaluated tests. Based on these findings, the tests have been recommended for future use in human performance testing.

Researchers are encouraged to establish and examine the recommended selection criteria for all measures prior to conducting data collection. The procedures and methods discussed in this paper, and illustrated in the example, represent what we consider the minimum necessary evaluation criteria for the selection of effective performance measures. Our future research shall be aimed at identifying additional metrically sound performance measures. Other research shall concentrate on establishing the sensitivity of each task to various environmental agents. Application of these methods not only promotes sound research tools, but also provides for increased testing flexibility. Evaluated tests may be added to a growing menu of similarly researched measures. Eventually, this base of information can be used by researchers to combine measures in various ways. Different batteries of tests may then be formed for application to specific measurement needs in assessing the effects of environmental aspects

REFERENCES

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1984). Performance Evaluation Tests for Environmental Research (PETER): The good, bad, and ugly. Proceedings of the 28th Annual Meeting of the Human Factors Society, San Antonio, TX, 11-15.

Baker, E. L., Letz, R. E., & Fidler, A. T. (1985). A neurobehavioral evaluation system for occupational and environmental epidemiology: Rationale, methodology, and pilot study results. Journal of Occupational Medicine, 25, 125-130.

Banderet, L. E., & Burse, R. L. (1984). Cognitive performance at 4500 feet simulated altitude. Paper presented at the Meeting of the American Psychological Association, Toronto, CANADA.

Cronbach, L. J. & Snow, R. E. (1977). Aptitudes and instructional methods. New York: Irvington.

Foree, D. D., Eckerman, D. A., Elliot, S. L. (1984). M. T. S.: An adaptable microcomputer-based testing system. Behavioral Research Methods, Instruments, & Computers, 16(2), 223-229.

Hanninen, H. (1979). Psychological test methods: Sensitivity to long term chemical exposure at work. Neurobehavioral Toxicology, 1, 157-161.

Jones, M. B. (1969). Differential processes in acquisition. In E. A. Bilodeau & I. McD. Bilodeau (Eds.), Principles of skill acquisition. New York: Academic Press.

Jones, M. B. (1972). Individual differences. In R. N. Singer (Ed.), The psychomotor domain. Philadelphia, PA: Lea & Febiger, 197-132.

Jones, M. B. (1980). Stabilization and task definition in a performance test battery (Final Report on Contract N00203-79-M-5089). NAMRL Detachment, New Orleans, LA. Also published as NBDL Monograph NBDL-M001. (NTIS No. AD A099987)

Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr. (1981). A video game for performance testing. American Journal of Psychology, 54, 143-152.

Kennedy, R. S., & Bittner, A. C., Jr. (1977). The development of a Navy performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity enhancement: Personnel performance assessment in Navy systems. Navy Personnel Research & Development Center, San Diego, CA. (NTIS No. AD A056047)

Kennedy, R. S., Wilkes, R. L., & Kuntz, L. A. (1986). Sensitivity of a notebook-sized portable automated performance test system. Paper presented at the Annual Behavioral Toxicology Society Meeting, Atlanta, GA.

Payne, D. L. (1982, February). Establishment of an experimental testing and learning laboratory. Paper presented at the 4th International Learning Technology Congress and Exposition of the Society for Applied Learning Technology, Orlando, FL.

Reid, G. B., Shingledecker, C. A., Nygren, T. E., & Eggemeier, F. T. (1981). Development of multidimensional subjective measures of workload. Atlanta, GA: IEEE Systems, Man & Cybernetics Society.

Smith, M. G., Krause, M., Bittner, A. C., Jr., Kennedy, R. S., & Harbeson, M. M. Performance testing with microprocessors: Mechanization is not implementation. Proceedings of the 27th Annual Meeting of the Human Factors Society, Norfolk, VA, 674--678.

Thorndike, R. L., & Hagen, E. P. (1977). Measurement and evaluation in psychology and education (4th ed.). New York: Wiley.

Thorne, D., Genser, S., Sing, H., & Hegge, F. (1983). Plumbing human performance limits during 72 hours of high task load. The Human as a Limiting Element in Military Systems, DRG Seminar Papers. Toronto, Ontario, CAN: Defense and Civil Institute of Environmental Medicine.

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Lane, N. E. (1986). Stability, reliability, and cross-mode correlations of tests in a recommended 8-minute performance assessment battery (TR No. EOTR 86-4 for NASA Contract No. NAS9-17326). Essex Corporation, Orlando, FL.

The Underestimation of Treatment Effects: Possible Causes

Harris R. Lieberman, Ph.D.

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Although efforts have recently been made to formally describe key parameters of performance tests, many critical characteristics of such tasks, and of the experimental design in which the tasks are embedded, remain controversial. This state of affairs produces a variety of unfortunate consequences. For example, because measurement of even simple behavioral tasks varies considerably from one laboratory to another, results that appear to be contradictory are in actuality the consequences not of substantive differences, but rather differences in methodology. Such problems are often magnified by the low statistical power of many studies and uncertainty regarding the underlying parameters actually assessed by apparently similar tasks. For these and other reasons it seems possible that the existence and magnitude of the effects of various experimental manipulations on task performance may have been underestimated. Some factors that could contribute to errors of this nature will be discussed in the context of research on exposure to high altitude and caffeine.

The premise of this paper, that the effects of treatment variables are frequently underestimated, is a difficult hypothesis to prove or disprove because the absence of an effect can never be proven. Although the question may therefore be moot from a logical perspective, from a purely practical viewpoint the underestimation of effects could result in considerable adverse consequences. In the case of any specific treatment, whether this type of error is important is a complex problem and will depend on a number of factors. Although, on purely scientific grounds, one would always wish to detect any effect present, this paper will view the problem from a more practical perspective, i.e. how such information could influence the policy decisions of an end user. Appropriate examples of users of performance data are military commanders and planners, health administrators, product development engineers and even astronomers. For example Cudaback (1984), from the astronomy department of a major university, recently found it necessary to review the psychological literature on exposure to hypoxia due to concerns related to the design of high-altitude observatories. Since many observatories are located above 4000 M, the question of human performance in hypobaric environments is an issue of considerable relevance to his profession. Cudaback (1984) discusses the effects of such exposure on human performance and health and concludes that the problem is larger than previously recognized and that the use of supplemental oxygen in working spaces at or above 4000 M is justified. Currently supplemental oxygen is not typically used in such observatories. It is apparent from his review that the issue of the threshold for serious acute impairment in performance at high altitude is controversial and has not been thoroughly investigated. Some

investigators have suggested little impairment is present until one reaches the 4000 M level or higher and others believe, based largely on anecdotal observations, that symptoms are present at altitudes as low as 2000 M. In aviation, civil pilots are required to use oxygen when spending more than 30 minutes above 3800 M but naval pilots flying in low performance aircraft during daylight must use it when operating above 3000 M. These discrepancies regarding operations at moderate altitudes are clearly the result of a lack of parametric studies at these altitudes. Such studies probably have not been conducted because it was assumed that deficits under these conditions would be impossible to document. At approximately 3700 M effects on mental performance are detected after 12 h on some but not on all tests administered (White, 1984). These results are further complicated by the apparent rapid recovery of performance that occurs after a day or even less of altitude exposure (Banderet, 1984). The consequences of underestimation of acute performance decrements could be substantial in aviation or as Cudaback believes, in astronomical observatories. Future studies at altitudes between 2000 and 4000 M are clearly called for.

Caffeine

An even more controversial area than the threshold for hypoxic-induced decrements in performance is the psychopharmacology of caffeine. Caffeine, a common food constituent and also a food additive, is generally considered by the public to be a stimulant that improves performance and increases alertness. However, reviews of the scientific literature on caffeine usually fail to reach any definitive conclusion concerning the behavioral effects of this substance. In spite of numerous studies on the effects of caffeine given in relatively high doses, there is little agreement on its effects on mood and performance at doses comparable to those found in single servings of ordinary foods. Foods that contain caffeine include coffee, tea, cola-beverages, and chocolate. The caffeine concentration of a cup of coffee can vary from approximately 60 to 112 mg depending on the type of coffee beans used and the method of preparation. A cup of tea and a typical 12 oz. serving of cola each contain about 40 mg (Roberts & Barone, 1983). Unfortunately, most investigations of caffeine's behavioral effects have used doses that are well above the amounts found in these foods. However, even at higher doses, the presence of behavioral effects of caffeine are difficult to consistently document because of the contradictory nature of many studies. The reasons for the lack of consistency across studies are numerous and some of them will be discussed below.

Caffeine's Effects on Performance

The classic work of Hollingworth (1912), who noted that caffeine enhances performance in some situations, was probably the first systematic evaluation of the behavioral effects of a food constituent on performance. Numerous additional studies performed since that time have failed to produce a consistent picture. The most thorough reviews of the literature conclude that performance effects are highly variable, with both improvement and impairment being reported (Sawyer et al, 1982; Dews, 1984). Another paper concludes that caffeine's behavioral effects are difficult to detect and "capricious" (Dews

et al, 1984). However, a great deal of this between- and even within-laboratory variability can be accounted for by methodological differences between studies and the failure of many investigators to take into account important confounding factors such as prior history of caffeine use or smoking (which substantially decreases the plasma half-life of caffeine). Errors in logic and misunderstanding in the meaning of statistical tests have also greatly confused the issues. In general, these kinds of disagreements, misunderstandings and errors in data interpretation have resulted in, we believe, substantial underestimation of the effects of caffeine on performance and mood state.

Task Selection and Comparison

As has been previously noted (Lieberman et al, 1986) selection of specific tests to be used in behavioral studies may be one of the most critical aspects of study design. Even if all other aspects of a study are conducted flawlessly, selection of inappropriate or insensitive tasks will result in a failure to detect significant treatment effects that are present. Sometimes selection of appropriate tests is primarily a matter of determining what behavioral functions are altered by the treatment in question. It is often not apparent in advance what parameter (i.e. sensation, reaction time (RT), memory and vigilance) will be affected by a particular treatment. Furthermore, tests that supposedly measure one cognitive function typically have a multifactorial substrate. For example, a "memory" test will also involve sensory processing, cognitive functions not related to memory and motor output. In other instances, what appear to be relatively minor differences between similar tests of the same behavioral parameter may be critical to detection of an effect.

The literature on caffeine's effects on performance provides numerous examples of the critical nature of test selection . Effects of caffeine on purely sensory tasks, like critical flicker fusion, are rarely observed (File et al, 1982). However, many investigators do report effects of moderate doses of caffeine on tests with substantial vigilance components (Regina et al, 1974; Clubley, 1979; Lieberman et al, 1986) while others fail to detect effects of even high doses of caffeine on vigilance tasks (Loke and Meliska, 1984). There appear to be a number of characteristics of vigilance tests that increase the probability of detecting effects of caffeine. One appears to be the duration of the task. The Continuous Performance Task (CPT), an adaptive test of visual vigilance, is a comparatively brief test and appears to be less sensitive to the effects of low and moderate doses of caffeine than the much longer duration Wilkinson auditory vigilance test (Clubley 1979, Lieberman et al, 1986). Long duration tests of simulated driving that have a substantial visual vigilance component have also detected effects of caffeine at moderate and high doses (Baker and Theologus, 1972; Regina et al, 1974). However, duration does not appear to be the only critical parameter that distinguishes vigilance-type tests that detect effects of caffeine from those that do not. A study conducted by Loke and Meliska (1984) failed to observe any effects of moderate and high doses of caffeine (195 or 325 mg.) even though their visual vigilance task was 90 minutes in duration. In their task a relatively large number of signal trials (22%) were employed, as opposed to only 2% in the

Wilkinson vigilance task (Wilkinson, 1970). This may have decreased the monotony of the task, a critical characteristic of vigilance tasks, which are purposely designed to produce boredom and fatigue (Holland, 1968). The fact that their subjects were detecting about 90% of the test stimuli as opposed to performance of about 50% correct on the Wilkinson vigilance task supports this interpretation and also suggests that a ceiling effect may have been present in the Loke and Meliska (1984) study.

In our laboratory we have consistently seen effects of low and moderate doses of caffeine on a modified version of the Wilkinson vigilance test (Lieberman et al, 1986; Lieberman et al, in press). In two separate crossover studies, each with 20 subjects participating, significant increases in vigilance were noted after every dose of caffeine administered compared to placebo. In one study doses of 32, 64, 128, and 256 mg were administered. In a second, doses of 64 and 128 mg were given. Other performance tests occasionally detected positive effects of caffeine (four choice visual RT and simple auditory RT) but the magnitude and consistency of the effects were not as great as those seen with the Wilkinson vigilance test (Lieberman et al 1986; Lieberman et al, in press).

Statistical Considerations

Another critical issue related to the possible underestimation of treatment effects is the statistical power of particular studies. The statistical power of a study is the probability of detecting an effect if one actually exists, i.e. rejecting the null hypothesis when it is false. It is often assumed that negative results (failure to detect treatment effects) are as likely to be "correct" as positive results. That is, if two identical studies are conducted and the result of one is statistically significant, and the other is not, then nothing has been learned. This is not the case. Nearly all investigators use the probability of correctly rejecting the null hypothesis as the basis for their statistical conclusions. This probability is usually set arbitrarily at $p<.05$ and is the probability of a Type I error. This is an error of claiming an effect is present when none exists. A Type II error is an error of denying the existence of an effect that is really present. The probability of a Type II error is not often formally considered by most investigators and is usually greater than 0.05. When it is taken into account, for example when the desired sample size is computed prior to initiation of a study, it is often set at 0.20. This is considered to be an appropriate level and means that one will detect an effect that is actually present only 80% of the time. Statistical power increases as the probability of a Type II error decreases (power = 1 - the probability of a Type II error). Since the probability of a Type II error is usually much greater than the probability of a Type I error, in many cases a negative study may actually be less likely to be "correct" than an equivalent positive study.

One of the key determinants of the probability of making a Type II error is sample size - as it increases the probability of this type of error declines. Negative studies with small sample sizes therefore convey little information. Conversely, and somewhat counterintuitively, positive studies with small sample sizes actually indicate that a potent effect is present, assuming the alpha level is maintained at $p<.05$. It should also be noted that

the results of a study are usually considered to be negative if an effect is present but the alpha level is only a fraction greater than 0.10. Many "negative" studies are therefore inconclusive for this reason, and do not convey information that warrants acceptance of the null hypothesis because the p values obtained approach significance. The magnitude of the effect in question is also of great importance in such instances. Of course, these statistical considerations have been oversimplified and a number of other issues are also relevant to this discussion. The interested reader should refer to texts such as Rosenthal and Rosnow (1985) for additional details. Unfortunately many literature reviewers tend to equally weigh positive and negative results and therefore reach equivocal conclusions. A more mathematical approach to the integration of results of different studies, termed meta analysis, has recently been developed (Rosenthal, 1980).

The consequences of underestimating the effects of caffeine on performance may be substantial. Currently many individuals believe that this substance is harmful and avoid it. However, if it does improve performance it may have considerable benefits at certain times for certain individuals. For example, applied studies have indicated that driving ability in general, and also maintenance of nighttime vigilance while operating a motor vehicle, is enhanced by moderate doses of caffeine (Baker & Theologus 1972; Regina et al, 1974). One can also conceive of a number of military tasks such as guard duty, radar or sonar monitoring, sustained flight operations and various intelligence gathering activities where increased vigilance would be quite helpful. Of course such potential benefits must be weighted against potential health risks that have been attributed to chronic high-dose consumption of caffeine (Jick et al 1973; Dobmyer et al, 1983).

Conclusion

Two treatments, hypoxia and caffeine, whose effects on performance may have been underestimated are discussed. It is apparent that numerous methodological inadequacies have contributed to confusion and inconclusiveness in the literature regarding their effects on performance. Examples of some other areas where similar problems could result in the underestimation of treatment effects include neurotoxicology, chronobiology, and nutrition. The underestimation of treatment effects in these areas and others could potentially have serious consequences.

References

Baker, W. J. & Theologus, G. C. (1972). Effects of caffeine on visual monitoring. Journal of Applied Psychology, 56, 519-523.

Banderet, L. E. & Burse, R. L. (1984, August). Cognitive performance at 4500 meters simulated altitude. Paper presented at the meeting of the American Psychological Association, Toronto, Canada.

Bittner, A. C., Carter, R. C., Krause, M., Kennedy, R. S., & Harbeson. M. M. (1983) Performance evaluation tests for environmental research (PETER): Moran and computer batteries. Aviation, Space, and Environmental Medicine, 54(10), 923-928.

Clubley, M., Bye, C. E., Henson, T. A., Peck, A. W. & Riddington, C. J. (1979). Effects of caffeine and cyclizine alone and in combination on human performance, subjective effects and EEG activity. British Journal of Clinical Pharmacology, 7, 157-163.

Cudaback, D. (1984). Four-KM altitude effects on performance and health. Publication of the Astronomical Society of the Pacific, 96, 463-77.

Dews, P. B. (1984). Behavioral effects of caffeine. History and use. In P. B. Dews (Ed.), Caffeine (pp. 86-103). Berlin: Springer-Verlag.

Dews, P. B., Grice, H. C., Neims, A., Wilson, J. & Wurtman, R. J. (1984). Report of fourth international caffeine workshop, Athens, 1982. Food and Chemical Toxicology, 22(2), 163-169.

Dobmeyer, D. J., Stine, R. A., Leier, C. V., Greenberg, R. & Schaal, S. F. (1983). The arrhythmogenic effects of caffeine in human beings. New England Journal of Medicine, 308, 814-816.

File S. A., Bond, A. J., Lister, R. G. (1982). Interaction between effects of caffeine and lorazepam in performance tests and self-ratings. Journal of Clinical Psychopharmacology, 2, 102-106.

Holland, J. G. (1968). Human vigilance. Science, 128, 61-67.

Hollingworth, H. L. (1912). The influence of caffeine on mental and motor efficiency. Archives of Psychology (NY), 3, 1-166.

Jick, H., Miettinen, O. S., Neff, R. K., Shapiro S., Heinonen, O. P. & Slone, D. (1973). Coffee and myocardial infarction. New England Journal of Medicine, 289, 63-67.

Lieberman, H. R., Spring, B. J., & Garfield, G. S. (1986). The behavioral effects of food constituents: strategies used in studies of amino acids, protein, carbohydrate and caffeine. Nutrition Reviews, 44(suppl.), 61-70.

Lieberman, H. R., Wurtman, R. J., Garfield, G. S., & Coviella, I. L. G. (in press). The effects of caffeine and aspirin on mood and performance. Clinical Psychopharmacology.

Loke, W. H. & Meliska, C. J. (1984). Effects of caffeine use and ingestion on a protracted visual vigilance task. Psychopharmacology, 84, 54-57.

Regina, E. G., Smith, G. M., Keiper, C. G. & McKelvey, R. K. (1974). Effects of caffeine on alertness in simulated automoble driving. Journal of Applied Psychology, 59, 483-489.

Roberts, H. R. & Barone, J. J. (1983). Biological effects of caffeine. History and use. Food Technology, 37(a), 32-39.

Rosenthal, R. (1980). New directions for methodology of social and behavioral science (no. 5). San Fransisco: Jossey-Bass, Inc.

Rosenthal, R. & Rosnow, R. (1985). Contrast analysis: focused comparisons in the analysis of variance. New York: Cambridge University Press.

Sawyer, D. A., Julia, H. L. & Turin, A. C. (1982). Caffeine and human behavior: arousal, anxiety, and performance effects. Journal of Behavioral Medicine, 5(4), 415-439.

White, A. J. (1984). Cognitive impairment of acute mountain sickness and acetazolamide. Aviation, Space, and Environmental Medicine, 55, 598-603.

Wilkinson, R. T. (1970). Methods for research on sleep deprivation and sleep function. In E. Hartmann (Ed.) Sleep and Dreaming (pp, 369-381). Boston: Little Brown.

THE USE OF SUBJECTIVE MEASURES FOR BASIC PROBLEM-DEFINITION

Ilse Munro, M. A. and MAJ Terry M. Rauch, Ph. D.

US Army Research Institute of Environmental Medicine

Natick, MA 01760-5007

One of the hazards of research is missing an effect because the appropriate measure was not used. This can easily occur in military field tests. It is difficult to predict which parameters will be affected by the complex test conditions. Moreover, it is not always possible to use the most sensitive measures. Methods that encumber the soldier, intrude on the scenario, or have little face validity are often deemed unacceptable. Since overlooking or falsely reporting the absence of an effect could have serious military consequences, steps were taken to develop a simple, sensitive method for obtaining comprehensive data.

A recent series of field tests indicated this method should be based on self-report measures. The tests simulated extended operations in areas contaminated with chemical agents. Regardless of the varied test conditions, soldiers showed inadequate endurance in the full protective ensemble (MOPP 4). Broad-based subjective measures showed that the problem was consistently related to a small number of physical symptoms: difficulty breathing, painful breathing, shortness of breath, headache, and nausea (Munro et al., in press). These effects were not detected by objective means. Respiration and respirator resistance could have been measured, but were not; the focus was on heat stress parameters (e.g., core temperature and heart rate). Headache and nausea could not have been measured objectively; they are essentially subjective phenomena. The subjective measures, therefore, had a dual function. They served as imprecise substitutes for objective measures and provided valuable information in their own right.

However, it was also apparent from these tests that standard self-report measures do not provide all the necessary information. In addition to determining how a soldier feels, it is important to know how he performs. The MOPP 4 tests suggested that the military tasks typically selected for field tests may be better suited for training, their original intent, than for test purposes. Few performance decrements were observed in any of the tests -- despite the arduous conditions -- and none of the observed decrements were seen across tests (Headley, Brecht-Clark, Feng, & Whittenburg, in preparation). Unfortunately, the subjective measures used did not include systematic self-rating of performance.

MILITARY ABILITIES QUESTIONNAIRE

The first step in developing a self-report battery, therefore, was to design a means by which soldiers could assess their own performance problems. To this end, a 90-item questionnaire was constructed. Thirty items were based on general abilities that are used in various combinations in the performance

of military tasks. Nine physical factors (Fleishman, 1964), eleven psychomotor factors (Fleishman, 1962), and ten cognitive factors (Dunnette, 1976) were selected. The remaining 60 items reflected attributes a soldier needs to function within a military organization. Since it was not possible -- or productive -- to cover all the behavioral requirements of military occupations, only those that were thought to make the difference between the success or failure of a mission were included. These were selected from a critical incident analysis performed for the Navy by Borman, Dunnette, & Johnson (1974).

The general abilities and critical behaviors were translated into terms readily understood by soldiers participating in field tests. The wording was as concrete as possible without restricting any item to a particular scenario. Modifications were made to ensure that all items were applicable to Army officers and enlisted personnel.

The inclusive nature of the questionnaire presented the problem of what to do with items that do not apply to a given test situation. From previous experience, it was obvious that neither the test soldiers, who tend to overuse an "N/A" option when it is offered, nor investigators, who do not always know the relevant parameters in advance, should select a subset of items. In order to allow soldiers to respond to all items, two separate five-point scales were provided. Each was accompanied by verbal descriptors of level of difficulty. One scale was to be used if the rating was based on direct observation, the other if the rating was hypothetical. It was assumed that soldiers, knowing how they feel in a given situation, could use their previous experience to make valid estimates when necessary. The two scales could be analyzed separately or combined for repeated measures analysis. Programming the questionnaire on portable computers (GRiD Compass II, GRiD Systems Corporation, Mountain View, CA) eliminated potential confusion of the two scales. Each item appeared on the screen by itself, accompanied by the two mutually-exclusive scales.

The questionnaire was first administered in a field training exercise (FTX) conducted with augmented Special Forces A-teams in a remote, mountainous region of Vermont (Askew et al., in preparation). The FTX was undertaken to test a prototype ration developed for use in covert operations lasting up to 30 days. Because of space and weight considerations, the compact ration (RLW-30) provides only about half the calories of a ration such as the MRE (Meal, Ready-to-Eat), which served as the control. This test offered an ideal opportunity to explore the properties of the new questionnaire. Nutritional effects are particularly subtle and likely to be missed (Lieberman, Spring, & Garfield, 1986). Moreover, the conditions of testing were conducive to negative results. A group design was used, subject assignment to the two rations was not random, the groups engaged in separate operations under a scenario that changed from day to day, and most psychological and physiological testing occured -- at best -- during weekly sessions at a base camp or a laboratory far removed from the test site. If the questionnare were able to detect effects in this test, there would be reason to believe it had general utility.

Table 1 shows a summary of the findings from the Military Abilities Questionnaire administered in the RLW-30 test. Data were collected during five weekly sessions, and items were rated according to experience over the previous week. Difficulty ratings from the two scales were combined. Pre-test data were submitted to one-way analysis of variance (ANOVA), and a separate two-way repeated measures analysis of variance was performed for data from the four sessions conducted over the course of the month-long test. Only items showing effects significant at the 0.05 level are presented.

TABLE 1
MILITARY ABILITIES QUESTIONNAIRE
Summary of Significant ($p \leq 0.05$) Findings in RLW-30 Test

PRE-TEST DIFFERENCES	TEST DIFFERENCES	
Group	Week	Group / Group x Week
Physical Abilities		
- Coordinate body while moving	- Push/pull a heavy object - Lift a heavy object - Maintain balance	- Lift a heavy object
Psychomotor Abilities		
	- Coordinate arms/legs - Respond quickly	- Track an object - Type/use telegraph
Cognitive Abilities		
		- Remember unrelated bits of information - Orient self/object
Critical Behaviors		
- Notice performance problems - Identify critical performance problems - Help others improve their performance - Work under duress without complaining - Avoid making fun of others	- Support the policies of superiors - Show respect for superiors - Follow orders - Present a good image of the Army - Be professional with civilians - Cooperate with civilians - Identify critical performance problems - Help others improve their performance - Improve own performance - Take charge of emergencies - Risk own safety for others - Observe regulations on restricted items	- Find new solutions to a problem - Observe regulations on equipment/personnel use - Complete all parts of a task - Work under duress without complaining - Improve the morale of others - Avoid making fun of others

The findings show that the questionnaire was, in fact, able to detect effects attributable to different aspects of the test. (1) Subject assignment. The team believed to be the best able to handle any adverse effects had been assigned to the prototype ration. Six questionnaire items showed significant pre-test group differences, and five indicated that the RLW-30 group experienced less difficulty under garrison conditions than did the MRE group ("coordinate body while moving" was the exception). (2) General test effects. Isolation under somewhat harsh environmental conditions could be expected to progressively affect all soldiers. Seventeen items showed that soldiers experienced increased difficulty over the test weeks. Peak difficulty typically occurred in the third week, perhaps due to increased physical demands during that time. (3) Treatment effects. It was expected that differences resulting from caloric adequacy would not be manifest in the same manner for the two groups over the four test weeks. Eleven items showed group differences, most of which involved interactions with time in the test. Typically, difficulty ratings decreased over time for the MRE group, while increases were seen for the RLW-30 group through the first three weeks. Only three items (last items, third column, Table 1) showed the RLW-30 group consistently fared better, and these can be explained in terms of pre-test differences. Potential performance decrements in the RLW-30 group were seen in each of the four categories of items.

The results indicated that the questionnaire is not only sensitive but suitable for its intent: filling in gaps left by other measures. The item "lift a heavy object", for example, suggested that group differences in physical abilities may have been most pronounced early in the test. This could not have been detected by physical fitness tests, which were administered only pre- and post-test. In addition, psychomotor and cognitive items showing group differences indicate that abilites not measured by psychological performance tests (tracking, tapping, memory, and orientation) may have been affected by the new ration. Finally, a look at all the critical behaviors listed in Table 1 shows that there are many important performance attributes (e.g., the ability to operate in emergencies or arrive at novel solutions) that are never directly measured by either psychological or military data collectors.

SUBJECTIVE PROBLEM-DEFINITION BATTERY

The Military Abilites Questionnaire was used in the RLW-30 test in conjunction with two standard questionnaires, the Environmental Symptoms Questionnaire (ESQ) (Kobrick & Sampson, 1979) and the Profile of Mood States (POMS) (McNair, Lorr, & Droppleman, 1981). This three-component battery appears to hold promise as a simple method of obtaining comprehensive data under field conditions.

The primary drawback of this battery lies in the interpretation of the collected data. One problem is sheer volume. The three questionnaires together total 222 items. The Military Abilities Questionnaire has no factor structure as yet, and the generality of the ESQ and POMS factors can be questioned. The ESQ factors were derived from studies conducted in cold, hypobaric environments (Sampson, Cymerman, Burse, Maher, & Rock, 1983). When a factor analysis was performed on the combined ESQ data from the MOPP 4 tests described above, a different structure emerged. The POMS may have similar problems, since the factors are based on data from civilian rather than military populations. Most field tests do not use a sufficient number of subjects for valid factor analysis, so the alternative to using the given

factors is intuitively integrating individual significant effects. In addition, the meaning of the effects is not certain. The Military Abilities Questionnaire has not been submitted to validation studies. However, even a questionnaire such as the ESQ provides data that can be variously interpreted. When one group of subjects has significantly higher ratings on items related to breathing distress (as seen in the MOPP 4 tests), does this mean that they are actually having greater problems breathing, that they perceive they are having greater difficulties, or that they are simply more willing to report either real or perceived breathing difficulties?

Despite problems of interpretation, there is a clear need for such a battery. Military field tests are the best means available for approaching the complexity that exists under real operational conditions. The suggested battery can ensure that this complexity is not lost in the pursuit of objectivity and precision. Once the basic effects are defined, greater precision can be attained under laboratory conditions by using objective tests. For many of the questionnaire items, there is a direct connection with an appropriate objective measure. One third of the Military Abilities Questionnaire, for example, is derived from factors that are based on laboratory tests. Even when objective measures do not exist or follow-on studies are not conducted, subjective data can provide valuable information. Perception is important in and of itself. In the MOPP 4 tests, soldiers reporting breathing difficulties removed their respirators and withdrew from the test. Under actual chemical attack, they could have exposed themselves to a toxic agent regardless of the validity of their belief.

REFERENCES

Askew, W. E., Munro, I., Teves, M., Siegel, S., Popper, R., Rose, M., Hoyt, R., Martin, J., Lieberman, H., Shaw, C., Reynolds, K., & Engell, D. (in preparation). Nutritional status and physical and mental performance of soldiers consuming the Ration, Lightweight or the Meal, Ready-to-Eat during a 30 day field training exercise (Technical Report). Natick, MA: US Army Research Institute of Environmental Medicine.

Borman, W. C., Dunnette, M. D., & Johnson, P. D. (1974). Development and evaluation of a behavior-based Naval officer performance assessment package. Minneapolis: Personnel Decisions, Inc.

Dunnette, M. D. Aptitudes, abilities, and skills. (1976). In M. D. Dunnette (Ed.). Handbook of industrial and organizational psychology. Chicago: Rand McNally College Publishing Company.

Fleishman, E. A. (1962). The description and predicition of perceptual-motor skill learning. In R. Glaser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press.

Fleishman, E. A. (1964). The structure and measurement of physical fitness. Englewood Cliffs, NJ: Prentice-Hall.

Headley, D. B., Brecht-Clark, J. M., Feng, T. D., & Whittenburg, J. A. (in preparation). The effects of the chemical defense ensemble and extended

operations on performance and endurance of combined arms crews (Technical Report). Alexandria, VA: US Army Research Institute.

Kobrick, J. L., & Sampson, J. B. (1979). New inventory for the assessment of symptom occurance and severity at high altitude. Aviation, Space and Environmental Medicine, 50, 925-929.

Lieberman, H. R., Spring, B. J., & Garfield, G. (1986). The behavioral effects of food constituents: Strategies used in studies of amino acids, protein, carbohydrate and caffeine. Nutrition Reviews, 44 (Suppl.), 61-70.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1981). EITS manual for the Profile of Mood States (POMS). San Diego: Educational and Industrial Testing Service.

Munro, I., Rauch, T. M., Tharion, W., Banderet, L. E., Lussier, A. R., & Shukitt, B. (in press). Factors limiting endurance of armor, artillery, and infantry units under simulated NBC conditions. Proceedings of the Army Science Conference.

Sampson, J. B., Cymerman, A., Burse, R. L., Maher, J. T., & Rock, P. B. (1983). Procedures for the Measurement of Acute Mountain Sickness. Aviation, Space and Environmental Medicine, 54, 1063-1073.

ADDENDUM

Human subjects participated in these studies after giving their free and informed voluntary consent. Investigators adhered to AR 70-25 and ASAMRDC Regulation 70-25 on Use of Volunteers in Research.

The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Army position, policy, or decision unless so designated by other official documentation.

MOOD STATES AT 1600 AND 4300 METERS HIGH TERRESTRIAL ALTITUDE

Barbara L. Shukitt, B.A. & Louis E. Banderet, Ph.D.

US Army Research Institute of Environmental Medicine
Natick, MA 01760-5007

When unacclimatized individuals are exposed to high terrestrial elevations (above 3000 m) for several hours, they often experience considerable subjective discomfort as well as some functional disability This disorder is referred to as acute mountain sickness (AMS). Two different inventories to assess symptom occurrence and severity of AMS have been used in past studies, the General High Altitude Questionnaire (GHAQ) and the Environmental Symptoms Questionnaire (ESQ) (Evans, 1966; Kobrick & Sampson, 1979; Sampson & Kobrick, 1980; Stamper, Kinsman & Evans, 1970). AMS is characterized by symptoms such as headache, dizziness, loss of appetite, nausea, fatigue, insomnia, irritability, depression, and difficulty with thinking (Carson, Evans, Shields & Hannon, 1969; Houston, 1983). The number, severity, rapidity of onset, and duration of AMS symptoms vary from person to person. Generally, AMS symptoms are most severe during the first or second day at altitude and then gradually recede over the next 2 - 4 days (Carson et al., 1969; Hansen, Harris & Evans, 1967; Ward, 1975).

Unfortunately, no one standardized scale has been utilized to measure mood changes at altitude. Personal anecdotes imply that ascent to altitudes between 2500 - 5500 m produce two predominate reactions - euphoria and depression. Initially, there is a stage of euphoria which is accompanied by a feeling of self-satisfaction and a sense of power. After a while, however, this initial stimulation is followed by depression. With time at altitude, the person also may become quarrelsome, irritable, and apathetic (Van Liere & Stickney, 1963).

The present study was part of a larger investigation (Evans, Robinson, Horstman, Jackson & Weiskopf, 1976) which examined whether a combination of staging, temporary residence for a few days at a moderate altitude before ascent to a higher altitude, plus the administration of acetazolamide would improve AMS symptomatology at altitude. In this investigation Evans et al. (1976) found that almost all symptoms of AMS were prevented at 4300 m by this treatment strategy. Acetazolamide has also been used in previous studies (Cain & [illegible], 1966; Forwand, Landowne, Follansbee & Hansen, 1968) as a pretreatment for AMS.

In this same study, Banderet (1977) assessed mood periodically but only reported altitude results after 19 hours at 1600 and 4300 m. This interval was chosen because AMS symptoms are usually most severe at this time. People in both the control and treatment groups became less friendly and clear thinking and more sleepy and dizzy at 4300 m. No mood changes were found at 1600 m. Therefore, Banderet found the Clyde Mood Scale to be sensitive to high altitude.

Since only one point (19 hours) was reported in this previous study, this effort looked at the time course of these mood states in the earlier data base (Banderet, 1977); i.e. changes in mood over a period of two days at 1600 m and a period of four days at 4300 m. Morning - evening differences in mood were also examined.

METHOD

Subjects: The subjects were 16 female and 19 male fully-informed volunteers, ranging in age from 18 to 28 years, from Fort Sam Houston, Tx (200 m). All were medically screened and were excluded if they were born at an altitude over 1000 m, had resided for more than 1 month at an altitude over 1000 m in the last 3 years, or had sojourned to an altitude over 3000 m within 3 months prior to the study.

Mood questionnaire: The Clyde Mood Scale (Clyde, 1963) was used to assess the subjects' moods. This scale was designed to measure human emotions. It consists of 48 adjectives - e.g. "kind", "dependable", "alert", "lonely", "tired" - rated on a four-point scale, i.e. "not at all", "a little", "quite a bit", and "extremely". Prior statistical analysis has shown that the 48 adjectives cluster into 6 principal mood factors - friendliness, aggressiveness, clear-thinking, sleepiness, unhappiness, and dizziness. Its sensitivity to high altitude has been shown previously (Banderet, 1977; Banderet, Shukitt, Kennedy, Houston & Bittner, in press).

Procedures: The present study was part of a larger investigation (Evans et al., 1976) in which subjects were randomly assigned to the control (n = 17) or treatment (n = 18) group and then studied for 2 weeks at 200 m. The first (control) group then proceeded to 4300 m (Pikes Peak, Co) within 5 hours in pressurized commercial aircraft and cars. The second (treatment) group proceeded in pressurized aircraft to 1600 m (Denver, Co). Since this study was double-blind only those days during which the subjects were on placebo are reported.

Each subject's moods were self-rated twice daily using the Clyde Mood Scale administered in a computer card, Q-sort format (to facilitate scoring). Moods were assessed initially at 200 m on Days 9 and 10 of the study, each day at the 1600 m staging site, and each day at 4300 m. The mood scale was always given after the morning and evening meals.

A two-way repeated measures analysis of variance was used to analyze for morning-evening, group, and interaction effects. Paired t-tests were used to analyze the baseline - altitude differences. Greater description and detail of the analyses are described elsewhere (Shukitt & Banderet, in press). A more stringent significance level of $p \leq .01$ was chosen to compensate for the several multiple comparisons used to analyze the data. Morning and evening values for each principal factor were calculated for each administration at 200, 1600, and 4300 m.

RESULTS

No differences were found between the morning and evening administrations on any of the factors at 200 and 1600 m. Therefore, these morning - evening values were averaged to produce a daily value. The morning - evening values for 4300 m were also pooled to produce a daily value to be comparable with the data for 200 and 1600 m. However, at 4300 m there was a significant day by time interaction found on the dizziness factor and the aggressiveness factor and a significant time effect was found on the sleepiness factor. At 200 m, a baseline value for each factor was calculated by averaging the four administrations since no differences were found between days 9 and 10.

Figures 1 and 2 show the daily value for each mood factor at 200, 1600, and 4300 m. The day 0 value reflects the day of ascent to high altitude, day 1 the first full day at altitude, etc. The scores for 200 m are the values for all subjects for the last two days at sea level (no differences were found

between groups at this altitude). Values for 1600 m were obtained from the second group mean scores on day 0 and day 1 since placebo was given on these days. Values for 4300 m were obtained from the first group mean scores on day 0 - day 3. Those mood states which were significantly different from baseline are marked with a double asterisk.

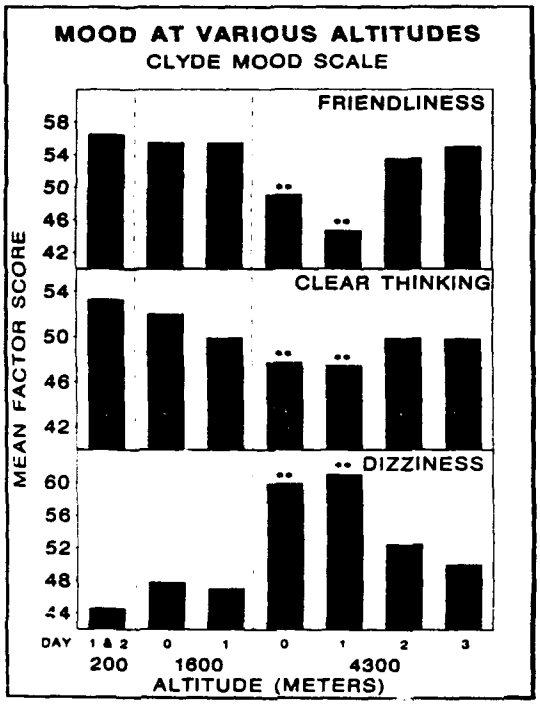


Figure 1. Factor scores for friendliness, clear thinking, and dizziness at 200, 1600, and 4300 m. A double asterisk indicates a significant difference from 200 m at $p \leq .01$.

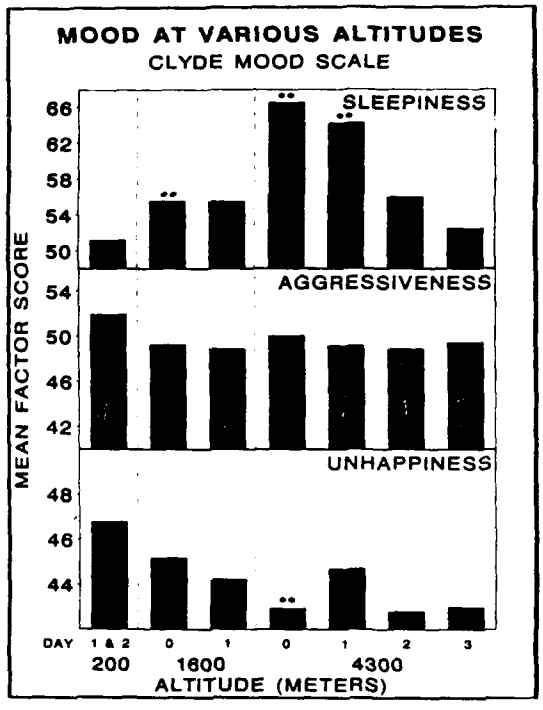


Figure 2. Factor scores for sleepiness, aggressiveness, and unhappiness at 200, 1600, and 4300 m. A double asterisk indicates a significant difference from 200 m at $p \leq .01$.

Friendliness mean scores are shown in Figure 1. Values at 4300 m on day 0 ($p \leq .002$) and day 1 ($p \leq .001$) were less than at 200 m. By day 2, however, friendliness was similar to that at 200 m. Friendliness did not change at 1600 m. Clear thinking factor scores are also shown in Figure 1. Subjects were less clear thinking on day 0 ($p \leq .005$) and on day 1 ($p \leq .004$) at 4300 m than at 200 m. Clear thinking returned to its original value by day 2. No differences were found at 1600 m. Figure 1 also shows the mean scores for the dizzy factor. On day 0 and day 1, subjects were more dizzy at 4300 m ($p \leq .002$ and $p \leq .001$, respectively) than at 200 m. No differences in dizziness were shown at 1600 m.

Sleepiness mean scores are shown in Figure 2. Subjects were more sleepy on day 0 ($p \leq .000$) and on day 1 ($p \leq 002$) at 4300 m than at 200 m. By day

2, however, sleepiness had recovered. Subjects also reported themselves as more sleepy at 1600 m on day 0 ($p \leq .008$). There were no differences for aggressiveness either at 4300 or 1600 m (Figure 2). Unhappiness, shown in Figure 2, had only one difference. At 4300 m, subjects were less unhappy (i.e., more happy) on day 0 ($p \leq .005$) than at 200 m. No differences in unhappiness were seen at 1600 m.

DISCUSSION

These results indicate that the Clyde Mood Scale is sensitive to altitude. Few mood changes occurred at 1600 m but several mood factors were affected at 4300 m. Moreover, mood factors at altitude showed distinctive changes with time and morning - evening moods were similar. This finding suggests that time of day need not be a major consideration in planning when to administer the Clyde Mood Scale, either at altitude or sea level.

No mood changes, except increased sleepiness, were observed at 1600 m. Armstrong (Van Liere & Stickney, 1963) noted that the most frequent symptom occurring at an altitude of 3700 m was sleepiness. However, at 4300 m sleepiness was third and at 4900 m it was fifth. This finding suggests that sleepiness may be the predominate symptom experienced at lower altitudes but at higher altitudes it is less important compared to other symptoms. This finding is also noteworthy because no other changes in mood were observed at 1600 m. These results imply that people flying on pressurized aircraft or traveling to such altitudes for work, recreation, or residence will experience few mood changes other than increased sleepiness (fatigue).

Five of the six factors on the Clyde Mood Scale were significantly different from baseline at 4300 m on Day 0. This finding is of special importance since these administrations were taken only 1 and 4 hours after ascent to altitude and they did not differ. It appears the Clyde Mood Scale can be used to measure mood changes even during the first few hours of hypoxia or high altitude studies, which is important since many of these studies are of short duration.

On day 1, 18 - 28 hours after ascent to altitude, all mood changes were greatest, except for sleepiness which was most severe 1 - 4 hours (day 0) after ascent. This finding supports the description of "high-altitude disease" by Monge which states that the first symptom to appear at altitude is a feeling of generalized fatigue, which bears no relation to the amount of work performed (Van Liere & Stickney, 1963). However, by day 2, 42 - 52 hours after ascent, all mood changes had recovered to baseline values.

Therefore, our results support previous research with AMS symptomatology since the most severe changes occurred 1 - 2 days (18 - 52 hours) after altitude ascent and then gradually subsided over the next 2 - 4 days. The Clyde Mood Scale detected changes in mood over time at altitude. Banderet (1977) found that this mood scale was also sensitive to changes produced by the treatment strategy (staging plus acetazolamide); this strategy resulted in improved moods at altitude. At 4300 m, subjects in the treatment group rated themselves as more friendly and less dizzy and sleepy than the control group subjects. Since the Clyde Mood Scale appears sensitive to high altitude effects, as well as treatment effects, it can be used in future high altitude studies as a subjective measure for mood.

SUMMARY

Personal anecdotes imply that ascent to high altitude causes mood changes such as depression, apathy, and drowsiness. Also, behaviors at high altitude suggest that people are more argumentative, irritable, or euphoric. Since systematic and quantitative studies assessing the effects of altitude on mood are few, mood was assessed in this study at two different altitudes and times of day using a standardized scale.

Self-rated moods were determined twice daily using the Clyde Mood Scale with 19 males and 16 females. Baseline (control) mood states were determined at 200 m. Moods were then assessed at 4300 m with one group and at 1600 m with the second group. Friendliness, clear thinking, dizziness, sleepiness, and unhappiness were affected at 4300 m. Only sleepiness changed at 1600 m. At altitude mood changes were different from baseline the day of arrival (1 - 4 hours), most severe after one day (18 - 28 hours), and back to baseline levels by day 2 (42 - 52 hours). Few time of day (morning - evening) differences were found. Therefore, this mood scale appears useful for assessing the effects of different altitudes on mood states.

REFERENCES

Banderet, L.E. Self-rated moods of humans at 4300 m pretreated with placebo or acetazolamide plus staging. Aviation, Space, and Environmental Medicine, 1977, 48(1), 19-22.

Banderet, L.E., B.L. Shukitt, R.S. Kennedy, C.S. Houston, and A.C. Bittner Jr. Cognitive performance and affective responses during a prolonged ascent to 7600 m (25,000 ft) simulated altitude. (Submitted for review).

Cain, S.M., and J.E. Dunn II. Low doses of acetazolamide to aid accommodation of men to altitude. Journal of Applied Physiology, 1966, 21(4), 1195-1200.

Carson, R.P., W.O. Evans, J.L. Shields, and J.P. Hannon. Symptomatology, pathophysiology, and treatment of acute mountain sickness. Federation Proceedings, 1969, 28(3), 1085-1091.

Clyde, D.J. Manual for the Clyde Mood Scale. Biometric Laboratory, University of Miami, Coral Gables, Fl., 1963.

Evans, W.O. Measurement of subjective symptomatology of acute high altitude sickness. Psychological Reports, 1966, 19, 815-820.

Evans, W.O., S.M. Robinson, D.H. Horstman, R.E. Jackson, and R.B. Weiskopf. Amelioration of the symptoms of acute mountain sickness by staging and acetazolamide. Aviation, Space, and Environmental Medicine, 1976, 47(5), 512-516.

Forwand, S.A., M. Landowne, J.N. Follansbee, and J.E. Hansen. Effect of acetazolamide on acute mountain sickness. The New England Journal of Medicine, 1968, 279(16), 839-845.

Hansen, J.E., C.W. Harris, and W.O. Evans. Influence of elevation of origin, rate of ascent and a physical conditioning program on symptoms of acute

mountain sickness. Military Medicine, 1967, 132(8), 585-592.

Houston, C.S. Going Higher: The Story of Man and Altitude. Burlington: Charles S. Houston, M.D., 1983.

Kobrick, J.L., and J.B. Sampson. New inventory for the assessment of symptom occurrence and severity at high altitude. Aviation, Space, and Environmental Medicine, 1979, 50(9), 925-929.

Sampson, J.B., and J.L. Kobrick. The Environmental Symptoms Questionnaire: Revisions and new field data. Aviation, Space, and Environmental Medicine, 1980, 51(9), 872-877.

Shukitt, B.L., and L.E. Banderet. Mood States at 1600 and 4300 Meters Terrestrial Altitude. (Submitted for review).

Stamper, D.A., R.A. Kinsman, and W.O. Evans. Subjective Symptomatology and cognitive performance at high altitude. Perceptual and Motor Skills, 1970, 31, 247-261.

Van Liere, E.J., and J.C. Stickney. Hypoxia. Chicago: The University of Chicago Press, 1963.

Ward, M. Mountain Medicine. London: Crosby Lockwood Staples, 1975.

ADDENDUM

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Human subjects participated in these studies after giving their free and informed voluntary consent. Investigators adhered to AR 70-25 and USAMRDC Regulation 70-25 on Use of Volunteers in Research.

EFFECTS OF VARIOUS ENVIRONMENTAL STRESSORS ON COGNITIVE PERFORMANCE

L.E. Banderet, Ph.D., B.L. Shukitt, B.A., E.A. Crohn, B.A.,
R.L. Burse, Sc.D., D.E. Roberts, Ph.D., & A. Cymerman, Ph.D.
US Army Research Institute of Environmental Medicine
Natick, MA 01760-5007

Rigorous testing instruments and psychometric methods are required to assess the effects of environmental stressors upon cognitive performance. Optimal instruments should be: 1) stable and sensitive, 2) given with minimal training and familiarization, 3) administered in a short time, 4) appropriate for test subjects with varied abilities, 5) useful in different environments, and 6) available in alternate forms for repeated assessment.

This paper summarizes six cognitive performance studies with environmental stressors which illustrate our approach and methodology for assessing environmental effects. The stressors included: hypobaric hypoxia, cold, dehydration, and atropine. The paper describes both our research findings and factors we surmise to be critical to the success of this approach.

METHOD

Subjects

A total of 87 men served as fully-informed medical research volunteers. Eighty were military personnel; seven were civilians.

Assessment Metrics

Cognitive performance was assessed with nine tasks. The Computer Interaction, Tower, and Map Compass tasks were developed in our laboratory (Banderet, Benson, MacDougall, Kennedy, & Smith, 1984; Jobe & Banderet, 1984); the other six tasks were adapted from the Navy's Performance Evaluation Tests for Environmental Research (PETER) Program (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984; Carter & Sbisa, 1982). All tasks were generated by computer and printed, off-line, on a laser copier. Each task had 15 alternate forms. Task descriptions and sample items were as described elsewhere (Banderet, Lieberman et al., 1986; Banderet, MacDougall et al. 1986; Banderet, Shukitt, Kennedy, Houston, & Bittner (in review)).

Procedures

Experimental conditions, number of subjects, and elapsed times for cognitive assessment for each study were as shown in Table I. Except for the Dehydration Study, all were repeated-measures experiments. The Inspired Air, Operation Everest II, and Tyrosine Evaluation studies investigated high altitude exposure in a hypobaric chamber.

Repeated testing procedures and methods were similar to those for the PETER Program (Bittner et al., 1984; Jobe & Banderet, 1984). Initially, subjects were trained and given extensive practice with performance feedback. To insure performance was stable and near-maximal, each task was completed

STUDY	N	CONDITIONS	ELAPSED TIME OF REPORTED MEASURES	REFERENCES
Inspired Air	23	4600 m 23 °C + 20% RH	1 or 6, 14 or 19, 24 or 29, 38 or 43 h	Banderet & Burse, 1984
Atropine	7	2 mg Atropine 40 °C + 20% RH	2.0 to 2.5 h	Banderet & Jobe, 1984
Cold & Dehydration	36	-24 °C + 4 mph winds Restricted Fluid Intake	50 & 54 h	Banderet, MacDougall, Roberts, Tappan, Jacey, & Gray, 1986
Dehydration	18[1]	2% Dehydration (body weight) 20 to 27 °C	9 h	Banderet, MacDougall, Roberts, Tappan, Jacey, & Gray, 1986
Operation Everest II	7	4600, 5500, 6400, 7600, 600, 600 m (23 °C + 75% RH)	8, 15, 24, 31, 39, & 41 Days	Banderet, Shukitt, Kennedy, Houston, & Bittner (in press)
Tyrosine Evaluation	24	4700 m + 15 °C (50% RH) Placebo	1.0 to 4.5 h	Banderet, Lieberman, Francesconi, Shukitt, Goldman, Schnakenberg, Rauch, Rock, & Meadors, 1986

NOTE: The predominate stressor in each study is listed first in the CONDITIONS column.

[1] These subjects were also in the Cold and Dehydration Study.

Table I.--Conditions for our studies of environmental stressors and their effects upon cognitive performance.

12-18 times before subjects were evaluated experimentally. All performance tasks were timed. The Tower, Computer Interaction, and Map Compass tasks were given typically for 5-6 min; all other tasks, for 3-4 min. Each task's actual duration, number of practice administrations, and other specifics were as described in the publications cited.

OUTPUT (number of problems attempted per minute) and ERRORS (number of problems wrong per minute) were calculated for each task. On tasks with limited response alternatives, ERRORS were adjusted to penalize for careless responding. A third performance measure (CORRECT) was calculated to reflect the combination of both problem solving and error rates. CORRECT (number of problems correct per minute) also included the adjustment for careless responding.

Statistical analyses were performed with Analysis of Variance and Student's t (one-tailed comparisons) statistics. Significance levels were $p \leq 0.05$.

RESULTS

The effects of practice on several cognitive tasks during baseline conditions are shown in Figure 1. Each task was practiced seventeen times in 9 days. Practice improved performance 30% (Coding) to 160% (Grammatical Reasoning) above initial values. Although increased practice resulted in diminishing gains in performance, performance was still improving even after 17 administrations.

Some environmental effects have dramatic timecourses. Figure 2 shows data from the same study after subjects were exposed to 4600 m altitude. Each cognitive task was significantly impaired (13-27%) from baseline values 1 or 6 hours after ascent. Impairments on Number Comparison (20%) and Addition (27%) were the greatest. With more time at altitude, performance returned to baseline values on most of the tasks, i.e. Coding, Grammatical Reasoning, Pattern Recognition, Pattern Comparison, and Computer Interaction.

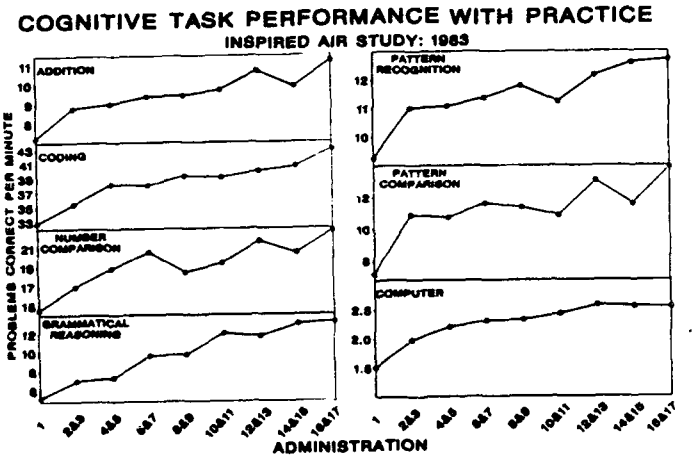


Figure 1.--Performances on seven cognitive tasks with practice.

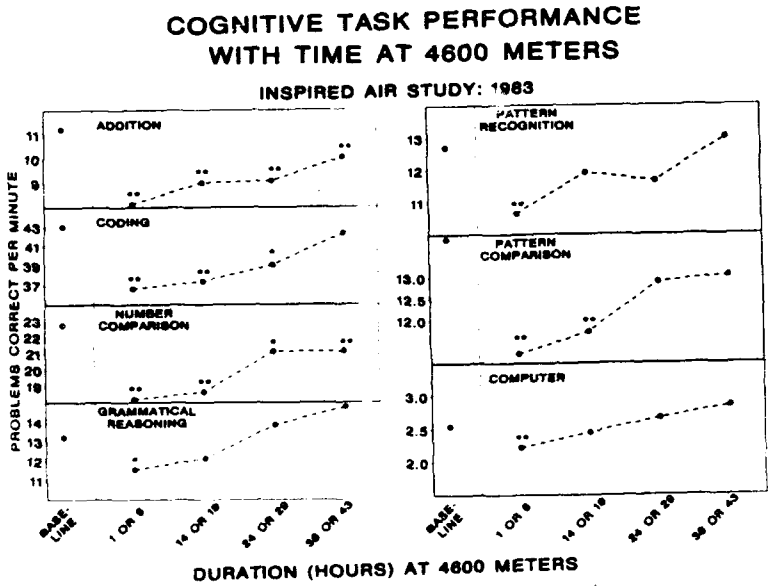


Figure 2.--Performances on cognitive tasks at 4600 meters after varied durations at altitude. Performance impairments, significantly different from baseline, are indicated with an asterisk above each data point.

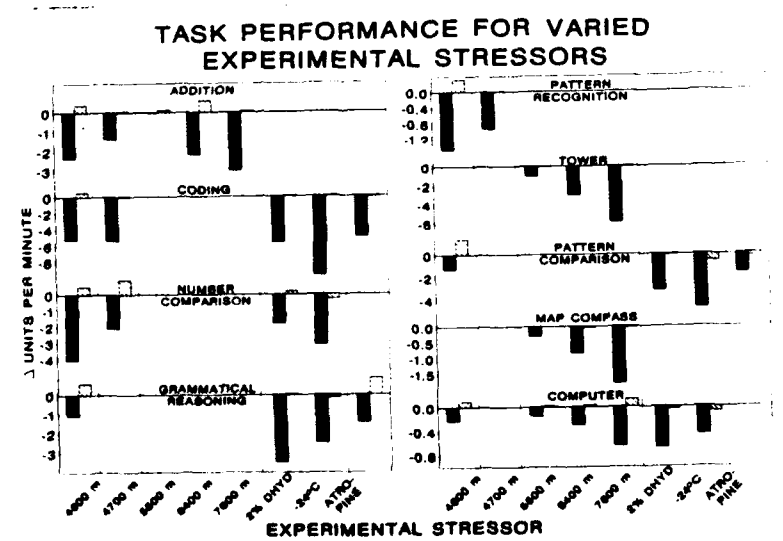


Figure 3.--Changes in task output or errors on nine cognitive tasks for varied stressors.

Cognitive performance was sensitive to a variety of stressful conditions. Impairments in cognitive performance are shown in Figure 3 for all stressors that we investigated. CORRECT, the measure influenced by both OUTPUT and ERROR rates, is not shown; however, it decreased significantly from baseline in all studies with the exceptions of Grammatical Reasoning (Dehydration and Cold Studies), Grammatical Reasoning (Atropine, $p \leq 0.10$), and Pattern Comparison (Atropine, $p \leq 0.10$). All nine tasks were not used in each study; bars are shown for those that were. Changes in OUTPUT are shown as solid bars; changes in ERRORS, as hatched bars. This figure shows slower problem-solving rates were responsible for the performance impairments observed for these varied stressors. ERRORS contributed little. Such OUTPUT impairments at 5500, 6400, and 7600 m increased linearly with increased altitudes during Operation Everest II.

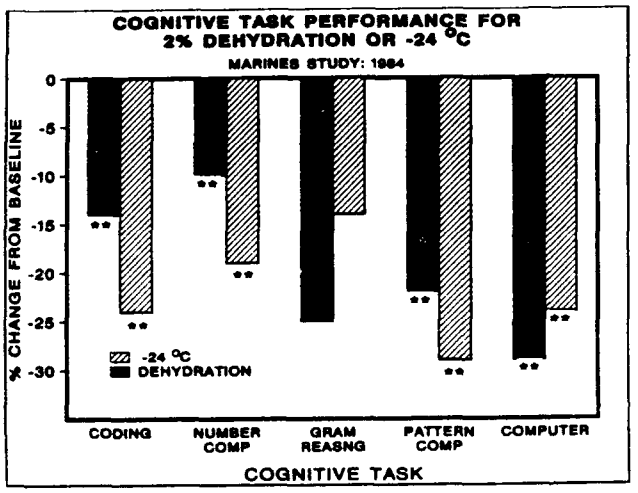


Figure 4.--Percent change from baseline on five cognitive tasks following 2% dehydration or exposure to -24°C and 6 km per hour winds.

The effects of 2% dehydration or windy cold upon five cognitive performance tasks are shown in Figure 4. Tasks involving verbal, spatial, or psychomotor processes were impaired 12-28% of baseline performance by these stressors. Grammatical Reasoning was not.

DISCUSSION

Impairments in cognitive performance were demonstrated for a variety of environmental stressors. Altitude impaired performance on all tasks at 4200 - 7600 m altitude. Furthermore, task performance at altitude was never significantly improved above baseline. With 2% dehydration or windy cold most tasks were impaired; however, Grammatical Reasoning was not. Atropine (2 mg) decreased Coding performance; however, impairments on Pattern Comparison and Grammatical Reasoning were marginally significant.

Impairments in performance resulted from a slowing of OUTPUT rather than increased ERRORS. This was a general finding across the stressors of hypoxia, dehydration, cold, and atropine. This is a very robust finding since our error adjustment exaggerated ERRORS, e.g. errors were doubled on tasks with only two response alternatives. Even with this exaggeration of actual errors, performance changes resulted from a slowing of problem solving.

The effects of altitude had a distinctive timecourse (Fig. 2). After 1 or 6 h at 4600 m all seven tasks were impaired; at 14 or 19 h four were impaired. At 38 or 43 h only two were still decremented. This information is critical for choosing appropriate times to evaluate environmental or treatment effects in altitude studies. It may also explain the negative findings in some earlier altitude studies.

These measures of cognitive performance can also be used to evaluate treatment effects. In data reported elsewhere (Banderet, Lieberman, et al., 1986) tyrosine, an amino acid, resulted in enhanced performance on the Addition, Coding, and Tower Tasks in a hypoxic and cold environment. Performances of the tyrosine-treated subjects did not differ from placebo-treated subjects on the Map Compass, Number Comparison, and Pattern Recognition tasks.

Our data demonstrate that cognitive performance deteriorates with environmental stressors. The fact that such impairments result with well-practiced and overlearned tasks suggests the sensitivity of our methodology. Adequate levels of stressors, enough subjects, practiced tasks with demonstrated stability and sensitivity, appropriate time sampling, and the establishment of near-maximum performance before experimentation are believed critical to our approach.

SUMMARY

Rigorous testing instruments and psychometric methods are required to assess the effects of environmental stressors upon cognitive performance. This paper presents findings and illustrates our methodology for evaluating the effects of several types of environmental stressors. Various cognitive performances were investigated experimentally with paper and pencil tasks in repeated-measures paradigms for several high altitudes, an altitude-treatment strategy, dehydration, cold, and atropine in a hot environment.

Cognitive performance was impaired on most tasks by each stressor. Impairments were usually due to decreases in the rate of performance rather than increased errors, e.g. problem solving rates decreased linearly from 4500-7600 m (15,000 - 25,000 ft) high altitude during a 40-day progressive exposure. Recovery of performance during 2 days at 4600 m depended upon the task; not all tasks improved fully. A treatment strategy (tyrosine) minimized altitude-induced performance impairments on some tasks.

Our results suggest even well-practiced and overlearned tasks deteriorate with environmental stressors. Adequate stressor levels, enough subjects, practiced tasks with demonstrated stability and sensitivity, appropriate time sampling, and the recruitment of maximum performance before experimentation are critical factors for our approach.

REFERENCES

Banderet, L.E., K.P. Benson, D.M. MacDougall, R.S. Kennedy, & M. Smith. (1984). Development of cognitive tests for repeated performance assessment. In Proceedings of the 26th Annual Meeting Military Testing Association, Munich, Federal Republic of Germany, 375-380.

Banderet, L.E., & R.L. Burse. (1984, August). Cognitive performance at 4600 meters simulated altitude. Paper presented American Psychological Association, Toronto, Canada.

Banderet, L.E., & J.B. Jobe. (1984). Effects of atropine upon cognitive performance and subjective variables (Technical Report No. T15/85). Natick, MA: U.S. Army Research Institute of Environmental Medicine.

Banderet, L.E., H.R. Lieberman, R.P. Francesconi, B.L. Shukitt, R.F. Goldman, D.D. Schnakenberg, T.M. Rauch, P.B. Rock, & G.F. Meadors III. (1986, in press). Development of a paradigm to assess nutritive and biochemical substances in humans: A preliminary report on the effects of tyrosine upon altitude- and cold-induced stress responses. In The Biochemical Enhancement of Performance: Proceedings of a Symposium, Lisbon, Portugal.

Banderet, L.E., D.M. MacDougall, D.E. Roberts, D. Tappan, M. Jacey, & P. Gray. (1986). Effects of hypohydration or cold exposure and restricted fluid intake on cognitive performance. In Predicting Decrements in Military Performance Due to Inadequate Nutrition: Proceedings of a Workshop, Washington, D.C.: National Academy Press, 69-79.

Banderet, L.E., B.L. Shukitt, R.S. Kennedy, C.S. Houston, & A.C. Bittner, Jr. Cognitive performance and affective responses during a prolonged ascent to 7600 m (25,000 ft) simulated altitude. (submitted for review).

Bittner, A.C. Jr., R.C. Carter, R.S. Kennedy, M.M. Harbeson, & M. Krause. (1984). Performance evaluation tests for environmental research: Evaluation of 112 measures (Report NBDL84R006 or NTIS AD152317). Naval Biodynamics Laboratory: New Orleans, LA.

Carter, R.C., & H. Sbisa. (1982). Human performance tests for repeated measurements: Alternate forms of eight tests by computer (Report NBDL8213003). Naval Biodynamics Laboratory: New Orleans, LA.

Jobe, J.B., & L.E. Banderet. (1984). Cognitive testing in military performance research. In Proceedings of a Workshop on Cognitive Testing Methodologies, Washington, DC: National Academy Press, 181-193.

ADDENDUM

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

Human subjects participated in these studies after giving their free and informed voluntary consent. Investigators adhered to AR 70-25 and USAMRDC Regulation 70-25 on Use of Volunteers in Research.

COMPUTERIZED ADAPTIVE TESTING HARDWARE/SOFTWARE DEVELOPMENT FOR THE U.S. MILITARY: INTRODUCTION AND OVERVIEW

· W. A. Sands*

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

Armed Services Vocational Aptitude Battery

The Armed Services Vocational Aptitude Battery (ASVAB) is a paper-and-pencil aptitude test battery used by all the U.S. military services for both enlistment eligibility screening and for subsequent classification and placement into entry-level training. The paper-and-pencil version of the battery (P&P-ASVAB) includes eight power tests and two speeded tests. Administration time for P&P-ASVAB is about three and one-half hours.

The P&P-ASVAB is administered under two large-scale testing programs. The Production Testing Program uses the battery in over 60 Military Entrance Processing Stations (MEPS) and over 900 Mobile Examining Team (MET) sites located across the country. The Student Testing Program is used in about 14,000 high schools. These two testing programs are quite large, each one involving between 800,000 and 1,000,000 persons annually.

Computerized Adaptive Testing

As suggested by the name, Computerized Adaptive Tests (CAT) differ from paper-and-pencil tests in administrative mode. They also differ in the way in which items are selected for administration. In the usual paper-and-pencil test administration, all examinees are given the same items in the same sequence. A CAT instrument, on the other hand, dynamically tailors item selection for each examinee during the course of test administration.

Typically, at the beginning of a test, no information is available on the examinee's ability level. Therefore, the examinee is assumed to have average ability and the initial item selected for administration is one of medium difficulty. If the examinee responds correctly, the ability estimate is raised, and a more difficult second item is selected. If the examinee answers this second item incorrectly, the ability estimate is lowered somewhat through the updating procedure. As a result, an easier item is selected as the third question. This process of selecting an item, scoring the examinee's response, updating the ability estimate, and choosing the next item for administration continues until some stopping rule is reached. This termination

* The opinions expressed here are those of the author and do not necessarily represent those of the Department of the Navy.

criterion may be either a pre-specified number of items (fixed length testing), or the administration of items until the ability estimate meets a pre-specified level of precision (variable-length testing), or a combination of the two approaches.

CAT Version of ASVAB

The Computerized Adaptive Testing (CAT-ASVAB) Program has two broad objectives. The first is to develop a system to automate the administration, test scoring, and computation of the Armed Forces Qualification Test (AFQT) score and various other composite scores that are derived from ASVAB and used by the individual military services. Such a system must be usable in both the fixed-base MEPS and in the portable testing environment of the MET sites, and must interface with the existing score reporting system. The second objective of the CAT-ASVAB program is to evaluate the suitability of CAT-ASVAB as replacement for the P&P-ASVAB in the Production Testing Program.

Accelerated CAT-ASVAB Program

As described in a paper presented to this conference last year (Sands, 1985), the original approach to the development of the system has been drastically changed. The current emphasis is on field-testing CAT-ASVAB as soon as possible, and has become known as the Accelerated CAT-ASVAB Program (ACAP). In line with this orientation, we are procuring off-the-shelf, commercially-available microcomputer hardware. Software design, development, test, and evaluation will be accomplished in-house at NAVPERSRANDCEN.

OVERVIEW

This symposium covers various aspects of the hardware and and software development in support of ACAP. Ms. Jones-James will set the stage by summarizing an MTA paper from last year that describes the ACAP network and computer hardware (Tiggle and Rafacz, 1985).

The first paper, entitled "Design and Development of the ACAP Test Item Data Base," will be presented by the author, E. Wilbur. She will describe the method used to place the test items onto the computer-based delivery system so that two alternate item banks are available.

The second paper, "Development of the Test Administrator's Station in Support of ACAP," was written by B. Rafacz and will be presented by E. Wilbur. This paper describes the tasks and responsibilities of the test administrator before, during, and, after the test session, and the software designed to assist the test administrator.

The third paper, "Design and Development of the ACAP Test Administration Software," will be presented by the author, G. Jones-James. After a brief description of functional requirements for the system, she will discuss both a networking and a stand-alone operating environment. Finally, she will present an overview of the software development for the examinee test station.

The last paper, "Communication of Computerized Adaptive Testing Results in Support of ACAP," was written by J. Folchi, and will be presented by G. Jones-James. This paper will present the ACAP functional specifications and equipment configuration for the Data Handling Computer (DHC). The procedures involved in data collection, data distribution, and failure recovery will be described. Finally, some possible extensions for the DHC are outlined.

After a period for questions from the audience, a hands-on demonstration of the ACAP version of CAT-ASVAB on the Hewlett-Packard Integral Personal Computer will be presented.

DESIGN AND DEVELOPMENT OF THE ACAP TEST ITEM DATA BASE

Elizabeth R. Wilbur†

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

The Navy Personnel Research and Development Center (NAVPERSRANDCEN) is currently the lead laboratory for a joint service program to develop a Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). This computerized battery is designed to replace the operational, paper-and-pencil version (P&P-ASVAB) currently used for the selection and classification of U.S. military enlisted personnel.

Prior to a full-scale deployment effort, NAVPERSRANDCEN will evaluate the fielding of CAT-ASVAB under the Accelerated CAT-ASVAB Program (ACAP). During ACAP, CAT-ASVAB will be deployed within the U.S. Military Entrance Processing Command (USMEPCOM) at two Military Entrance Processing Stations (MEPS) and all of the Mobile Examining Team (MET) sites within the two MEPS. A requirement of ACAP is to develop test item banks from which two alternate test forms will be developed; each form will contain items for the nine power tests and two speeded tests. Over 2000 test items were included in the source item banks; each form of each of the power tests is approximately 100 items in length, while the speeded tests range from 50 to 85 items. During CAT-ASVAB test administration, items for each alternate test form will be on the Hewlett Packard Integral Personal Computer (HP IPC), the computer chosen for ACAP test administration (Tiggle & Rafacz, 1985; Rafacz, 1986).

Problem

While the final location of the ACAP Test Item Bank was to be the HP IPC, the three components of the source item files resided at different locations at the outset, and were in various stages of development. Text portions containing the stem and response alternatives of the items and the item parameter files were stored on the VAX minicomputer at NAVPERSRANDCEN. The graphics components of the items, however, were only available as black-and-white line drawings in the experimental booklets used in the development of the Omnibus Item Pool for ACAP (Prestwood, Vale, Massey, & Welsh, 1985). Because item calibration was performed using the items from the experimental booklets, the task for test items requiring graphics was to create a computer graphic version of the drawings on the HP IPC comparable to the booklet drawings. Then the graphics would be merged with item text and parameters into a format compatible with the ACAP test administration software (Jones-James, 1986). In addition, the completed items had to be sized to fit within the ACAP delivery system memory requirements. That is to say, all test items comprising one alternate test form, including supplementary data files and test administration software, must fit into Random Access Memory (RAM) of the HP IPC. Within ACAP, an examinee testing station will have available 1.5 megabytes (MB) of RAM; (Rafacz, 1986).

† The opinions expressed here are those of the author and do not necessarily represent those of the Department of the Navy.

Maintenance of item security is essential. Steps taken to ensure this included: 1) storage of items on an electronic RAM disc during production (all RAM storage is volatile; i.e., data are lost at power-down), 2) storage of items permanently on microfloppy discs only, not on hard discs, 3) encryption of the test item files on the microfloppy discs, 4) minimization of the number of discs containing items, 5) monitoring copies of discs and experimental booklets, and 6) securing the experimental booklets and discs containing items in a locked container.

METHOD

The method used to develop the ACAP test items and the resulting alternate forms for use in CAT-ASVAB test administration involved four phases. Initially, specifications for the display of test items on the ACAP delivery system screen were developed by the psychometric community. Using these specifications as guidelines, a computer hardware and software system was developed for producing items containing graphics for the HP IPC. Once this system, called the Image Capturing System, was developed, actual production of the graphic items could commence. Finally, the alternate test forms were developed on the HP IPC in an optimal format for all types of items, including those with, and without, graphics.

Item Display Specifications

A font selection program was written to assist NAVPERSRANDCEN researchers from psychometrics and human factors in the development of display specifications for the ACAP test items. This software package facilitated the choice of the font, format and screen characteristics used for item display. A 7x11 standard font was selected for the text display. Criteria used in font selection were size and quality of the letter form, readability and spacing of the letters, and the ease of distinguishing one letter from another. For the tests of Mathematics Knowledge, Arithmetic Reasoning and General Science, a set of control characters was designed to write a graphic representation of fractions, radical signs, exponents and chemical subscripts. Screen characteristics determined to be optimal for examinees included amber foreground letters on a black background for items with text only. For graphic items, the portion of the screen occupied by the text was as described above, while the graphic portion of the screen was an amber background with the black line drawing in the foreground. The screen layout of the items varied for the subtests, depending on the length of the items and the presence of graphics in the subtest. Margins were minimized for tests with densely packed text or long items (e.g., Paragraph Comprehension, Arithmetic Reasoning and General Science). Wider margins and spaces between alternatives were used for subtests with short items or with special characters such as subscripts, superscripts, and fractions (e.g., Word Knowledge and Mathematics Knowledge).

The Image Capturing System

The development of an efficient graphic Image Capturing System was the first step in producing graphics for the ACAP test items (Bodzin, 1986). The original system consisted of the following components: a) an IBM Personal Computer XT (IBM PC/XT) microcomputer, b) a Datacopy 700 optical scanner (i.e., a digitizer), c) the Word Image Processing System (WIPS) software for the IBM, d) an HP IPC, e) Datacomm communications software for the HP IPC, and f) in-house software written for the HP IPC. The IBM PC/XT was chosen for its compatibility with the WIPS software, a necessary interface with the Datacopy 700, and was configured with 2.5 megabytes of RAM. The HP IPC was configured with five megabytes of RAM. The additional RAM in the two machines not only increased the speed of

production, but improved item security by having items stored only in RAM during production. Permanent storage of the items was on floppy discs for both machines.

Recently, a Hewlett Packard Vectra computer has replaced the IBM PC/XT in the Image Capturing System. The Vectra uses an Intel 80286 microprocessor and, operating at 8 Megahertz (Mhz), is a much faster machine than the IBM PC/XT which operates at 4 Mhz. The increased speed will greatly reduce the time for future item production and modification of the existing item bank. In addition, the Vectra is able to directly prepare 3.5-inch discs for the HP IPC. (The standard size for IBM PC/XT is 5.25 inches.) The 3.5-inch discs, prepared in MS-DOS format on the Vectra, can then be converted to HP-UNIX format using the Oswego software utilities on the HP IPC. This eliminates the need for the transfer of items between computers via telecommunications.

Critical for administration of the ACAP test items is the device used for display of the items. The HP IPC screen display is 512 (horizontal) pixels by 255 (vertical) pixels with a uniform resolution of 2.8 pixels per millimeter in both vertical and horizontal directions; the screen size is 9 inches measured diagonally, 8 inches wide by 4 inches high. The colors available are amber and black, with one as foreground and the other as background color.

Graphic Item Production

The next step was to create a computer graphic version of the approximately 400 drawings in the Omnibus experimental booklets (see Figure 1). Five of the nine power tests contained graphics associated with some or all of the items. These tests were Automotive Information (AI), Shop Information (SI), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electrical Information (EI).

Item graphics were scanned from the Omnibus experimental booklets using the Datacopy 700 optical scanner, WIPS software, and the IBM PC/XT. Scanning time was approximately six seconds per image. The image bit-map representation was stored in a RAM disc file on the IBM PC/XT and saved with a surrounding border useful for subsequent editing purposes. The optimal size for the display of each image on the HP IPC screen was calculated on the HP IPC, based upon the shape of the image (Bodzin, 1986). In addition, due to differences between the IBM and HP IPC microprocessors, binary word sizes were included in the image size calculations. The WIPS software was then used to scale the image to the target size, preserving the original aspect ratio of the drawing in the booklet. The graphic images were stored with a unique item identification number (UID) on IBM PC/XT discs. The loss of information inherent in the scaling process resulted in degradation of the quality of the graphics. This necessitated restoring the graphic image to the original booklet quality using the WIPS graphics editor on the IBM PC/XT. The editing time for each of the 400 images ranged from fifteen minutes to six hours, with the average time approximately one and one-half hours per image. In the future, with the HP Vectra in the Image Capturing System, capturing, scaling and editing time will be reduced dramatically (currently estimated at a 20% reduction on the average).

After editing, the graphic images were transferred to the HP IPC via a serial communications line and the Datacomm software. A routine was written on the HP IPC to eliminate the border around each image and to save the optimal image size for the HP IPC screen, thereby reducing the storage requirements for the graphics. In addition, the image file was reformatted for use in the ACAP test administration program.

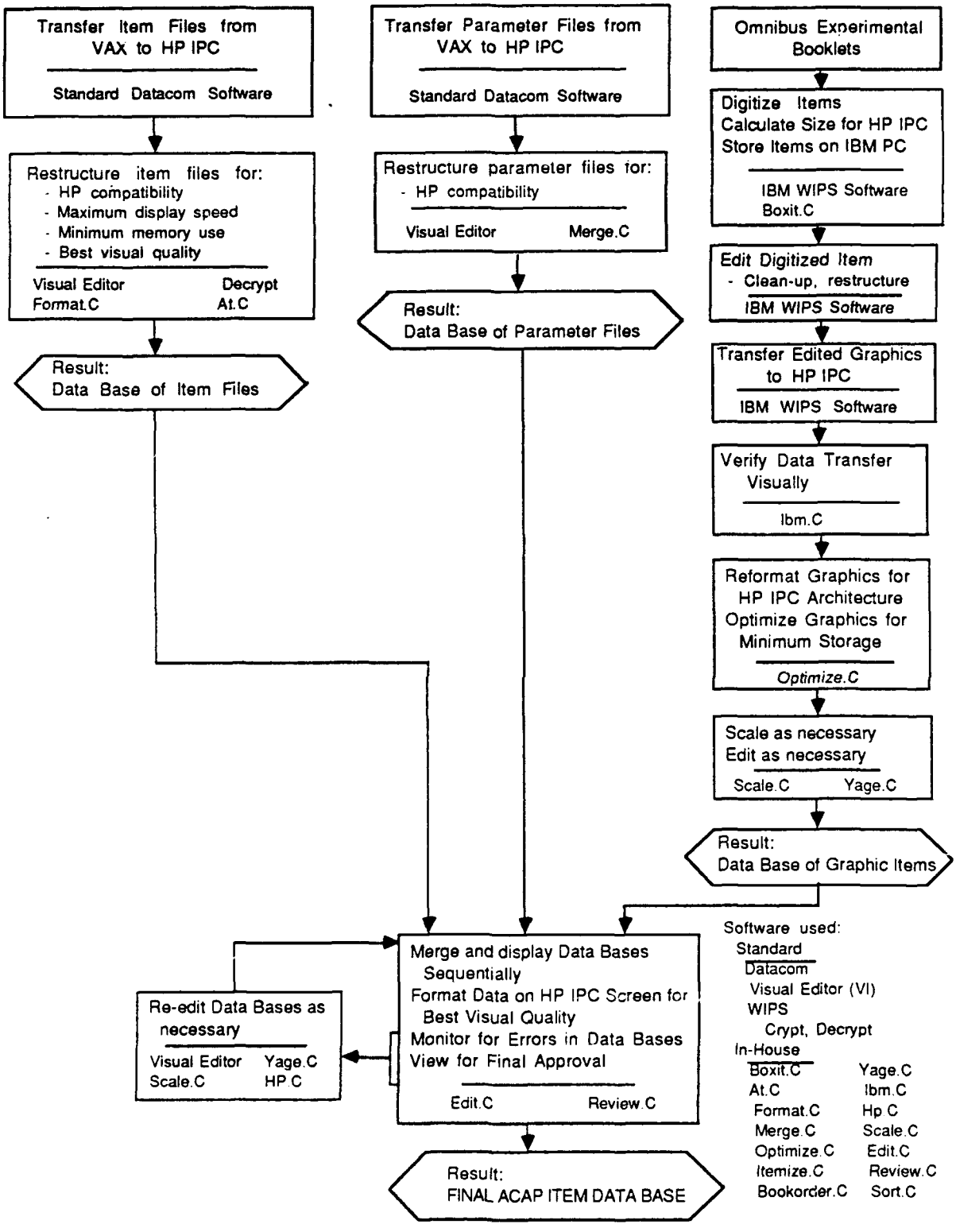


Figure 1. ACAP Test Item Bank Development

ACAP Alternate Test Forms

In preparation for the final processing phase to produce ACAP alternate test forms, the item text files and parameter files were transferred from the VAX to the HP IPC via serial communication line (see Figure 1). Software routines were written to reformat the files and to eliminate data fields not necessary for the purposes of ACAP test administration. The items were then reviewed for typographical errors and corrected. A review of the items for content and quality standards, and for sensitivity to minority issues, was conducted by Educational Testing Service and NAVPERSRANDCEN researchers. Those that passed the item content, quality and item sensitivity reviews were retained in the test item banks.

In the final step of production, item text, graphics and parameters were merged using a specially developed software package, the Item Image Editor. The graphics image was centered either above or to the left of the text. Items were reviewed again for errors. As the items were stored, the graphics components of the items were compressed reducing the storage requirements to approximately 60 percent of the original size.

After the production phase, the items were compiled into the two alternate test forms. Each form contained nine power tests and two speeded tests to be administered during CAT-ASVAB testing. For security purposes, final versions of the alternate test forms were encrypted and stored on HP IPC micro-floppy discs designated as ACAP System Discs. The procedure for use of the ACAP System Discs by an ACAP Test Administrator during CAT-ASVAB test administration is documented by Rafacz, 1986.

SUMMARY

Initially, the item bank development phase of the Accelerated CAT-ASVAB Program involved preparing the individual components (text, graphics, and parameters) of candidate test items for subsequent optimal storage on the ACAP computer system. After processing to achieve the psychometric requirements for item display, the final forms of the item text, graphics, and parameters were merged to create an ACAP item. With the compression of the graphics components of the items, storage requirements for the test items were reduced to 60 percent of the original size. Selected items were merged to create two alternate test forms according to CAT-ASVAB psychometric requirements. These forms will be displayed during ACAP test administration on the Hewlett Packard Integral PC.

REFERENCES

Bodzin, L. J. (1986). *An image capturing and editing system for the HP-Integral Computer.* A Naval Oceans Systems Center, Code 943, unpublished manuscript.

Jones-James, G. (1986). *Design and Development of ACAP Test Administration Software.* Paper presented at the 28th Military Testing Association (November, 1986).

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an Adaptive Item Pool.* (AFHRL-TR-85-19). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rafacz, B. A. (1986). *Development of the Test Administrator's Station in support of ACAP.* Paper presented at the 28th Military Testing Association (November, 1986).

Tiggle, R. B. and B. A. Rafacz. (1985). *Functional requirements for the Accelerated CAT-ASVAB Program (ACAP) Development System.* A NAVPERSRANDCEN, Code 63, unpublished manuscript.

DEVELOPMENT OF THE TEST ADMINISTRATOR'S STATION IN SUPPORT OF ACAP

Mr. Bernard A. Rafaczt

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

ABSTRACT

This paper describes the duties of the Test Administrator (TA) for the purposes of the Accelerated CAT-ASVAB Program (ACAP). A description of the ACAP computer hardware will be provided, followed by the duties of the TA in the three phases of operation: a) computer equipment transport and setup at the testing site, b) examinee test administration duties, and c) computer equipment takedown and transport. Finally this paper will discuss the software development to automate the TA functions on the ACAP computer hardware.

The Navy Personnel Research and Development Center (NAVPERSRANDCEN) is involved in a major system development effort that concerns the research, development, test, and evaluation of a Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). This is a joint-service effort that intends to implement CAT-ASVAB on a nationwide distributed processing network for the selection and classification of enlisted personnel. The CAT-ASVAB network of computers will allow the United States Military Entrance Processing Command (USMEPCOM) to adaptively administer the ASVAB to applicants for military service.

Prior to a full-scale deployment effort, NAVPERSRANDCEN will be refining the operational concept and the functional specifications required of the CAT-ASVAB computer network through the Accelerated CAT-ASVAB Program (ACAP). The goal of ACAP is a limited deployment of CAT-ASVAB at two of the 68 Military Entrance Processing Stations (MEPS), and at all of the Mobile Examining Team (MET) sites within the two MEPS. The deployment within ACAP will be accomplished with computer hardware and software that matches, as closely as possible, the functional specifications for CAT-ASVAB as documented in a CAT-ASVAB Stage 2 Full Scale Development (FSD) Request for Proposal (RFP) dated 6 June 1985. A recent paper (Tiggle and Rafacz, 1985) describes the commercially available computer hardware that will be used for ACAP.

Current CAT-ASVAB Operational Concept

Briefly, the CAT-ASVAB Stage 2 FSD specifications describe the operational deployment of CAT-ASVAB within USMEPCOM as follows: At each MET site, a Test Administrator (TA) will be required to move a collection of transportable packages containing computer equipment, identified as a Local CAT Network (LCN), from a storage facility to the MET testing site. Each LCN consists of one TA Station, a collection of Examinee Testing (ET) Stations, and one or more packages that contain peripheral support equipment (e.g., power cords and network cables). Once at the testing site, the TA will be required to configure the stations into a network to support an automated monitoring capability for the TA. The resulting network of computers will accomplish examinee assignment to testing stations, the adaptive administration of nine power tests, the non-adaptive administration of two speeded tests, and the collection of a host of examinee testing information. Once testing is completed, the TA will either telecommunicate the data, or send them on an appropriate media via registered mail to the responsible MEPS. The LCN will then be reconfigured into transportable packages, and returned to the storage facility.

Aptitude testing also occurs at the MEPS. The equipment is normally stationary, but must be identical to MET site equipment. Identical LCN components for the MEPS and MET sites are

† The opinions expressed here are those of the author and do not necessarily represent those of the Department of the Navy.

necessary to support equating CAT-ASVAB with P&P-ASVAB, to minimize software and hardware maintenance requirements, and for the sharing of equipment among all testing sites. The MEPS CAT-ASVAB equipment also includes a Data Handling Computer (DHC) which would compile the data received from the MEPS/MET testing sites within the MEPS. A portion of these data would be telecommunicated to the currently-operational MEPS UNIVAC System 80 computer system, and the full complement of testing data would be sent to USMEPCOM Headquarters (HQ). At USMEPCOM HQ, the data from each of the MEPS would be compiled and all of the data sent periodically to NAVPERSRANDCEN. Folchi (1986) describes the operation of the DHC for purposes of ACAP.

ACAP LCN Transportable Packages

Within ACAP, for each set of n ($n \leq 30$) ET Stations in an LCN, n+3 packages must be transported by the TA from the equipment storage site to the testing room. The following outline includes a description of the components in each transportable package:

ET Stations. A total of n packages is required with each package containing the following components: (a) an examinee testing station consisting of the Hewlett Packard (HP) Integral PC [8 MHz CPU with 512 KB of main internal RAM, 256 KB of ROM (containing the UNIX System 5.0 kernel), one microdisc drive (710 KB capacity), an adjustable electroluminescent screen with a resolution of 512 (horizontal) X 255 (vertical) pixels (screen size is 9 in. measured diagonally; 8 in. wide x 4 in. high), HP-IL network board installed in Slot A, and a 1 MB RAM board installed in Slot B]; (b) a dust cover (with three pockets) to protect the testing station during travel; (c) an Examinee Input Device (EID) developed from the current HP Integral keyboard; and (d) a 110 VAC power cord and a 1 meter HP-IL network cable, stored in a pocket of the dust cover.

NOTE: The total computer program directly addressable RAM memory is 1.5 MB. Total weight of each ET Station package is 25 lbs. This package includes a padded carrying handle and measures 19 in. wide x 13 in. high x 7.5 in. deep.

TA Station. Two packages are required with the first package containing the following components: a TA test session monitoring station consisting of the HP Integral PC with the same characteristics as the ET Station, mentioned above in (a), (b) and (d). This station also includes a built-in 150 cps inkjet printer, a RAM Expansion Box (REB) board installed in Slot B, a typewriter style, 128-character ASCII keyboard (with 8 function keys and a numeric keypad), and a 110 VAC power cord and several 1 and 5 meter HP-IL network cables stored in a pocket of the dust cover.

NOTE: The front panel that protects the screen is the keyboard. Total weight of the package is 26 lbs.; it includes a padded carrying handle and measures 18 in. wide x 13 in. high x 7.5 in. deep.

The second package within the TA Station includes the REB unit with 3 MB RAM installed in three of the five available expansion slots. In addition, the package contains the REB to TA Station cable, and a 110 VAC power cord. Components of this package are transported in a light-weight aluminum case measuring 21 in. wide x 14 in. high x 7 in. deep. The package weight is 21 lbs.

Peripheral Equipment. The final package is a cloth carrying bag, used to transport any peripheral equipment needed in support of the aforementioned equipment or test session administration. Materials or equipment that may be found in this package are: (a) several 110 VAC power strips (six outlets) with voltage surge protectors (includes a 6 ft., 3-prong plug power cord), (b) several 110 VAC extension cords, (c) several two-prong to three-prong plug adaptors, (d) pencils and scratch paper, (e) printer paper for the TA Station, (f) spare TA and ET Station discs, and (g) a collection of discs containing testing data (identified as DataDiscs).

NOTE: The ACAP System Master (and Backup) discs contain encrypted test item files and will be secured by the Test Administrator. These discs would not normally be included in any of the aforementioned packages. Jones-James (1986) includes a complete description of the data files on the ACAP discs [three System Master discs (labeled System Disc A, B or C), a TA Station disc (labeled TA Disc), ET Station discs (each labeled ET Disc), and a set of discs containing testing data, identified as DataDiscs].

Test Administrator Duties

Initially, if the testing equipment is conveyed to the testing site in a vehicle, the TA carries all testing equipment into the testing room. If the equipment is secured at the testing site, the TA carries the equipment into the testing room, as necessary.

TA Setup Duties. The TA identifies the TA Station and sets it on the floor, near the table reserved for this station and carries the case that contains the REB to the TA's table and places it on that table. The carrying case containing the peripheral equipment is also placed on, or near, the TA's table.

The REB unit, its AC power cord, and the REB cable (that connects to the TA Station) are removed from their carrying case and placed on the TA's table. The power cord and the REB cable are connected to the rear panel of the REB. The REB is then moved to a position on the TA's table that is convenient for subsequent test monitoring purposes.

The TA Station is then placed on top of the REB, and the following actions taken: (a) once the dust cover is removed, the power cord and any HP-IL network cables are removed from a pocket of the dust cover, (b) the top lid of the computer is unlocked and rotated to the rear of the station. This permits the keyboard to be released and placed on top of the table, in front of the TA Station, (c) the cable to the keyboard is then installed in an HP-HIL slot found on the front of the TA Station, (d) the REB cable is then connected into the appropriate receptacle found on the rear panel of the TA Station, (e) the AC power cord is installed into an AC adapter found on the rear panel of TA Station, (f) two 5-meter network cables are then installed into the HP-IL network board found on the rear panel of the TA Station (the other ends of these cables are placed at the locations of two adjacent ET Stations), and (g) paper is installed into the printer of the TA Station.

The TA then removes the transport disc from the TA Station disc drive, places this disc into the computer's storage compartment, removes the TA Disc from the same compartment, and installs this disc into the disc drive. The TA then uses a single power strip and/or extension cord to connect both the TA Station and the REB to a 110 VAC power source.

At this time, the TA turns on the power switch to the REB, and then the power switch to the TA Station. Once the initial boot-up process is completed, the TA inserts ACAP System Disc A into the disc drive when instructed. This commences the installation of a portion of the test item banks into RAM of the TA Station.

NOTE: If for some reason the TA Station should fail the bootup process, the TA may continue the testing session in a standalone mode of operation. In this mode, any other ET Station will serve to perform the TA Station test administration functions (to be discussed); all stations are interchangeable up to mode of operation. However, it will not be possible to broadcast software and data files to the ET Stations. In that case, each ET Station's bootup process will also have to include the installation of test administration software and data files. The net effect will be a longer time period prior to commencing testing than if the designated TA Station had not failed.

While System Disc A data is being processed at the TA Station, the TA proceeds to setup ET Stations. For each station, the steps include: (a) carrying an ET Station to a testing booth, (b) removing the AC power cord and the HP-IL network cable from the pockets of the dust cover, and removing the cover, (c) unlocking the top lid on the ET Station, removing the front (screen) protective panel (which includes the EID), and rotating the lid to the rear of the ET Station, (d) removing the transport disc from the disc drive, removing the ET Disc from the computer's storage compartment, and installing this disc into the disc drive, (e) rotating the top lid back to its original position and locking it in place, (f) installing the power cord in the AC adaptor found on the rear panel of the station, and (g) installing a network cable into the HP-IL receptacle on the rear of the station.

Once all of the ET Stations have been setup according to the aforementioned procedure, the TA would then connect each ET Station to a power strip and each power strip to a 110 VAC electrical outlet. Electrical extension cords will be available in the event the power cord on a power strip is not of sufficient length to reach an electrical outlet.

The TA will walk from one ET Station to another turning on the power to these units. Once all units have been powered up, the TA will confirm that each station has passed the boot-up procedure.

If not, that ET Station is powered down, removed from the network, and recorded for repair.

NOTE: During the process of setting up ET Stations, the "buzzer" may sound at the TA Station. This is a cue to the TA to install the next ACAP System disc, as instructed by that station.

Finally, all stations have passed the bootup procedures and all of the test item banks, and supporting data files, have been installed in RAM disc files of the TA Station. The TA Station will confirm the (electronic) integrity of the network for the LCN and proceed to download CAT-ASVAB test administration software and test item data files to each ET Station, as appropriate. In the event the integrity of the network cannot be confirmed, the TA will continue the testing session in a standalone mode of operation, as discussed in the *NOTE* of the previous page.

TA Examinee Administration Duties. At this point, the LCN is ready for actual CAT-ASVAB test administration. However, prior to admitting examinees into the testing room, the TA must identify to the TA Station: (a) all ET Stations available for testing, and (b) those examinees expected to be processed for testing during this session. Case (a) should be executed first. This involves maintenance on a set of ET Station ID numbers recorded on a file of the TA Disc. Using menu-driven software on the TA Station, the TA can either: a) *CREATE* a new set of ID numbers, b) *ADD* to the current set, c) *DELETE* from the set, or d) *LIST* (on the screen or printer) all currently recorded ID numbers. The TA will use function keys on his/her keyboard to select the appropriate option from a menu. For example, if an ET Station failed the bootup process, and is therefore not included in the LCN, the TA should select the option to *DELETE* the ID number for that station from the current list. (A demonstration included with the presentation of this paper will illustrate the ease with which a TA may perform such file maintenance activities.)

Once all the ET Stations in the LCN have been correctly recorded, the TA is now ready to identify examinees to be tested during a selected session. Note that this process may be accomplished at any time prior to the beginning of the testing session, at the convenience of the TA. Initially, this involves declaring the date and time of the target testing session. The subsequent interactive dialogue between the TA and the TA Station, and corresponding file maintenance activities, are now synchronized for the session. Again, using menu-driven software, the TA must choose one of the following options: a) *PROCESS* examinees for the target testing session [i.e., *Create* a new set of examinees to be tested in the testing session, *Add* to the current set, *Delete* from the set, or *List* (on the screen and/or printer) all examinees recorded for testing in the target testing session.], b) *ASSIGN* examinees to ET Stations, c) *SUBMIT* examinee personal data, d) *COLLECT* examinee testing data from the ET Stations, or e) *RECORD* all examinee testing data from all stations in the LCN onto a DataDisc. These functions should be performed sequentially, and include all of the steps necessary to record examinee testing data onto a DataDisc for subsequent transmission to the DHC at the MEPS. Option a)-*PROCESS*-records all examinees to be tested in terms of Name and Social Security Account Number (SSAN). Option b)-*ASSIGN*-randomly assigns examinees to the currently recorded set of ET Station ID numbers, while Option c)-*SUBMIT*-permits the TA to record any personal data that may be desired on the examinees. Option d)-*COLLECT*-permits the TA Station to receive the full complement of examinee testing data recorded during testing from individual ET Stations, and, finally, Option e)-*RECORD*-compiles all of the examinee testing data onto a single DataDisc. *PROCESS* and *ASSIGN* must be performed prior to examinees entering the testing room, *SUBMIT* is performed during testing, *COLLECT* at the conclusion of each examinee's test, and *RECORD* can only be performed once ALL examinees have completed testing.

Prior to examinees entering the testing room, the TA will have verified that the set of ET Station ID numbers are current, identified (by Name and SSAN) the examinees to be tested, and *ASSIGNED* those examinees to the ET Stations. The TA will also be in possession of the printout from the TA Station that lists the examinees and their ET Station assignments. Examinees may now be admitted into the testing room, and, as they arrive, the TA will direct each examinee to his/her assigned station. Simultaneously with their arrival, the TA will collect the USMEPCOM 714-A form (recording various personal information) from each examinee. Once all examinees are seated, the TA will provide a short (verbal) description of the CAT-ASVAB testing program, and reconcile any Privacy Act information and/or forms that need to be completed. At this point, the examinees will be instructed to press the ENTER key on the EID, commencing the test administration process [see Jones-James (1986)].

During examinee test administration, the TA will have to complete the entry of examinee personal data at the TA Station, if not already completed. In addition, as examinees complete testing, the TA will *COLLECT* testing data from their station, where they will be automatically recorded on a RAM disc file at the TA Station. However, the most critical duties of the TA, at this time, include responding to "HELP" requests from examinees, and attending to certain failure recovery procedures in the event some station in the LCN fails. In the former situation, an examinee has pressed the HELP key on the EID, triggering an interruption of the testing process at that station. The TA would be informed of this situation on the screen of the TA Station, and also by the examinee as he/she is instructed to "raise your hand" when needing HELP. The TA would then walk to the subject ET Station and be assisted there by an on-screen dialog at that station that will resolve the difficulty.

NOTE: If a station in the LCN should fail during test adminstration, failure recovery procedures are available to assist the TA. If an ET Station should fail, the TA will instruct the examinee at the subject station to wait for an ET Station to be free. At that time, the TA removes the ET Disc from the failed station, and inserts that disc into the disc drive of the now available station. As the ET Disc contains a log of the examinee's testing history, as well as a backup of current testing data, it is possible to restart the examinee at the new station at the beginning of the first non-completed test without any loss of data from previous tests. In addition, should the TA Station fail, any ET Station can serve as the "new" TA Station. Because of the interchangeability of stations in an LCN, it is almost impossible that the testing session not be completed; i.e., excluding failure of electrical power.

TA Takedown Duties. The testing session has been completed at the testing site and all Examinees have been excused. It is now necessary for the TA to prepare a DataDisc; this disc will be mailed to the parent MEPS via registered mail, together with the examinees' 714-A forms. The TA will accomplish this task by selecting the *RECORD* option in a previously discussed menu. Note that only one DataDisc is necessary to record all examinee testing data for any one testing session.

It is now necessary for the TA to configure the equipment comprising the LCN into transportable packages. The procedure for performing this function is essentially the reverse of the operations discussed in the section titled **TA Setup Duties**, and will not be detailed here. As a final step, the TA carries these packages from the testing room to a vehicle for transport to a storage site. If the equipment is stored at the testing site, the packages are then secured at that location.

TA Station Software Development

All of the software development for the ACAP computer system is being developed in the 'C' programming language. The use of this language was motivated in large part due to it being native to the UNIX operating system available on the HP Integral PC, and by certain characteristics of 'C' which greatly aid software development, performance, and testing. These considerations include: a) support of structured programming, b) portability, c) execution speed, d) concise definitions and fast access of data structures, and e) real time system programming.

The software development effort proceeded using a top-down, structured design approach. Initially, TA Station functional requirements were developed and documented, resulting in a macro-level design for subsequent software development. Primitive routines and procedures (e.g., a routine to permit the TA to submit a valid SSAN) were identified and then tested as "standalone" operations. Simultaneous with this effort, detailed (hard-copy) interactive screen dialogues were developed. Then, using the primitive routines, main stream code was developed that automated the interactive dialogues and underlying file maintenance functions. This software was then thoroughly tested. Finally, the main stream code was interfaced with specially-developed LCN networking protocols to permit communication between the TA and ET Stations in an LCN. To date, software to support the TA duties (previously discussed) that lend themselves to automation has been developed and tested for use in an LCN environment. In addition, networking functions that include the downloading of software and data files to ET Stations are currently being developed, as well as an elementary monitoring capability. In the months to come, the monitoring capability will be enhanced, an automated HELP function will be installed, and the process of moving testing data from ET Stations to the TA Station at the conclusion of testing will be automated. Currently, the TA manually moves the ET Disc (containing the testing data) to the TA Station at the conclusion of an examinee's test. In addition, it is anticipated that

modifications to the TA Station software will be requested by consumers once they review the product.

To give the reader an indication of the ease with which the TA may use the software, consider Figure 1, a typical interactive screen. The situation displayed in Figure 1 is a menu of options for file maintenance of ET Station ID numbers as described above. The TA keyboard is locked out except for the function keys located at the top of the keyboard; pressing an invalid key results in a low-level buzzer sounding. The function keys align one-on-one with the eight boxes at the bottom of the screen in Figure 1. Selection of function keys f1 through f4 results in immediate transfer to a set of interactive screens that support the functions of *CREATE, ADD, DELETE, or LIST* for ET Station ID numbers. Key f5-*STATUS*-gives the TA a status report on ET Stations in the LCN; examinees being tested, station ID, current test ID, total testing time, expected time to complete the testing session, etc. Function key f6-*INVERT*-toggles the screen background from amber to black, and conversely. Function key f7 provides for on-line HELP for the TA, in the event the TA is confused and wishes assistance. Finally, function key f8 returns control to the most recently executed menu, in the event the TA really did not want to be at this menu. In other words, the TA has the opportunity to change his/her mind, without penalty; all data file maintenance activities must be confirmed before updating occurs.

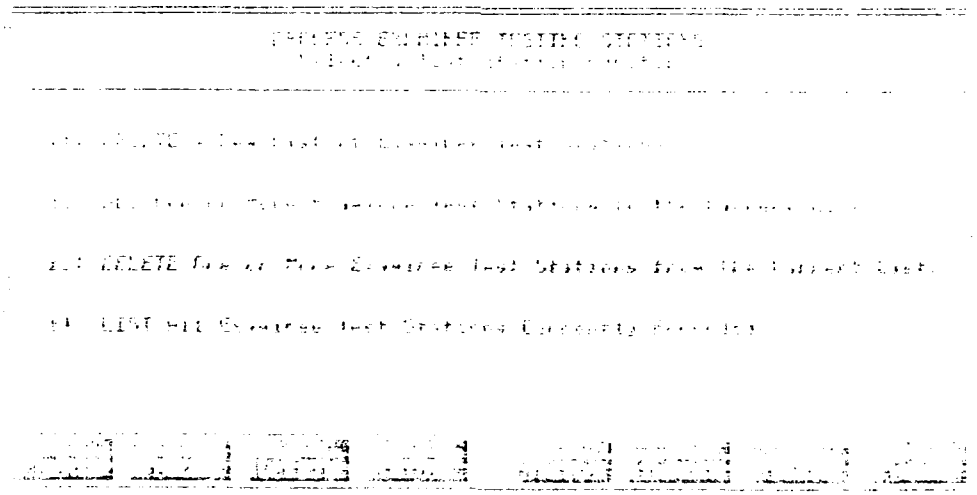


Figure 1: A typical TA Station software interactive screen

REFERENCES

Folchi, J. S. (1986). *Communication of computerized adaptive testing results in support of ACAP.* Paper presented at the 28th Annual Conference of the Military Testing Association.

Jones-James, G. (1986). *Design and development of the ACAP test administration software.* Paper presented at the 28th Annual Conference of the Military Testing Association.

Tiggle, R. & Rafacz, B. A. (1985, October). Evaluation of three local CAT-ASVAB network designs. *Proceedings of the 27th Annual Conference of the Military Testing Association.* San Diego, CA: Navy Personnel Research and Development Center.

DESIGN AND DEVELOPMENT OF THE ACAP TEST ADMINISTRATION SOFTWARE

Ms. Gloria Jones-James†

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

The Navy Personnel Research and Development Center (NAVPERSRANDCEN) is developing a Computerized Adaptive Testing (CAT) version of the paper-and-pencil Armed Services Vocational Aptitude Battery (P&P-ASVAB). As the lead laboratory for this joint service effort, NAVPERSRANDCEN is responsible for the research and development of CAT-ASVAB. Presently, NAVPERSRANDCEN is refining operational and functional concepts for the CAT-ASVAB project under the Accelerated CAT-ASVAB Program (ACAP). ACAP activities to date have involved the configuration of off-the-shelf computer equipment to be used for the delivery system, the development of examinee test administration software, and the development of Test Administrator (TA) Station software to support examinee testing.

The mission of ACAP is to automate CAT-ASVAB testing at selected Military Entrance Processing Stations (MEPS) and Mobile Examining Team (MET) sites within the United States Military Entrance Processing Command (USMEPCOM). The area of operation for ACAP is currently limited to two of the 68 MEPS sites and their associated METs. ACAP tests will be administered on the Hewlett Packard Integral PC (HP IPC). A description of the ACAP computer equipment is provided by Rafacz (1986). Test item banks to be used within ACAP were compiled from the Omnibus Item Pool and are divided into two unique forms. This is described by Wilbur (1986). ACAP data collection efforts are divided into three stages: a) Pre-Test, b) Score Equating/Provisional Operational Check (SE/POC), and c) Initial Operational Test and Evaluation (IOT&E).

ACAP has undergone a Pre-Test at the MEPS in San Diego, California. A CAT-ASVAB test was administered in the MEPS followed by a questionnaire or an interview on the examinee's perception of the test. ACAP tests were administered on a Local CAT Network (LCN) consisting of a Test Administrator (TA) Station networked with "n" Examinee Test (ET) Stations ($n \leq 30$), as described by Rafacz (1986). Data obtained from the Pre-Test will be used to evaluate ACAP's human-machine interactive software (i.e., the examinee's comprehension of test instructions, ease of use, etc.). Evaluation of the data obtained from the Pre-Test will suggest modifications to the ACAP software for the SE/POC stage.

Scope

This paper will address the status of ACAP as it pertains to examinee test administration software development. First, a brief description of ACAP requirements will be provided as follows: a) CAT-ASVAB functional specifications, as identified by Rafacz and Tiggle (1985); and b) Psychometric requirements incorporated in the current ET Station software. Second, an overview of the current ACAP networking and standalone software environments will be presented, followed by an overview of ET Station test administration software development.

ACAP Requirements

CAT-ASVAB Functional Specifications. The current ACAP efforts are being directed at the implementation of CAT-ASVAB functional requirements and the collection of psychometric data. For a detailed description of the functional requirements, see Rafacz and Tiggle (1985). A few noteworthy functional requirements incorporated into the ACAP system development are:

a. Portability. ACAP deployment system portability, a major requirement, is met through the implemention of the current software on the HP IPC microcomputer. The HP IPC is a 25-pound,

† The opinions expressed here are those of the author and do not necessarily represent those of the Department of the Navy.

transportable, self-contained, microcomputer system, described by Rafacz (1986).

b. Communications. The ACAP data communications link between the TA Station and ET Stations is satisfied by the LCN design and the HP IPC. The LCN design enables CAT-ASVAB tests to be administered individually to examinees stationed at any one of up to 30 ET Stations within the LCN environment.

c. Recovery. Recovery requirements for an ET Station during testing are supported in the ET Station design; i.e., a failure in the network or an ET Station will result in the use of a back-up mode of operation (i.e., the standalone mode of operation). The standalone mode of operation enables an examinee to resume testing at the beginning of the first non-completed test within the CAT-ASVAB test battery at the first available ET station. In addition, the standalone mode of operation can be used when only one or a few examinees are to be tested.

d. Security. The security requirements for the CAT-ASVAB Test Item Bank (TIB) are described in Design #3 from Rafacz and Tiggle (1985). Briefly, TIBs are deencrypted from the ACAP System Disc, stored at the TA Station in Random Access Memory (RAM) files and then downloaded to the ET Stations within the network. Once received by an ET Station, the test items are read into a string array for the purpose of random access during test administration. The ET Station software retains control of the HP system prohibiting user access to the data, except through the actual testing program. At the conclusion of testing, the HP is powered off and the RAM based data arrays (i.e., TIB data) are cleared from memory.

Psychometric Requirements. The psychometric data collection for ACAP will be used to support verification of CAT-ASVAB item parameters and equating CAT-ASVAB to P&P-ASVAB. Psychometric requirements implemented in the current ACAP software include, but are not limited to, the following:

a. All of the interactive screen dialogues for CAT-ASVAB test administration as defined by Rafacz and Moreno (1986).

b. Two test forms, in which a form is randomly assigned to an ET Station. Each form consists of all the necessary items to support nine power tests administered adaptively and two speeded tests administered sequentially.

c. Power test administration requirements (item selection and scoring) based on Owen's Bayesian technique and the infotable procedure.

d. The use of power test item exposure rate control (Sympson and Hetter, 1985).

e. Administration of seeded test items in a pre-determined order among "standard" power test items.

f. Administering a "group" of seeded tests containing items as described in "e" above. A seeded group will consist of from one to three sets of 20 power test items.

g. Speeded test administration of Numerical Operations (NO) and Coding Speed (CS) tests. A set of speeded tests for NO and CS are randomly selected from the three sets of NC and two sets of CS items and administered sequentially for each form.

h. Implementation of explicit (i.e., examinee presses the "HELP" key) and implicit (i.e., "HELP" generated via examinee errors or time-outs) "HELP" requests by the examinee, as described by Rafacz and Moreno (1986).

ACAP Modes of Operation

ACAP operating environments shall consist of the standalone (back-up) and networking (predominant operating environment) modes of operation. In order to incorporate many of the CAT-ASVAB functional requirements into ACAP, the design required that the software be divided into three programs: 1) the "Boot-up" program for hardware verification, 2) the "TA" program for the Test Administrator's functions, and 3) the "ET" program for the examinee test administration functions. The following discussion will summarize the interaction among these three programs as they relate to the standalone and networking modes of operation.

ACAP Standalone Mode. In this mode of operation, failure recovery procedures are activated in the event of software or hardware failures of the LCN or in the event that a TA or ET Station fails during testing. The major differences between the networking and the standalone modes of operation concern the TA Station's ability to transfer software and testing data, to transfer examinee information

[i.e., name and Social Security Account Number (SSAN)], and to monitor an examinee's progress during the test. In the standalone environment, the TA Station assigns examinees to an ET Station; the Test Administrator subsequently requests the printing of an assignment list. Prior to examinees entering the testing area, the Test Administrator would have manually assigned (by SSAN) the examinees to the appropriate ET Stations, as noted in the assignment list. In the networking mode, this would have been accomplished automatically as a session file is broadcast to the ET Stations.

In the networking environment, all data are expeditiously broadcast through the LCN, while in the standalone environment all data must be manually loaded into each ET Station. This distinction between the networking and standalone environments is resolved through ACAP System Discs (SD). All software and test items are loaded and the examinee's SSAN is entered at the ET Station keyboard in the standalone environment. Each ET Station in a standalone environment will be initialized using three of four discs; one ET Disc (one for each ET Station) and two of the three ACAP System Discs (identified as SD-A, SD-B, and SD-C). The ET Disc contains "Bootup" software, the "HP-IL Driver" networking software and certain graphic files. ACAP SD-A and SD-B contain form-specific test items and psychometric data (i.e., exposure control parameters and infotables, etc.) for Form 1 and Form 2 respectively. Additional seeded items, common data, and the ET Software are stored on SD-C.

Once an ET Disc is inserted into the ET Station disc drive and powered on, the "Bootup" software (in both networking and standalone modes) is activated. In the standalone mode, the Bootup program on the ET Disc randomly determines the form to be loaded onto the ET Station and stores the form number on a file. The "Bootup" program then instructs the Test Administrator to load ACAP SD-C and activates the ET Software. The ET Software loads the common data on ACAP SD-C and then locates the file in RAM containing the form number and instructs the Test Administrator to load the appropriate ACAP System Discs for the selected form. The test items are then read and decrypted from either ACAP SD-A or SD-B and loaded into arrays, as they are during the networking mode of operation. This procedure is repeated for each ET Station to be used during testing. After the testing data are loaded, the Test Administrator is instructed to enter the assigned examinee's SSAN at each ET Station, as identified on the TA Station printed assignment list. After the SSAN is entered the testing proceeds identically to testing in the networking mode of operation, with some exceptions. These exceptions include: a) the monitoring function at the TA Station and b) automated transfer of examinee testing data at the conclusion of testing to the TA Station.

ACAP Networking. The networking mode of operation facilitates the broadcasting of ET Station software and testing data from the TA Station to the ET Stations in the LCN via the HP-Interface Loop (HP-IL). Data are broadcast over the HP-IL loop from the TA Station to each station in the LCN loop and returned to its source (i.e., TA Station). Each ET Station within the LCN is uniquely addressable which allows the data to traverse the loop and only be retained by the target ET Station(s). Other ET Stations, not identified by the prerequisite addresses will simply pass the data on to the next station in the loop. After the data are returned to the TA Station, they are checked for errors. If, after several unsuccessful broadcasts, errors are still detected, the LCN loop is terminated. For purposes of ACAP, a failure of the LCN will be recoverable through the standalone mode of operation as described above.

After the TA and ET Stations are configured into a LCN, as described by Rafacz (1986), the TA Station waits to receive the set of ET Station ID Numbers for all ET Stations in the LCN. For all ET Stations, a CAT-ASVAB test form (i.e., Form 1 or Form 2) is randomly assigned to the ET Station IDs. Following the randomization of form to ET Station IDs, all ET Stations are assigned as "listeners" and broadcasted their respective form number and ET Software. Upon arrival at the ET Stations, the ET Software is executed after which all ET Stations are assigned as listeners and the common data is broadcast. Following the receipt and storage of all common data in RAM, the ET Station waits for the TA Station to broadcast the form-assigned testing data. The TA Station assigns "listener" status to all ET Stations assigned to a specific form and proceeds to broadcast the form data. After all ET Stations for one form have successfully stored their data they are set to "non-listener" and the alternate form specific data are broadcast to the other ET Stations now assigned as "listeners". The session file (containing examinee identifiers and station assignments) is then broadcast to all ET Stations and the TA Station subsequently monitors the ET Stations in the LCN as described by Rafacz (1986).

In the preceding scenario, the ET Station initial Bootup program loads and activates the HP-IL Networking software. This software waits to receive the ET Station (CAT-ASVAB test administration) software from the TA Station. After the TA Station has broadcast this software to all ET Stations in the LCN, the Bootup program activates the software and terminates. Test administration software, once activated, signals the TA Station, through the HP-IL driver, that it is ready to receive the test item data. The TA Station initially broadcasts a "Rosetta" file listing the names of the files that the ET Station is to receive. Each ET Station then waits to receive the data (i.e., common data and form-specific TIBs) in the station input buffer. The broadcast data are received at the ET Station as follows: Common data items are received and loaded into RAM disc files and retained in RAM, followed by the TIBs that are loaded into ET Software dynamic array area. The session file that was received at the conclusion of the broadcast is then accessed. The examinee to be tested is identified and the CAT-ASVAB test is administered.

For the interested reader, Rafacz (1986) provides details on the duties of the Test Administrator, and how some of these functions have been automated on the TA Station for purposes of ACAP.

Test Administration Software

As stated previously, the goal of the ACAP software development is to implement CAT-ASVAB functional and psychometric requirements in accordance with the above standalone and networking modes of operation. To accomplish this goal, the ACAP software development used a top-down design structure in which the major functions performed by TA and ET Software were broken into modules. The modules are subordinates to the initial TA and ET programs. These modules have further subordinate modules or functions which perform individual tasks. This modular division of the TA and ET functions enhances the flexibility, performance, efficiency, readability, and maintainability of the software. The following section will describe ACAP software with special emphasis given to ET Software development.

ACAP Software. Software for both the TA Station and ET Station software is being developed on the HP IPC, utilizing the HP UNIX (HP-UX) System 5.0 operating system, based on the AT&T UNIX™† operating system. The computer programming language selected for software development within ACAP is 'C'. As a mid-level language, C lies between machine languages and high level languages such as assembler and Pascal, respectively. One of the many characteristics of 'C' which greatly aids in software development is the ability to provide easy access to memory, allowing the programmer the versatility to manipulate data structures and control processing within the CPU. Another feature of 'C' is the use of global and local variable declarations. This feature enables the present design modules to share global or common data while protecting their own local internal data.

ACAP Test Administration software is composed of TA and ET Station software as previously described. The TA Station software facilitates the following functions: LCN configuration, random assignment of forms to ET Stations, downloading data through the network, assignment of examinees to ET Stations, and monitoring examinees at the ET Stations (see Rafacz, 1986 for further details). Functions expedited by the ET Station software are: the acceptance of test data, identifying examinees, administration of CAT-ASVAB, and the transfer of examinee test results to the TA Station. The following sections will briefly describe the ACAP software and current status of the ET Station software.

Test Administration Software Development. ACAP's Test Administration software (identified as "ET Software") has been divided into two programs, "Bootup" and "ET Software". The ACAP modular design enables all functions for standalone and networking mode of operations to be divided into distinct modules. The modular design is illustrated in the initial Bootup program. In the Bootup program, the ET Station hardware is verified and the mode of operation is identified. Once the standalone or networking environment is determined, the Bootup program proceeds to call the appropriate module to load the ET software from either the network (via the HP-IL) or directly from the ACAP System Discs.

†UNIX™ is a trademark of AT&T Bell Laboratories.

The same modular approach is applied in the actual test administration ET software. In the ET software, the Test Item Banks (TIBs) are loaded, the examinee to be tested identified, and the test administered. In the ET software, the networking and standalone environments vary in the loading of the TIBs and identifying the examinee. If the mode of operation identified is networking, TIBs and examinee data are received via the HP-IL as described above. After the Test Administrator identifies the operating environment, the ET software simply switches to the appropriate module to load the TIBs and the examinee data. In either case, the data obtained are stored in the same manner and the appropriate test administration software modules are activated.

The ET software modular design enables the actual administration of the CAT-ASVAB test to remain constant across examinees and not be contingent on the LCN operational environment. Testing begins after the examinee verifies the SSAN (as recorded in the session file) and listens to a verbal description of the CAT-ASVAB testing program by the Test Administrator. Upon commencement of the test, the examinee is first given an orientation session covering the use of the computer with an Examinee Input Device (EID), followed by a brief practice session of the type of test items that will be administered. At this time, the ET software searches through a "test vector" describing the tests and obtains testing related parameters (i.e., test type, time limit for a test, number of items per test, etc.). For each test, the examinee is initially administered a familiarization and practice session in which the test is described and the examinee is given a practice item. After the practice session, the actual test is administered. This process, starting with the practice session, is repeated for each test identified in the test vector. At the completion of all tests, the ET software instructs the examinee to raise his/her hand for assistance. The Test Administrator then excuses the examinee from the testing area and transfers the examinee's test data to the TA Station for further processing.

In the scenario above, the ET software background processing performs many functions. Before a power test item is administered, it is first identified as adaptive or seeded. For an identified test item, the appropriate software modules are called to select an item, display the item, record the examinee response for the item, and finally, score the examinee's response. These modules are called until all items for a specific test are administered or the test time limit is expired.

Power test administration was easily divisible into unique modules. A "Select" module determines the type of item to be administered from the test "allocation vector" (i.e., a file containing 0's and 1's where 0 = a seeded test item and 1 = an adaptive item). When the item type is equal to adaptive, the "Select" module activates subsequent modules to access the "infotable" for the new ability level (initialized to 0 at the beginning of each test), generate a random number, determine if the item was not previously looked at, verify exposure rate, and display the item, if it meets the criteria.

NOTE: The criteria for display are as follows: If the random number (between 0 and 1) generated is greater than the item's exposure control value, that item is not to be displayed and the item is flagged to no longer be considered for display during the test. If the random number is less than or equal to the item's exposure control value, that item is considered to be a viable item for display. In the latter case, the item will be displayed if it was neither previously displayed nor considered for display. Finally, if the item is rejected (for some reason), the next item (for the current ability estimate) will be considered for display by the aforementioned criteria.

A power test item is scored using the Owen's Bayesian technique (Owen, 1969; 1975). Subsequently, once the examinee has answered the item with a valid response, a new ability estimate is generated for selection of the next test item. Processing continues with the selection of an item for the new ability estimate as described above. After an item meets the criteria for administration, the "Display Item" module is called. Once an item is displayed, a timing function is triggered. If the examinee does not respond to an item in a pre-determined time, an error message is displayed and the Test Administrator is called to assist the examinee.

When the item to be administered is a seeded item, the "Select" item module calls subsequent modules to sequentially retrieve a seeded item, display the item, and record the examinee's response, as described above with minor variations. For the seeded items, the "infotable", random number, and exposure control modules are not used. In addition, seeded items are scored as correct or incorrect, rather than using the Owen's Bayesian technique.

Future Software Development

The future plans for ACAP software include the modification of the current software to enhance testing procedures, implementing additional psychometric requirements, enhancing networking functions for monitoring capability, automating the HELP function, and moving testing data from the ET Stations to the TA Station. In addition, in the SE/POC stage, the Data Handling Computer (DHC) will be used to transmit data to USMEPCOM, as described by Folchi (1986).

REFERENCES

Folchi J. (1986). *Communication of Computerized Adaptive Testing Results in the U.S. Military.* Paper presented at the 28th Military Testing Association (November, 1986).

Owen, R. J. (1969). *A Bayesian approach to tailored testing.* (Research Bulletin 69-92). Princeton, New Jersey. Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of American Statistical Association, 70,* 351-356.

Rafacz, B. A. (1986). *Development of the Test Administrator's Station in Support of ACAP.* Paper presented at the 28th Military Testing Association (November, 1986).

Rafacz, B. A. and K. Moreno. (1986). *Accelerated CAT-ASVAB Program (ACAP) Development System Interactive Screen Dialogues for the Examinee Testing (ET) Station.* A NAVPERSRANDCEN, Code 63, unpublished manuscript (August, 1986).

Rafacz, B. A. and R. B. Tiggle. (1985). *Functional Requirements for the Accelerated CAT-ASVAB Program (ACAP) Development System.* A NAVPERSRANDCEN, Code 63 unpublished manuscript (November, 1985).

Sympson, J. B. and R. Hetter. (1985). *Controlling Item Exposure Rate In Computerized Adaptive Testing.* Proceedings of the 27th Military Testing Association (November, 1985).

Wilbur, E. R. (1986). *Design and Development of the ACAP Test Item Data Base.* Paper presented at the 28th Military Testing Association (November, 1986).

Communication of Computerized Adaptive Testing Results in Support of ACAP

John S. Folchi†

Computerized Testing Systems Department
Manpower and Personnel Laboratory
Navy Personnel Research and Development Center
San Diego, California 92152-6800

1. Introduction

The Navy Personnel Research and Development Center (NAVPERSRANDCEN) is currently involved in the development of a Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). The goal of the effort is deployment of a CAT-ASVAB system on a nationwide distributed processing computer network. When fully deployed, CAT-ASVAB would be used by the United States Military Entrance Processing Command (USMEPCOM) to select and classify enlisted service personnel.

Prior to deployment of a Full Scale Development (FSD) version of the CAT-ASVAB system, NAVPERSRANDCEN will refine the operational concepts and functional specifications of the CAT-ASVAB system through the Accelerated CAT-ASVAB Program (ACAP). The goal of ACAP is deployment of CAT-ASVAB at two of the Military Entrance Processing Stations (MEPS), and at all of the Mobile Examining Team (MET) sites under the jurisdiction of the selected MEPS. The hardware and software deployed during ACAP will satisfy most of the anticipated functional specifications of the FSD CAT-ASVAB system. Tiggle and Rafacz (1985) describe the computer system selected for purposes of ACAP.

This paper will discuss the role of the Data Handling Computer (DHC) during the Initial Operational Test and Evaluation (IOT&E) phase of ACAP and the extension of the ACAP DHC so it will meet the functional requirements of an FSD CAT-ASVAB system. Figure 1 depicts the ACAP communications network within which the DHC will operate. As shown there, a DHC is situated at USMEPCOM Headquarters (HQ) and at each of the MEPS involved in ACAP. The MEPS DHC will coordinate the transfer of ACAP data from the MEPS/MET testing sites to USMEPCOM's System 80 minicomputer located at the MEPS and to the DHC at USMEPCOM HQ. The DHC at USMEPCOM HQ will coordinate the flow of ACAP data from each of the MEPS DHC's to consumers at USMEPCOM HQ and the CAT-ASVAB Maintenance and Psychometric (CAMP) Facility at NAVPERSRANDCEN.

2. DHC Functional Specifications for ACAP

The most important functions of the DHC are data collection and data distribution:

(1) Data Collection - The DHC must collect the data that are generated at the testing sites and organize them so that they may be conveniently distributed to the users. After collecting and organizing the data, the DHC must store them until distribution takes place.

(2) Data Distribution - The DHC must distribute the data to consumers at the MEPS, USMEPCOM HQ, and the CAMP Facility. The DHC must separate data designated for USMEPCOM's operational requirements from research data designated for the CAMP Facility. The CAMP Facility must be provided with research data to evaluate the psychometric aspects of CAT-ASVAB. During ACAP, the DHC will send these data to the CAMP Facility via the USMEPCOM DHC. ACAP examinee scores will "count" for selection and classification purposes during IOT&E.

† The opinions expressed in this paper are those of the author and do not necessarily represent those of the Department of the Navy.

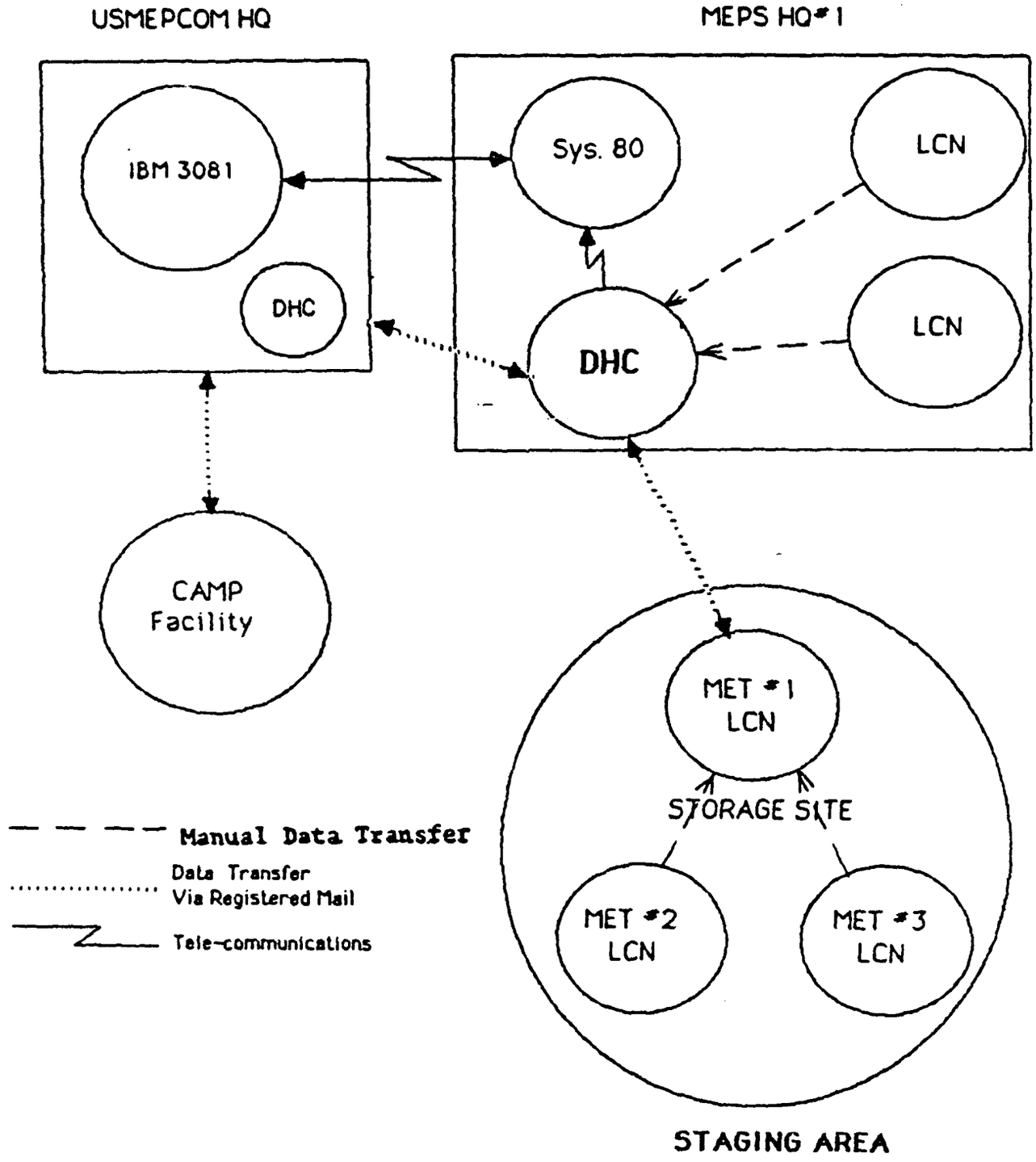


Figure 1: PROPOSED ACAP NETWORK FOR IOT&E

Therefore, USMEPCOM HQ must be provided examinee scores and other personal data needed for its operational requirements. During ACAP, the DHC will send these data to USMEPCOM HQ via the MEPS System 80. All data must be distributed in a timely manner, via communications media (e.g. magnetic tape, floppy disc, telecommunications, etc.) that can be read and processed by the consumers.

The DHC plays the role of what can be best described as a data controller or a data manager. It must insure that all examinee data are collected exactly once, to avoid the accumulation and distribution of redundant data. Similarly, the DHC must insure that the data are distributed exactly one time to each consumer, unless a consumer specifically requests that a particular data set be sent more than once.

3. DHC Hardware Configuration

This section describes the hardware of the Data Handling Computer:

(a) One Hewlett-Packard Integral Personal Computer (HPIPC) with 512 kilobytes (KB) of internal Random Access Memory (RAM) and one internal 710 KB microdisc drive. All examinee data collected at the MEPS/MET test sites will be moved to the DHC by means of 3.5" floppy discs, hereafter referred to as DataDiscs. Selection of the HPIPC insures that the DHC will be able to process the DataDiscs generated by the HPIPC hardware employed at the testing sites. Selection of the HPIPC also facilitates recovery from hardware failure, as will be discussed further in Section 5.

(b) Approximately 3 megabytes (MB) of CPU addressable external RAM in a RAM expansion box.

(c) One tape drive capable of processing 67 MB tape cartridges.

(d) Two 55 MB hard disks. Once such unit, hereafter referred to as the master disk, will hold all DHC software and data files. The other disk, hereafter referred to as the secondary disk, will be used only in the event the master disk fails. The DHC must be able to store all examinee data from the time they are received from a testing site until the consumers have verified that their data have been received from the DHC. A 55 MB disk will provide sufficient capacity for all examinee data on the DHC awaiting distribution to consumers and/or verification of receipt.

(e) One HP serial interface board for asychronous communications with the System 80.

4. Data Collection

As described by Rafacz (1986), the Test Administrator (TA) will create a DataDisc at the completion of each ACAP test session which contains all examinee data generated during the session. The DataDisc will be sent by registered mail to the parent MEPS from the test site.

Once the DataDisc has arrived at the MEPS, the DHC operator will initiate the Data Prep module. The Data Prep module will read the disc, separate USMEPCOM's operational data from the CAMP Facility's research data, and store all data on the DHC. As shown in Figure 2, the operational data are written to the COM file and the Operational Data Archive (ODA). The COM file contains all operational data on the DHC awaiting transmission to the System 80 and the ODA is an archive of all operational data received by the DHC in the last 30 days. The Data Prep module also copies the research data to the CAMP file, which contains all research data on the DHC awaiting transmission to USMEPCOM HQ.

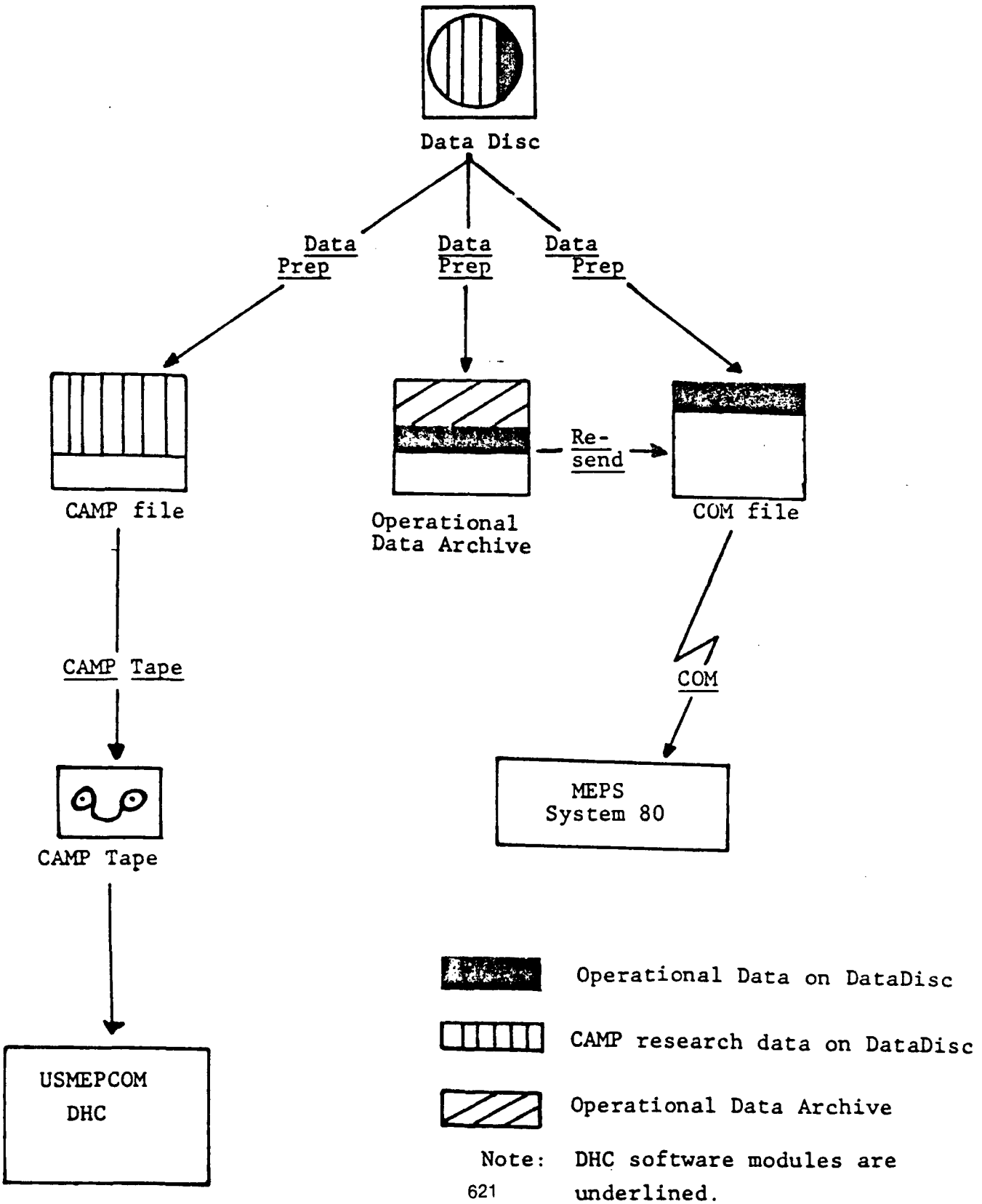
Once the Data Prep module has completed its work, the operational data and the research data are ready for distribution to the System 80 and the CAMP Facility respectively. The Data Prep module will have built-in checks to prevent any DataDisc from accidently being processed more than once, thus minimizing the accumulation of redundant data on the DHC.

5. Data Distribution

The COM module distributes the operational data in the COM file to the System 80 through an RS-232 communications link. The volume of testing and USMEPCOM's operational requirements are such that the operational data must be telecommunicated on demand. The COM software will have an

FIGURE 2

MEPS DHC Data Processing



error detection/correction algorithm that will insure all data are correctly transmitted to the System 80.

The Resend module will facilitate re-transmission of operational data that have already been sent to the System 80. The Resend module gives the DHC the capability to restore any data that have been sent to the System 80 but have subsequently been lost. All data designated by the DHC operator for re-transmission to the System 80 will be extracted from the ODA and written to the COM file. As described in the preceding paragraph, the COM module may then be used to transmit the COM file to the System 80.

To facilitate transfer of ACAP research data to the CAMP Facility and USMEPCOM HQ, the CAMP Tape module will consolidate all examinee data collected at the MEPS/MET testing sites onto a single tape cartridge. This tape, hereafter referred to as a CAMP Tape, may be generated upon demand. Once created, the tape cartridge will be sent to the DHC at USMEPCOM HQ. The USMEPCOM HQ DHC will consolidate all CAMP Tapes received from the various MEPS onto a single CAMP Tape and send that consolidated tape to the CAMP Facility. Using a procedure yet to be determined, the CAMP Facility will verify the integrity of the data on the consolidated tape and inform USMEPCOM HQ and the MEPS accordingly. A DHC operator will be permitted to delete the corresponding CAMP data file from the DHC only after being informed that the tape was verified.

6. DHC Failure Recovery

Because the DHC furnishes information necessary for urgent USMEPCOM operational requirements, the DHC must provide a simple, quick, and reliable means of restoring itself to operation in the unlikely event of hardware failure. During ACAP, failure recovery will be facilitated because much DHC hardware is completely interchangeable with that being employed for the Examinee Testing (ET) Station and Test Administrator's (TA) Station located at the MEPS testing site. [Please refer to Rafacz (1986) for more information on the TA Station and to Jones-James (1986) for more information on the ET Station]. Consequently, in many instances, failure recovery will be accomplished by simply borrowing ET and/or TA Station hardware from the MEPS testing area located in the same building as the DHC. Such interchangeability will minimize the cost and the amount of space required to store back-up hardware for ACAP.

Since neither the ET nor the TA Stations are equipped with a hard disk, the DHC operator cannot recover from a hard disk failure by substituting TA or ET Station hardware. Consequently, an identical secondary disk will be available at each MEPS as a back-up. Periodically, the DHC operator will back-up the contents of the DHC master disk onto a magnetic tape cartridge. This will protect the DHC from the loss of all data on the master disk. If the master disk should subsequently fail, the operator will recover by copying the back-up tape onto the secondary disk and then repeating all DHC activity since the back-up tape was created. At that point, the operator can resume normal DHC operations using the secondary disk as the new master disk.

7. Possible Extensions to Accomodate FSD CAT-ASVAB

This section will briefly discuss possible extensions which may be necessary in order for the DHC to satisfy the functional requirements of an FSD CAT-ASVAB system.

A critical question regarding the configuration of the FSD DHC concerns whether or not telecommunications will play a role in the movement of examinee data from the test sites to the DHC and from the DHC to the consumers. Telecommunication has the potential to deliver examinee data to the consumer much more rapidly than is possible under either current paper and pencil ASVAB or ACAP. The availability of a telecommunications link between the test sites and the DHC will enable these data to be transmitted to the DHC in a matter of minutes rather than days. Similarly, minutes rather than days would be required to transmit the data from the DHC to individual consumers. However, telecommunications will entail considerable costs. Should policy-makers decide that telecommunications will play a role in the movement of examinee data during FSD CAT-ASVAB, then an FSD communications system centered around the DHC will have to be built.

The FSD communications system will be required to satisfy strict functional specifications that are not required of the ACAP system. Because the Department of Defense (DoD) has recently

mandated that all future DoD data communications systems must subscribe to the Defense Data Network (DDN), all telecommunications in the FSD CAT-ASVAB system will probably be done through the DDN. In addition, Army Regulation 530-2, which requires encryption of all information passed between computer systems utilizing external communications systems, will apply to the FSD system.

All ACAP hardware that will interface with the DDN will probably have to be upgraded, inasmuch as computer systems that currently interface with the DDN are either mainframe systems or front-end processors for mainframes, while the current ACAP hardware consists of single-user microcomputer systems. In addition, the FSD DHC will also be required to have the security features of a multi-user system, in order for it to act as a host computer telecommunicating with multiple test sites.

The FSD communications software must also meet stricter standards than are being imposed upon ACAP. The protocols used by the FSD communications system must be compatible with the protocols used in the DDN. In addition, the software must allow the DHC to act as a host computer capable of telecommunicating with more than one testing site at any given time. However, such software already exists. NAVPERSRANDCEN recently evaluated the feasibility of adapting a currently-available electronic mail software package to CAT-ASVAB telecommunications requirements (see PURVIS Systems, 1986). The software evaluated was RMAIL, the electronic mail package currently being used by the US Navy Recruiting Command. The study concluded that such an adaptation was feasible and that RMAIL can satisfy the functional requirements of an FSD CAT-ASVAB system, including encryption and compatibility with DDN.

REFERENCES

Jones-James, Gloria (October 1986). *Design and Development of the ACAP Test Administration Software*. Paper presented at the 28th Annual Conference of the Military Testing Association

PURVIS Systems (March 1986). *Technical Assessment of the Software and Hardware Performance Characteristics of a Candidate CAT-ASVAB Telecommunications Capability*

Rafacz, Bernard A. (October 1986). *Development of the Test Administrator's Station in Support of ACAP*. Paper presented at the 28th Annual Conference of the Military Testing Association

Tiggle, Ronald and Rafacz, Bernard A. (November 1985). *Functional Requirements for the Accelerated CAT-ASVAB Program (ACAP) Development System*. An unpublished NAVPERSRANDCEN, Code 63, document.

ROTC CADET SUBPOPULATIONS: SCHOLARSHIP OR NOT

Allyn Hertzbach and Timothy W. Elig
U.S. Army Research Institute for the Behavioral and
Social Sciences Alexandria, Virginia 22333-5600

The Reserve Officer's Training Corps (ROTC) has been providing capable officers for the US Army since the Civil War. The US Military Academy (USMA) provided the majority of officers for the small peacetime Army maintained during the latter 19th century. But after the experiences of World Wars I and II, a larger standing peacetime Army required more officers than could be provided by the USMA, so that other methods for providing professional officers needed to be institutionalized. ROTC and the Officer Candidate School (OCS) provide these addtional officers. Currently, ROTC provides 75% of the Army officers that are commissioned each year, and the program has changed dramatically in the past decade. The purpose of this paper is to recommend a shift in the focus of cadet acquisition and retention research that recognizes two distinctive Army ROTC subpopulations: scholarship and nonscholarship cadets.

Snyder (1984) suggests that for all of the services, including the Army, ROTC programs were profoundly changed by the Viet Nam experience and the move toward an all volunteer force. The recruiting environment facing Army ROTC during the difficult days of the Viet Nam era necessitated a policy of increasing the number of scholarships to attract able college students. Previously, many colleges had required ROTC of its male students for one or two years, and this practice had given the ROTC unit on campus a chance to recruit successfully without requiring additional incentives, such as scholarships. A military career was an attractive career choice for students, so that the opportunity to become an Army officer itself encouraged participation.

The poor recruiting environment of the Viet Nam era appears to have abated, and a more favorable view of the military and the high cost of college have had the combined effect of encouraging much interest in winning an Army ROTC sholarship. Competition is keen, and there is little problem in enrolling male and female students with excellent academic backgrounds and good character and health. In fact, the ROTC scholarship student is very similar to the USMA cadets in the criterion measures used to evaluate cadet quality, such as SAT score, high school grades, leadership positions, and other honors (Snyder, 1986; Hertzbach, Gade, and Johnson, 1985). Currently, the Army has about 12,000 two, three, and four year ROTC scholarships authorized.

From the point of view of officer recruiting, the most important changes brought about by the increased use of scholarships in ROTC have been that: the program is more closely managed; cadets who will provide career service are sought; students pursuing scientific/technical educations are sought; and the term of obligated service is lengthened (Snyder, 1984). The last three changes are significant ones for the Army in that its ROTC officers are likely to be more competitive with USMA graduates and to serve longer tours than their

---

1 The views expressed in this paper are those of the authors and do not necessarily reflect the views of the US Army Research Institiute or the Department of the Army.

initial obligation, asssuming that the scholarship goal of cadet career service is met. But for the purposes of this paper, the implications of the scholarship program that are important are the ones that separate the ROTC scholarship student from the traditional, nonscholarship cadet.

Despite the increased reliance on scholarships ROTC is mainly a nonscholarship program (Table 1). Looking at the opening enrollment figures for the Army ROTC program for academic year 1985/86 illustrates this point, see Table 1. Though many of the Military Science 1 (MS1) nonscholarship students (mostly freshmen) attrite well before contracting (at the end of MS2), a sizeable portion of contracted cadets (MS3 and MS4), 53%, are nonscholarship. (Generally, military science level, 1 through 4, corresponds with the cadet's year of college, though not always.)

Table 1

Army ROTC Opening Enrollment for 1985/86, Military Science (MS) 1 - 4 by Scholarship Status

	Scholarship		Nonscholarship		Total	
ROTC/College Year	freq.[a]	%	freq.	%	Freq.	%
MS1 (Fresh.)	1,244	4	29,786	96	31,030	100
MS2 (Soph.)	2,101	17	10,306	83	12,407	100
MS3 (Jr.)	4,602	51	4,352	49	8,954	100
MS4 (Sr.)	3,911	43	5,182	57	9,093	100
Totals	11,858	19	49,626	81	61,484	100

[a] The number of scholarships rises because of the award of two and three year sholarships to cadets who entered college (and ROTC) without scholarships.)

There have been a number of efforts in the area of ROTC recruiting in which cadets and students have been surveyed and similarities and differences discussed (Kimmel 1985; Armstrong, Farrell, and Card 1979; Shaw, Graf, Davis, and Hertzbach 1984; and Hertzbach et al. 1985). In reviewing these efforts, I find a wealth of interesting and useful information, but there was little emphasis on the importance of scholarship status. There was much less reason to make this point before the scholarship program became as large a force in ROTC. Nor can researchers be faulted for ignoring this characteristic. Usually, their aims were unrelated to scholarship status, and the full effect

of the change in ROTC was not clear enough or directly a part of the researcher's experience. But the more familiar one becomes with survey responses of ROTC cadets, the more one begins to see that scholarship cadets do not have the same reasons for enrolling, career intentions, background and demogrphics as nonscholarship cadets.

Snyder (1986) sensibly suggests that the nonscholarship cadet is less talented than is the scholarship (or USMA) cadet and resembles the pre-Vietnam ROTC graduate. These cadets are drawn from the general college population which does not have as high achievement test scores or grades as can the scholarship cadets, which means that they probably reflect a greater range of abilities and potential, not that the nonscholarship group is without capable individuals. Many distinguished careers have been served in the Active and Reserve Army components by the pre-Vietnam ROTC officer. However, when considering how to identify and encourage nonscholarship students to participate in ROTC, the same strategies and approaches used for scholarship participation are obviously inappropriate. For one thing, the scholarship student usually must be aware of the opportunity well before entering college. The nonscholarship cadet need not know about the program until being on the college campus, though being able to establish an awareness of Army ROTC earlier is a worthy recruiting goal.

In order to illustrate the reasons that researchers need to be careful about discriminating between scholarship and nonscholarship cadets, three areas of concern are discussed that contribute to formulating effective officer acquisition and retention policy: ROTC scholarship officer attrition, modeling the ROTC enrollment decision, and specific market information.

As one of the main goals of utilizing ROTC scholarships is to provide higly capable career officers from ROTC, the attrition rates of ROTC scholarship officers need to be carefully studied and compared to USMA graduates and nonscholarship ROTC graduates. Recently, I discovered some unpublished data that suggest that resignations are relatively higher among ROTC scholarship holders than among USMA officers after their first obligated tour of service is completed. These data are sketchy; however, the resignation trends need to be clarified so that policy makers can determine if their goals for the scholarship program are being realized. In short, having the resignation information available by ROTC scholarship and nonscholarship status is imperative for tracking salient differences. If ROTC attrition data are lumped together, the averages could very well conceal important trends and information about ROTC scholarship and nonscholarship resignation rates.

Another more complicated issue that requires separation of ROTC scholarship cadets and nonscholarship cadets is found in the research aimed at the recruiting and marketing of ROTC. I have used a decision step model for conceptualizing the ROTC enrollment process. This model consists of enumerating the decisions points of ROTC enrollment and contracting. What quickly becomes apparent is that the progression of decisions is very different for scholarship cadets than for nonscholarship cadets. The decisions for enrollment and contracting of the scholarship cadet are made in high school or early in college and are qualitatively different than considerations of the

nonscholarship cadet, that is they are committing themselves for the next eight years to the Army, first parttime and then, in most cases, full-time, for a substantial financial incentive. The nonscholarship cadet can delay the commitment much longer, until midway through his college career; and there is much less financial incentive for participating in the ROTC program.

In building a meaningful model of nonscholarship ROTC enrollment, we have been concerned with reasons for enrolling, background information, career intentions and interests, attitudes towards the ROTC program, faculty, peers, and the effects of various influences, such as parents, friends, and advertising. Not only do these two subpopulations (scholarship and nonscholarship cadets) vary on many of these issues, but from a marketing point of view there is less need to know much of this market relevant/recruiting information for the scholarship cadet, though there are certainly uses for some of this information in evaluating the ROTC program. The Army ROTC scholarship program is getting more than enough high quality applications and acceptances. What we need to know is how to encourage students into the nonscholarship program who have the potential to become able Army officers in the Active and Reserve components.

Research findings indicate that the two subpopulations of cadets are different (Hertzbach, et al, 1985), and the differences are important in understanding the policies required for the acquisition and retention of scholarship and nonscholarship cadets. In fact, the four year scholarship cadet is sometimes different than the three and two year cadet, though as the sample of two year cadets was too small to be certain of the stability of differences, only three and four year comparisons were made. The sample of Military Science 2 (MS2) cadets who responded to the 1984 ARI/ROTC Demographic Survey are used for making scholarship/nonscholarship comparisons. The sample of MS2 cadets was composed of 540 nonscholarship cadets, 149 three year scholarship cadets, and 88 four year scholarship cadets from ROTC detachments across the country. (The scholarship and contracting status of these cadets was obtained the academic year following the survey data collection.) All of the differences reported are significant ($p<.03$), and the the chi square statistical test was used to calculate significance levels.

An example of the kind of difference between scholarship cadets and nonscholarship cadets that is important to marketing research is the distance between the home (parental) and the college attended. If the pool of potential cadets live relatively close to home, then the campus ROTC detachment can begin recruiting before the student arrives at college by visiting area high schools or other civic centers. Such a program has, in fact, been instituted because of the finding that more than half of sampled nonscholarship cadets (52%) lived within 100 miles of the college that they attended. Had this information not been segmented into nonscholarship and scholarship groupings, the proportions might have been masked because scholarship students are more likely to attend schools farther from home than does the rest of the sample of cadets. More than 40% of three and four year scholarship cadets attended a college more than 200 miles from their homes as compared to barely 30% of the nonscholarship cadets sampled. Only 29% of four year scholarship cadets attended college within 100 miles of their homes.

Other differences between the two groups of cadets concern their educational backgrounds. For example, scholarship cadets: Were far more likely to have been in college preparatory curricula than were nonscholarship cadets (84% to 63%, respectively); Were more likely to have better than a B average in high school (more than 90% as compared to 72%, respectively); Were more likely to have participated in a varsity sport in high school (82% to 68%, respectively); Were more likely to hold student government office in high school (39% to 28%, respectively); And were more likely to be members of the Honor Society in high school (more than 50% to 27%, respectively). Averaging these groups together would most likely misrepresent both of them. Though these differences seem unsurprising, they need to be tracked so that relevant information can be provided regularly to policy makers.

Another argument based upon observed differences between the nonscholarship and scholarship cadets concerns the analysis of responses to uncover more subtle relationships between the two groups. The researcher is often confronted with a finding that might have a misleading impact, such as the fact that scholarship cadets report that they do not regularly watch television (43% to 32%, respectively). If this difference were to be literally and straightforwardly interpreted, the Army might discourage television advertising, thinking that the more able audience is not watching enough television to justify the expense of ROTC television advertising. But a more careful analysis of televsion watching behavior reveals some inconsistency in scholarship cadets' reports of this behavior. They report similar rates of watching televised sports, as well as similar rates of watching such programming as situation comedies, network and local news, movies, late night programming, and drama as do nonscholarship cadets. If policy makers wanted to test this approach, they could focus the ads on the market using the scholarship/ nonscholarship information for greater precision and success.

One last set of characteristics that vary for scholarship and nonscholarship cadets is sources of influence. Four year scholarship cadets are far more likely to be made aware of ROTC in high shool, since they apply for the scholarship at that time (97% to 69%, respectively). All scholarship cadets are also more likely to be made aware of ROTC by their families than are nonscholarship cadets (60% to 38%, respectively), and the parents of scholarship cadets have a less neutral or negative view of an Army officer's career than do nonscholarship cadets'parents (12% to 22%, respectively). Nonscholarship cadets and three year scholarship cadets are more likely to discover Army ROTC after arriving on the college campus (21%) than the four year scholarship cadets (1%). Policy makers must consider how and when awareness occurs to plan effective strategies for recruitng, and this effort must be done with information that takes account of the nonscholarship and scholarship cadet differences. Researchers must provide policy makers with clarifying information that is tailored to the decisions that need to be made. For Army ROTC policies, that information must include scholarship and nonscholarship information.

Conclusion

The Army ROTC researcher needs to be aware of the importance of differences between the scholarship and nonscholarship cadet. He or she also needs to be aware of other demographic differences, but the scholarship or nonscholarship distinction is essential for all ROTC research. Within the scholarship category, no assumptions of similarity should be made between three and four year scholarship holders. Sometimes the three year scholarship cadet is more similar to the nonscholarship cadet, e.g., initial ROTC awareness, and other times that cadet is more like the four year cadet, e.g., academic and athletic background.

The Army ROTC program is composed of at least two different subpopulations. The scholarship student, particularly the four year scholarship cadet, is selected from an elite group of the nation's young people. He or she will often behave differently than the nonscholarship cadet. For recruiting purposes, as well as for retention purposes, policy makers should always ask themselves before promulgating new policy about the impact of that policy on each of these components of the cadet population. Both components are essential in order for the Army to successfully meet its expanded peacetime mission, as well as its main mission of being prepared for war.

References

Armstrong, T. R., Farrell, W. S., and Card, J. J. (1979). Subgroup differences in military-related perceptions and attitudes: implications for ROTC recruitment. (Research Report 1214) Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Hertzbach, A., Gade, P. A., and Johnson, R. J. (1985). The 1984 ARI ROTC cadet decision making survey: an overview (WP85-6). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Kimmel, M. J. (1985). Attitudes, preferences and career intentions of ROTC and non-ROTC students Proceedings of the 27th Annual Military Testing Association Conference, I, 188-193.

Shaw, C. D., Graf, J. G., Davis, R. L., and Hertzbach, A. (1984). The 1983 demographic profile study of reserve officers' training corps advanced corps cadets (DCSPER/TRADOC joint research project). Fort Monroe, VA: Deputy Chief of Staff for ROTC.

Snyder, W. P. (1984). Officer recruitment for the all-volunteer force: trends and prospects Armed Forces & Society. Vol 10, No. 3, Spring 1984.

Snyder, W. P. (1986). West point: facing tougher competition. Armed Forces Journal International (June 1986).

Relationships Among Precommissioning Indicators of Army Officer Performance[1]

Fumiyo T. Hunter
U.S. Army Research Institute

The Army Reserve Officers' Training Corps (ROTC) faces a challenge of increasing officer production while also assuring high leadership potential of the ROTC-trained officers (U.S. Department of the Army, 1986). This challenge is faced at a time of a continuing decline in overall college enrollment.

A critical objective in meeting the current challenge is to improve the effectiveness in assessing officer potential of cadets. An important step taken towards this end by the ROTC Program is the implementation of the Precommissioning Assessment System (PAS) (U.S. Department of the Army, 1978). PAS specifies five dimensions for assessing officer-potential: academic, psychological/attitudinal, physical, medical, and leadership. However, the predictive power of measures representing these dimensions has not been systematically investigated. The present report examines relationships among several precommissioning measures assumed to be relevant to the PAS dimensions. This step would, in turn, facilitate validating the predictive efficiency of these and other measures against indices of officer performance.

Analysis of precommissioning measures was guided by three basic questions: (1) What is the degree of association among the precommissioning measures? Near redundancy would mean that one measure can be substituted for another as a predictor; very low correlation would suggest each measure may be related to different aspects of officer performance. (2) Are the patterns of associations among the measures similar or different for cadets with ROTC scholarships and those without? Cadets receiving ROTC scholarships have already passed a screening process to assess officer potential. The intercorrelations among the precommissioning measures may be greater for this group than for the non-scholarship group if the screening process successfully selects "well-rounded," high-quality candidates. Alternatively, the intercorrelations may be lower for this group than for the non-scholarship group due to range restrictions in their data. (3) Do those cadets who score in the "marginal range" on

---

[1]Any conclusions in this report are not to be construed as official positions of the U.S. Army Research Institute or the Department of the Army unless so designated by other authorized documents.

standardized achievement tests differ substantially from higher-scoring cadets in terms of other potential predictors? This question was addressed since ROTC has considered setting screening standards based on achievement test scores.

Procedure

Analyses reported in this report were based on available data generated from two separate projects undertaken by ROTC during 1984 and 1985: (1) ROTC-wide administration of three achievement tests (Missouri College English Test, Nelson-Denny Reading Test, and Stanford Achievement Test of Mathematics) to assess cadets' basic skills (Hunter, 1986) and (2) a demographic sample survey, primarily to be used for marketing research (Hertzbach, Gade, & Johnson, 1985).

The sample for this report, consisting of 793 cadets, was created by extracting all cases for whom both the survey and the achievement test data were available, with no further sampling constraint applied. Most of these cadets participated in the survey project in the summer of 1984 just before their senior year and were enrolled in the Military Science (MS) IV class during the senior year. The subgroup composition of the sample was compared to those reported in the ROTC Enrollment Report for the school year 1984-1985 (U.S. Army Training and Doctrine Command, 1984) and the Achievement Testing Program report (Hunter, 1986). The sample constituted about 10% of the MS IV population. Of the four ROTC regions, Regions 1, 2, and 3 are over-represented, and Region 4 (the smallest region) is not included, in the sample. However, the gender and ethnic group composition of the sample closely approximates that of the total MS IV population.

Items in the two data sources judged to be relevant to the PAS dimensions were extracted. Leadership experiences in high school (6 items) and college (7 items), such as being an officer of student government/organization or class or an editor of a school publication, were used to represent the leadership dimension. The physical dimension was measured by participation in high school sports teams and winning varsity letters and sports awards (12 items). The psychological/attitudinal dimension was based on patriotic reasons for pursuing military career (3 items) and is labeled Army Career Orientation in this report. A cadet's score on each scale was the sum of items for which the keyed answer was given. The academic dimension was assessed by high school and college grade point averages, Scholastic Aptitude Test (SAT), American College Testing Program Examination (ACT), and standardized achievement tests (average of ROTC percentile scores from the three tests mentioned earlier).

Results and Discussion

First, the descriptive statistics of each measure were examined by scholarship status and are shown in Table 1. As

expected, the scholarship group scored higher than the non-scholarship group (see note under Table 1) on all measures except for Army Career Orientation and High School Sports.

Table 1
Means, Standard Deviations, and *t*-test Results for Scholarship and Non-scholarship Groups

Measure		n	Mean	SD	t	p
HS-LDSHP:	Non-S.	493	.97	1.18		
	S.	291	1.31	1.30	-3.72	< .001
HS-SPORTS:	Non-S.	493	2.24	1.74		
	S.	291	2.35	1.75	- .08	n.s.
CO-LDSHP:	Non-S.	479	.75	.83		
	S.	285	1.28	1.13	-7.49	< .001
ARMY-OR:	Non-S.	477	9.79	2.08		
	S.	283	9.77	2.21	.13	n.s.
HS-GPA:	Non-S.	487	3.87	.70		
	S.	286	4.27	.67	-8.57	< .001
CO-GPA:	Non-S.	446	5.28	2.20		
	S.	232	6.28	2.33	-5.53	< .001
SAT:	Non-S.	299	4.70	1.80		
	S.	215	3.52	1.75	7.38	< .001
ACT:	Non-S.	168	5.38	2.32		
	S.	118	4.02	2.14	5.05	< .001
AchTests:	Non-S.	494	44.89	23.53		
	S.	291	60.31	23.47	-8.88	< .001

Note: HS-LDSHP=High School Leadership Experiences, HS-SPORTS=High School Sports, CO-LDSHP=College Leadership Experiences, ARMY-OR=Army Career Orientation, HS-GPA=High School Grade Point Average, CO-GPA=College GPA, and AchTests=achievement test average. Non-S.=Non-scholarship, S.=Scholarship. Instead of reporting the actual SAT and the ACT scores, cadets selected one of 10 score range categories, higher test scores being associated with lower category numbers, resulting in the lower category number means for the scholarship group.

Table 2 shows correlations between the academic measures for the total sample. As expected, the standardized tests produced fairly high intercorrelations. The HS-GPA showed a stronger association with the standardized tests than the CO-GPA did. The correlations between the GPAs and the standardized tests were generally lower than those among the standardized tests, suggesting that school and standardized test performance might be considered as distinct indices of academic competence.

Table 2
Correlations Between Academic Measures (Total Sample)

	HS-GPA	CO-GPA	SAT	ACT
CO-GPA	.17***			
SAT	.40***	.16***		
ACT	.34***	.22***	.49***	
AchTests	.38***	.15***	.71***	.66***

Note: The N for these correlations ranged from 773 to 139 due to missing or non-existent data. For example, few ROTC cadets take the ACT. *** p < .001.

Although the means of all academic measures indicated higher performance for the scholarship group, the patterns of correlations were similar for the scholarship and the non-scholarship groups with one exception. The r between the high school and college GPAs was .06 (n.s.) for the non-scholarship group and .22 (p < .001) for the scholarship group.

Table 3 shows the correlations between the non-academic measures, and also GPAs, by scholarship groups. Overall, the magnitude of association is relatively weak, suggesting that these non-academic measures represent different attributes of individual cadets. Each of these variables could make a unique contribution to the prediction of officer job performance.

Table 3
Correlations Between Non-academic Measures by Scholarship Groups

		HS-LDSHP	HS-SPORTS	CO-LDSHP	ARMY-OR	HS-GPA
HS-SPORTS:	Non-S.	.25***				
	S.	.42***				
CO-LDSHP:	Non-S.	.22***	.03			
	S.	.22***	.11			
ARMY-OR:	Non-S.	.03	-.03	.09*		
	S.	.17**	.12*	.00		
HS-GPA:	Non-S.	.18***	.00	.05	.03	
	S.	.26***	.12*	.20***	.03	
CO-GPA:	Non-S.	.02	-.03	.06	-.03.	.06
	S.	.17**	-.02	.10	-.04	.22***

Note: S.=Scholarship, Non-S.=Non-scholarship.
* p < .05, ** p < .01, *** p < .001.

The patterns of correlations for the scholarship and the non-scholarship groups were generally comparable. However, the correlations between HS-LDSHP and HS-SPORTS, HS-GPA and CO-LDSHP, and HS-GPA and CO-GPA were greater for the scholarship than the non-scholarship group (p < .05, based on tests of difference between correlations). These results may provide support for the screening procedure for ROTC scholarships. The greater convergence among the officer-potential measures suggests that the

scholarship recipients tend to be more "well-rounded," and behave more consistently over time/situations, than the non-scholarship cadets.

The correlations between the standardized tests and the non-academic measures (not shown in tables) were consistently near-zero, regardless of the scholarship status, suggesting that the latter measures may provide sources of variance quite separate from standardized academic measures in predictor validation research.

The final set of analyses covered in greater detail the group of cadets whose average achievement test score was in the "marginal range", i.e., scores 11 through 30. This range was selected, purely for the purpose of this report, to illustrate the subgroups of the sample that may be weak in basic academic skills but who may possess other aspects of officer-potential. (Currently, there is no sound information on the relationships between these test scores and officer performance to determine what would be unsatisfactory, marginal, and satisfactory score ranges for the predictor measures.)

Table 4 presents the percentages of the total sample averaging below 11, 11-30, or above 31, by dichotomized categories of the other measures. Most of the scholarship cadets averaged above 31. The SAT and ACT are not included since they correlate strongly with the achievement tests, and Army Career Orientation is excluded since the majority of the sample indicated very positive ratings.

Table 4
Percentages of Total Sample by Measures of Officer Potential

Measures		Achievmt Test Average Score Range		
		< 11	11-30	> 30
HS-LDSHP:	No experience	2⅔	8⅔	44⅔
	1 or more positions	3	11	33
CO-LDSHP:	No experience	2	7	29
	1 or more positions	3	11	48
HS-SPORTS:	No experience	1	4	15
	1 or more teams	4	14	61
HS-GPA:	Below "B"	2	5	15
	"B" or above	3	12	62
CO-GPA:	Below "B"	0	1	0
	"B" or above	5	17	76

Note: The N for the total sample varied from 673 to 779 due to missing data.

About 5, 18, and 77% of the total sample averaged in the below-11, 11-30, and above-30 ranges, respectively. Over half of

the "marginal group" had high school leadership experiences (11% vs. 8%). The percentages were very similar for college leadership experiences. The majority of this group were members of high school sports teams, and most of them earned GPAs of "B" or above in high school and/or college. Based on 9,000 commissionees per year from the ROTC Program (which is a reasonable estimate for the next few years), about 1,600 might perform in the "marginal range" on standardized tests. However, of these cadets, 1,000-plus might have leadership experiences, interest in physical activities, and/or sound school performance records. Since these projections are based on a small and non-representative sample, they are strictly tentative. However, these results do point to the feasibility of expanding the range of attributes which are measured before commissioning and of systematic validation of the assessment measures.

Standardized tests offer uniformity in assessment procedure and norms to serve as evaluation guidelines. Performance on various standardized tests is strongly associated. If a standardized test were to be used to measure academic competence, the choice would largely depend on logistical considerations, e.g., cost, availability, and administration time. High school and college GPAs should be further examined; they may capture aspects of academic competence, e.g., effort and organizational skills, which are not measured as well by standardized tests. Future work will include development of more non-academic measures of officer potential, as well as multi-dimensional officer performance criterion measures.

References

Hertzbach, A., Gade, P. A., Johnson, R. J. (1985). The 1984 ARI Cadet Decision Making Survey: An overview of results (Working Paper 85-6). Alexandria, VA: U.S. Army Research Institute.

Hunter, F. T. (1986). ROTC Achievement Testing Program: School years 1983 - 1985 (Research Report 1429). Alexandria, VA: U.S. Army Research Institute.

U.S. Army Training and Doctrine Command. (1984). Opening enrollment report school year 1984-85 Army Reserve Officers' Training Corps. Fort Monroe, VA: Author.

U.S. Department of the Army. (1978). Review of Education and Training for Officers (RETO): Implementation plan (Vol. 2). Washington, DC: Office of the Deputy Chief of Staff for Operations and Plans.

U.S. Department of the Army. (1986). ROTC Study final report (Volume 1). Washington, DC: Reserve Officers' Training Corps Study Group.

Issues Concerning ROTC Intervention Programs[1]

Paul Twohig
U.S. Army Research Institute for the
Behavioral and Social Sciences

The major tasks of the ROTC system are to select and develop future officers. Generally, the development programs focus on providing military knowledge and skills and build on previously developed cognitive, social and leadership skills. Most entering cadets exceed minimum requirements in non-military skills, but some students may need improvement to meet the standards in key skills, e.g. communication. From time to time the ROTC system has tried special interventions to ensure that more cadets pass minimum standards and to improve the skills of other cadets substantially beyond the minimum.

This report describes two such interventions, the difficulties in such interventions, and what must be done to maximize the strength of interventions. One program is the Leadership Enrichment Program (LEP) which was designed to improve cognitive and communicative skills (Twohig, Rachford, Savell and Rigby, in press). The other is the Leadership Assessment Program (LAP) which was designed to assess and aid the development of a wide range of leadership competencies (Rogers, Lilley, Wellins, Fischl and Burke, 1985).

Before describing these programs the factors that make conducting interventions in the ROTC system difficult should be pointed out: the large number (well over 300) and geographic dispersion of ROTC departments, department and host school heterogeneity, instructor turnover, host school control of curriculum changes, and the need to tradeoff Military Science requirements with intervention requirements within a limited number of classroom hours. The major resource the system has is the set of ROTC instructors. ROTC instructors have worked hard and creatively once convinced of the value of an intervention.

The Leadership Enrichment Program was largely based on Feuerstein's Instrumental Enrichment Program (FIE) (Feuerstein, 1980). The FIE program is designed to train a wide variety of general cognitive skills and there is evidence for its effectiveness (Savell, Twohig and Rachford, in press). It involves a set of paper-and-pencil problem-solving tasks and an interactive classroom teaching style. One of the key techniques is that of bridging, in which abstract cognitive skills are applied in a specific topic area. For example, the skill of categorization might be applied to organizing the characteristics of past battles or classifying combat vehicles. The Leadership Assessment Program is an Assessment-Center approach in which cadets take part in simulations of Lieutenant tasks, including an in-basket exercise and a group planning exercise.

---

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

The LEP program has been implemented twice, academic years 1982-1983 and 1984-85, each time in about 12 schools (Twohig et al., in press). The first implementation was a pilot to define what was required for a strong implementation and was successful in that the second implementation reached the goal levels of interventions, e.g. numbers of hours of LEP taught. Unfortunately the second intervention, which was planned for three years, was stopped after one year due to a shift in priorities towards basic skills evaluation in ROTC.

Similarily, the LAP program was piloted with a few officers who were to be trainers and in a few schools. But it was not implemented widely in the system and was initially cancelled due to resource conflicts. Currently though, the program is being revived and is being used by a number of schools.

The LEP and LAP programs are similar in that they have ambitious goals - to lead to major improvements in the performance of young leaders - and they require substantial training of the instructors. They both require active involvement by students allowing instructors to get a better view of their strengths and weaknesses. Both programs use precious personnel and curriculum-hours. And both were initially cancelled because of resource tradeoffs.

It is important to look at what is required to have strong implementations with such programs to help evaluate whether they are practical:

o Program support must be gained at all levels in the ROTC system - from headquarters through the regions to the instructors and Professors of Military Science.

o Support is often best obtained through providing some training in the technique. In both programs, initially skeptical officers became believers in the value of the program after taking the role of students during training.

o Programs should be pilot-tested in a few schools and should be implemented more widely in a series of "try-and-revise" iterations.

o The ROTC departments in the pilot may have to be forgiven some of the Military Science requirements. For example, the LEP program uses up classroom hours and ROTC departments may then have less than the specified numbers of hours in some Military Science topics. Of course, evaluation measures must be used to ensure there is no actual loss of Military Science learning.

o Methods have to be defined for evaluating the effects of the programs, including the use of control groups.

Ideally, we would have information on the effectiveness of these programs to compare to the effort required. But due to their early cancellation, we do not have meaningful quantitative results. We only have indications of the programs' potential value from student and instructor observations based on surveys and interviews.

In both programs the trained instructors thought that both they and the students benefited. The instructors thought that their trainer-assessor role helped them develop their own skills. Most (above 80%) of the students thought the programs were valuable. In the LEP project the instructors found that the training aided in teaching some Military Science topics and there were improvements in writing. Instructors in both programs believed that they got an improved insight into their students' strengths and weaknesses.

No firm conclusions can be drawn about the potential for conducting complex interventions in ROTC. But anyone planning such an interventions should face up to the difficulties involved. Other approaches to reaching program goals should be considered. For instance, the LAP approach has been extended to self-study modules for students (Burke and Davis, 1985). Also it may sometimes be more cost effective to provide funds to schools for the provision of training in specific areas. In any case those conducting future interventions should review past efforts for lessons learned.

References

Burke, W.P. and Davis, P.K. (1986). The Leadership Improvement Modules of the Precommissioning Leadership Assessment Program. (ARI Research Report RR-1425). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Feuerstein, R. (1980). Instrumental Enrichment: An intervention program for cognitive modifiability. Baltimore: University Park Press.

Rogers, R.W., Lilley, L.W., Wellins, R.S., Fischl, M.A. & Burke, W.P.(1982). Development of the Precommissioning Leadership Assessment Program (ARI Technical Report TR-560). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Savell, J.M., Twohig, P.T. and Rachford, D.L. (in press). Empirical status of Feuerstein's Instrumental Enrichment technique as a method for teaching thinking. Review of Educational Research.

Twohig, P.T., Rachford, D.L., Savell, J.M. and Rigby, C.K. (in press). Implementation of a cognitive skills training program in ROTC: The Leadership Enrichment Program. (ARI Research Report). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Issues Involved in Establishing Basic Skills Standards[1]

Laurel W. Oliver
U.S. Army Research Institute

In recent years, the leadership of both civilian and military organizations has expressed concern that the basic skills deficiencies of some of their junior colleagues have impaired the effectiveness of their organizations. Army concern about basic skills deficiencies in its young officers led the U.S. Army Reserve Officers' Training Corps Cadet Command (formerly the Office of the Deputy Chief of Staff for the Reserve Officers' Training Corps) to administer achievement tests to Reserve Officers Training Corps (ROTC) cadets. The Army Research Institute (ARI) was tasked to analyze the achievement test data. The findings of the ARI data analysis replicated those of other civilian and military testing projects--i.e., the average scores of some minority groups (specifically, Blacks and Hispanics) were markedly lower than those of the white, non-Hispanic majority group (Hunter, 1986).

Such results pose a problem for organizations. If an organization sets relatively high achievement test standards to insure the basic competence of its employees in this area, it runs the risk of decreasing the representativeness of its work force. Accordingly, the purpose of this paper is to propose a framework within which the problem of adverse impact can be ameliorated and employee competence in basic skills enhanced. To this end, I shall discuss some of the issues involved in setting basic skills standards for employee selection and then offer recommendations for dealing with these issues.

Pertinent Issues

Before proceeding to recommendations, there are several pertinent issues to be considered by organizations seeking to enhance the competencies of their employees. These issues are addressed below.

Issue 1: The Role of Achievement Testing in the Total Assessment Process.

This paper focuses on issues relating to achievement test standards. However, it must be emphasized that achievement testing is only a part of the assessment process. In Army ROTC, for example, the emphasis is on the "whole person." The Cadet Command has implemented a revised and expanded Precommissioning Assessment System (PAS) as the primary measure for quality control in selecting cadets. The PAS comprises 12 elements (e.g., medical exam, grade point average, basic camp performance), no one of which can by itself be used to eliminate a cadet from the ROTC program.

---

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

Issue 2: Social Policy.

It should be clearly recognized that adverse impact is a matter of social policy. It is not a question of test construction, statistical analysis, or any other technical consideration. In discussing fairness in selection procedures (an issue which has been extensively discussed in the educational measurement literature), Novick (1984) concluded, "It seems to be widely accepted that there are no psychometric solutions to questions of fairness to groups or individuals and that only a consensus of value judgments can solve these problems" (p. 230).

The failure to acknowledge the social and political implications of setting test standards tends to impede the resolution of the problem. In a discussion of "Selection Theory for a Political World," Cronbach (1982) has asserted that an "accommodation among individual, corporate, and national interests can be worked out more easily if the tensions are acknowledged" (p. 38).

Issue 3: Achievement vs. Aptitude Testing.

Tests such as the Scholastic Aptitude Test (SAT) or the American College Testing Program test (ACT) are generally considered measures of aptitude--i.e., tests which assess a person's potential for academic achievement. Achievement tests, on the other hand, are designed to measure how much a person has learned about a specific subject. The advantage of achievement over aptitude tests in selection is that one can presumably acquire specific knowledge more easily than one can acquire "intelligence" or "academic potential." If an organization's purpose is to insure minimum competence in basic skills, testing those skills directly is a more practical approach than is using aptitude tests.

Issue 4: Determining Achievement Test Standards.

Criteria for cut scores. Setting standards for an achievement test (or for any other kind of selection device) involves deciding where the cut score should be set. (The "cut score" is the level at or above which people "pass.") In order to set useful standards, it is essential to have criteria or guidelines for setting the standards. After a perusal of the cut score literature, the author concluded that there were four principal factors to be considered in determining basic skills standards:

(1) The minimal knowledge or skill required. The cut score should not be set so low that people who lack the needed proficiency in the basic skill would pass.

(2) Current and anticipated personnel requirements. In setting the cut score, one must take into account how many people are and will be required by the organization. If a large proportion of applicants must be selected, setting high cut scores is impractical.

(3) Expense of testing. The time, money, and aggravation of a testing program should not exceed its benefits.

(4) Subgroup representation. Most organizations strive to have their workforce representative of society as a whole.

Trade-offs in optimizing criteria. When these criteria are applied to setting test standards, it is apparent that optimizing any one of them may adversely affect one or more of the other three. To increase the basic skills competence of employees by setting high standards, for example, would decrease the number of applicants that could be selected.

One solution to the subgroup representation problem, employed by some colleges and universities for admissions standards, is to set different standards for different groups. From a psychometric standpoint, this approach insures that the most qualified persons in all groups are selected. All persons within each group are rank ordered, and it is determined how many will be selected from each group. Selection is then made from the top down until the allotted number of persons in each group has been selected.

However, using a multiple standard is generally considered unacceptable in the area of personnel selection. In an article entitled, "The Realities of Employment Testing," Tenopyr (1981) has flatly stated that "...having different critical scores for groups is not a viable employment policy for either a public or a private employer" (p. 1121). Flaugher (1978), in a discussion of the various definitions of test bias, concurs that the dual standard (for minority and majority groups) approach is unpalatable to many, "violating as it does the treasured principle of equal opportunity" (p. 672).

Issue 5: Remediation/Development Issue.

Setting achievement test score standards at any level will mean that some people will not pass them. The higher the cut score, the fewer the people who can attain the standard. However, many people may be "marginal"--that is, they almost but not quite attain the cut score. For these individuals, activities such as tutoring, extra classes, study skills workshops, etc. may be helpful in developing the desired competencies. For example, efforts to strengthen academic weaknesses of ROTC cadets or prospective cadets have been undertaken by 20 of the 21 Historically Black Colleges (HBCs) which house Army ROTC programs. And the University of Texas has had underway for several years a program to strengthen the cognitive skills of marginal students.

Issue 6: Implications of Government and Professional Guidelines for Achievement Test Standards.

The "Uniform Guidelines on Employee Selection Procedures" is a document developed by the federal government. The provisions of Title 7 of the 1964 Civil Rights Act, the 1972 amendment of the Civil Rights Act, and various court decisions concerning these procedures resulted in a certain amount of confusion concerning selection procedures. The purpose of the Uniform Guidelines was to construct a set of common guidelines upon which the various federal agencies could agree.

Careful consideration of these guidelines could help ensure the fairness of test-related decisions, such as those concerning achievement test standards. For example, when adverse impact results from selection procedures (such as the use of tests), the Uniform Guidelines require that these procedures be validated--that is, the tests must be shown to be related to success on the job.

In addition to the Uniform Guidelines, there are two other documents which provide guidelines for the construction and use of tests. These are "Standards for Educational and Psychological Testing," developed jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council for Measurement in Education (NCME), and "Principles for the Validation of Selection Procedures," developed by the Division of Industrial/Organizational Psychology of the APA. These two sets of guidelines represent uncounted hours (ranging over months and years) of effort by dedicated professionals to develop guidance based on the latest developments in testing research and practice.

Issue 7: Need for Empirical Data.

Sometimes decisions must be made even though the decision maker lacks valid information on which to base those decisions. However, virtually every aspect of human resources management requires the analysis of jobs and the identification of job requirements. Valid job analysis data are essential for the proper functioning of selection, classification, training, and performance appraisal systems. As an ARI report noted, "Without first assessing job requirements, selection instruments and training programs will fail to meet their objectives" (Wellins, Rumsey, & Gilbert 1980).

Only when the job requirements have been determined can appropriate measures of performance be developed. Once valid and reliable measures of performance are available, it is possible to use them as criteria for predictors of job success. To establish the relationships among these factors (job requirements, performance criteria, and predictors), requires research and, ideally, longitudinal data. With the increasing automatization of personnel information, it is becoming more feasible to collect the longitudinal data needed for evaluating selection and classification procedures as well as developmental programs.

Recommendations

Consideration of the issues discussed above leads to the framework outlined below. This framework comprises a set of recommendations, some of which are more easily implemented while others would require a longer time period and/or a greater investment of resources.

Short-Term Recommendations.

1. Apply the same standards to all. All persons would be subject to the same requirements. No subgroups would be treated differently.

2. Set a mandatory minimum score and a higher desired minimum score. Setting a relatively low (at, say, the 5th percentile) score will eliminate those individuals with truly low scores, while specifying a higher desired minimum will give those with lower scores a standard at which to aim.

3. Average over several tests. Averaging over two or more tests allows persons with relatively higher scores on one test to compensate for lower scores on another test. This procedure increases the reliability of the procedure and, for some test score ranges, is of greater benefit to minority than majority subgroups (Hunter, 1986).

4. Insure that remediation/developmental activities are available. While remedial/developmental activities may not help every student (especially poorly motivated ones), they are potentially of great benefit to the motivated, marginal student.

Long-Term Recommendations.

1. Evaluate remedial/developmental procedures. The remedial/developmental programs that students participate in should be evaluated to assess their usefulness. Some remedial strategies will be more helpful than others, and these should be identified and their use encouraged.

2. Conduct job analyses of target jobs. Job analysis is required for the proper implementation of the two following recommendations.

3. Identify/develop meaningful measures of job performance. Reliable and valid measures should be developed which tap various aspects of job performance. These should not be limited to academic measures, but should measure all critical facets of performance.

4. Identify/develop meaningful predictor measures. These measures should represent all facets of performance and be validated against the performance measures identified or developed in No. 3 above. (Validation involves determining the relationship between the predictor measures and performance.)

5. Follow employees over time. To determine factors associated with effective performance and to evaluate the effects of various developmental programs and procedures, longitudinal follow-up is needed. The Officer Longitudinal Research Data Base (OLRDB), now being established at ARI, will be a useful vehicle for conducting such research on Army officers.

Concluding Remarks

While the short-term recommendations outlined above will be of advantage to an organization, it is the long-term recommendations that can lead to tangible, measurable benefits. Unfortunately, organizations want to find "quickie" solutions rather than committing the resources needed to explore the problem in depth. The issues discussed above, however, are not amenable to the "quick fix" that is so ardently desired by many organizations. As John Campbell (1983) has cautioned: "There are no quick fixes, and nothing will substitute for careful problem analysis and long term commitment to painstakingly worked out solutions" (p. 11).

References

Campbell, J. P. (1983). I/O psychology and the enhancement of productivity. The Industrial/Organizational Psychologist, 20, 6-10.

Cronbach, L. J. (1980). Selection theory for a political world. Public Personnel Management, 9, 37-50.

Flaugher, R. L. (1978). The many definitions of test bias. American Psychologist, 33, 671-679.

Hunter, F. T. (1986). ROTC Achievement Testing Program: School years 1983-1985. (Research Report 1429). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Novick, M. (1984). Federal guidelines and professional standards. In Readings in professional personnel Assessment (pp. 229-254). Washington, DC: International Personnel Management Association.

Tenopyr, M. L. (1981). The realities of employment testing. American Psychologist, 36, 1120-1127.

Wellins, R. S., Rumsey, M. G., & Gilbert, A. C. F. (1980) Analysis of junior officer training needs. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Socail Sciences.

BYLAWS 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

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 association 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 format request to the Chairman of the Steering Commttee, 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 Bylaws may be made at any annual conference of the Association.

B. Amendments of the Bylaws 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 July 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 Bylaws 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

Belgian Armed Forces Psychological Research Section

Canadian Forces Directorate of Military Occupational Structures

Canadian Forces Personnel Applied Research Unit

Defense Activity for Non-Traditional Education Support

Federal Republic of Germany Ministry of Defense

National Headquarters Selective Service System

Royal Australian Air Force Evaluation Division

U. S. Air Force Human Resources Laboratory

U. S. Air Force Occupational Measurement Center

U. S. Army Research Institute

U. S. Coast Guard Institute

U. S. Naval Education and Training Program Development Center

U. S. Navy Occupational Data Analysis Center

U. S. Navy Personnel Research and Development Center

MINUTES
MTA Steering Committee Meeting
28th Annual Conference
3-7 November 1986
Mystic, Connecticut

The meeting was opened at 1000 on 3 November by CDR Earl H. Potter III, 1986 Conference Chairman, and the attendees were introduced (see attached list). The Minutes of the 1985 meeting were read and a financial report for the 1985 Conference presented (see attached report).

PROCEEDINGS

The first item on the agenda was a discussion of the Conference Proceedings which drew in issues concerning the nature of the conference itself as well. The unit cost of the Proceedings in the last several years has approached $40-45,000 U.S. Some hosting organizations have funded these costs "out of hide" others have had "in house" resources to publish the Proceedigs. LTCOL Pinch noted that many conferences do not publish a Proceedings or, if they do, the Proceedings are abreviated. Dr. Holz suggested changing the format of the MTA Conference to include more exchange and dialogue and fewer paper presentations. COL Baker suggested a "tightening up" of the quality while COL Pinch and Dr. Tartell suggested keeping the Conference open for younger professionals. It was generally agreed that submitting the paper for review would aid in both the quality of the papers and publishing the Proceedings on time. Ms. Jones noted that the Proceedings represented an opportunity for professional development and a significant record of the Association's work. The consensus was that the Proceedings would be published for the 1986 and 1987 conference in the familiar format. In 1988 ARI will publish the Proceedings, but may change the format.

PAPER SUBMISSION

Following further discussion of the process of submitting papers for review, it was moved (Wiskoff) and seconded (Holz) that the call for papers require a three page draft with an abstract to be submitted for review. There was general agreement that presenters should bring copies of the paper to the Conference for distribution and that MTA should encourage the submission of papers to professional publications. Some concern was expressed that three pages was too short to include necessary information. It was agreed that a full paper would be longer than three pages and that three pages really represented an expanded abstract. COL Zypchen noted that the call should go out in March to allow presenters time to meet the greater requirement. The motion carried (16-3). Dr. Wiskoff suggested that we try the new submission requirement for one year and discuss the issues again in 1987.

BYLAWS

COL Zypchen moved that the Bylaws be amended to required a 1 July deadline vice 1 January deadline for the Harry H. Greer Award nomination (Article XI, Section A2). It was

generally agreed that this schedule was more reasonable and that it would facilitate a greater number of nominations. The motion carried unanimously.

FUTURE CONFERENCE SCHEDULES

COL Zypchen, as host, invited the members of the Steering Committee to the 1988 MTA meeting in Ottawa. It was noted that ARI will host the conference in November 1988 in Washington, D.C., the Air Force in San Antonio in 1989, and the Navy in 1990 on the East Coast.

The meeting was adjourned at 1110.

STEERING COMMITTEE MEETING

Name	Organization
COL Ronald C. Baker	USAFOMC/CC, Randolph AFB, TX
Dr. Lloyd Burtch	AFHRL/XO, Brooks AFB, TX
LCDR R. W. Clark	NODAC/WNY, Bldg 150, Anacostia, D.C.
MAJ Ron Dickenson	NDHQ/DPSRSC, Ottawa, Canada
Dr. John A. Ellis	Navy Personnel R&D Center, San Diego, CA
SQNLDR Ken Given	Royal Australian Air Force
Dr. Robert F. Holz	U. S. Army Research Institute, Alexandria, VA
Ms. Karen N. Jones	U.S. Coast Guard Institute, Oklahoma City, OK
CDR Jerrold E. Olson	Naval Education and Training Program Management Support Activity
LTCOL Franklin C. Pinch	DPSRSC/National Defence Hdqtrs, Ottawa, Canada
LTCOL Terry J. Prociuk	CF Personnel Applied Research Unit, Toronto, Canada
Dr. Hendrick W. Ruck	AFHRL/IDT, Brooks AFB, TX
Dr. Peg Smith	Naval Education Training Program Management Activity Support , Pensacola, FL
Dr. Friedrich W. Steege	FMOD - P II 4, Bonn, Federal Republic of Germany
Dr. J. S. Tartell	USAFOMC/OMY, Randolph AFB, TX
Dr. Ray Waldkoetter	U. S. Army Soldier Support Center
CDR R. J. Wilson	NODAC/WNY, Bldg 150, Anacostia, D.C.
Dr. Martin F. Wiskoff	Navy Personnel R&D Center, San Diego, CA
COL G. A. Zypchen	NDQH/DMOS, Ottawa, Canada

30 October 1986

1985 MTA Income and Expenses
27th Annual Conference

Starting balance		$7,179
Income:		
Registration fees	18,364	
Banquet tickets for guests	320	
Misc. income	34	
		$18,718
Total available		$25,897
Expenses:		
Hotel and banquet costs	9,050	
Printing of proceedings	10,600	
Misc. expenses Banquet tickets, badges, logos, envelopes, plaques, flowers, etc.	716	
Total expenses		$20,366
Balance transferred to Coast Guard		$5,531

MTA 28TH ANNUAL CONFERENCE STAFF

MTA President

RADM R. P. Cueroni

MTA Chairman

CDR Earl H. Potter III

Committee Chairpersons/Members

Program Committee

CDR Earl H. Potter III

Operations Committee

LT Robert R. Albright II
LCDR Timothy W. Hylton
Prof. Philip I. Mathew
CDR Earl H. Potter III
Ms. Rita J. Smith
Prof. David W. Weber
LT Robert D. Williamson
LT William L. Zack

Finance

LT Robert D. Williamson

Publications/Graphics

LT William L. Zack

Hospitality

LT William L. Zack

"A"

PATRICIA A. ALBA
MAXIMA CORPORATION
8301 BROADWAY, SUITE 212
SAN ANTONIO, TX 78209

PRISCILLA ANDERSON
ALLEN CORPORATION OF AMERICA
10080 CARROLL CANYON ROAD
SAN DIEGO, CA 92131

DR. JANE M. ARABIAN
ATTN: PERI-RS
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

1LT THOMAS O. ARTH
ATTN: MOAO
AIR FORCE HUMAN RESOURCES LABORATORY
BROOKS AFB, TX 78235-5601

DR. RICHARD D. ARVEY
INDUSTRIAL RELATIONS CENTER
574 MANAGEMENT & ECONOMICS BLDG
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

DR. CAROL V. ASCHERFELD
NETPMSA (CODE 3143)
SAUFLEY FIELD
PENSACOLA, FL 32509-5000

NANCY K. ATWOOD
U. S. ARMY RESEARCH INSTITUTE
P.O. BOX 8757
PRESIDIO OF MONTEREY, CA 93944 5011

"B"

DR. HERBERT G. BAKER
NAVY PERSONNEL R&D CENTER (CODE 62)
SAN DIEGO, CA 92152-6800

COL. RONALD C. BAKER
COMMANDER
USAF OCCUPATIONAL MEASUREMENT CTR
RANDOLPH AFB, TX 78150-5000

THOMAS B. BAKER
U.S. ARMY ORDNANCE CENTER & SCHOOL
440 DORIS CIRCLE
ABERDEEN, MD 21001

DR. LOUIS E. BANDERET
U.S. ARMY RESEARCH INSTITUTE OF
ENVIRONMENTAL MEDICINE
NATICK, MA 01760-5007

JEFF BARNES
HUMAN RESOURCES RESEARCH ORG (HumRRO)
1100 S. WASHINGTON ST.
ALEXANDRIA, VA 22314

LT(N) EDWARD G. BARNETT
CANADIAN NAVY
D COMDT (D)
CFFS (H)
FMO HALIFAX
HALIFAX, N.S. B3K 2X0

DR. WILLIAM M. BART
EDUCATIONAL PSYCHOLOGY
330 BURTON HALL
UNIVERSITY OF MINNESOTA
178 PILLSBURY DRIVE, S.E.
MINNEAPOLIS, MN 55455

LORI BAUMANN
ALLEN CORPORATION OF AMERICA
10800 CARROLL CANYON ROAD #A
SAN DIEGO, CA 92131

HELEN E. BELLETTI
ATTN: ATSA-TAC
U. S. ARMY AIR DEFENSE SCHOOL
FT. BLISS, TX 79916

EMILY BERRY
NAVY EDUCATION & TRNG PROG MGT
SUPPORT ACTIVITY
SAUFLEY FIELD
PENSACOLA, FL 32509-5000

MAJ. E. R. BLACK
CANADIAN FORCES PERSONNEL APPLIED
RESEARCH UNIT (CFPARU)
4900 YONGE STREET, SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

BRUCE M. BLOXOM
DEFENSE MANPOWER DATA CENTER
550 CAMINO EL ESTERO, SUITE 200
MONTEREY, CA 93940-3231

ARNOLD C. BOHRER
PSY RESEARCH S. B.A.F.
REKRUITERINGS-EN SELECTIE CENTRUM
SECTIE PSYCHOL. ONDERZOEK
BRUYNSTRAAT
1120 BRUSSEL, BELGIUM

TROY L. BOOKER
NAVAL EDUCATION & TRNG PROGRAM MGT
SUPPORT ACTIVITY (CODE 041-1)
SAUFLEY FIELD
PENSACOLA, FL 32509-5000

MAJ. GRAHAME BROWN, MA, RAEC
ARMY SCHOOL OF TRAINING SUPPORT
RAEC CENTRE
BEACONSFIELD
BUCKS HP9 2RP
U.K.

DR. LARRY D. BROWN
RESEARCH ANALYST
BUILDING T-251
U.S. ARMY TRAINING BOARD
FT. MONROE, VA 23651

DR. KENDALL J. BRYANT
NAVAL SUBMARINE MEDICAL RESRCH LAB
U. S. NAVAL SUB BASE
GROTON, CT 06349

LAWRENCE S. BUCK
PLANNING RESEARCH CORPORATION
1440 AIR RAIL AVENUE
VIRGINIA BEACH, VA 23455

GLEN R. BUDGELL
PERSONNEL PSYCHOLOGY CENTRE
PUBLIC SERVICE COMMISSION OF CANADA
300 LAURIER AVENUE, WEST
WEST TOWER
OTTAWA, ONTARIO
CANADA K1A 0M7

GARY R. BUNDE
NAVY TECHNICAL TRAINING CENTER
8237 LYRIC DRIVE
PENSACOLA, FL 32514

EUGENE F. BURKE
U.K. EXCHANGE PSYCHOLOGIST
AIR FORCE HUMAN RESOURCES LAB/MOAE
SAN ANTONIO, TX 78235-5601

ERNEST J. BURNELL
USMEPCOM
151 FOREST AVENUE / P.O. BOX 8190
PORTLAND MEPS
PORTLAND, ME 04104-8190

DR. LLOYD D. BURTCH
AIR FORCE HUMAN RESOURCES LAB/XO
BROOKS AFB, TX 78235-5601

JANICE BUSH
USMEPCOM
HBG, MEPS
NCAD, PA

AJIT S. BUTTAR
ATTN: ATZM-CM-ES
U.S. ARMY CHEMICAL SCHOOL / DOES
FT. MCCLELLAN, AL 36205-5020

"C"

WAYNE J. CAMARA
HUMAN RESOURCES RESEARCH ORG.
1300 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

CHARLOTTE H. CAMPBELL
HUMAN RESOURCES RESEARCH ORG (HumRRO)
295 WEST LINCOLN TRAIL BLVD
RADCLIFF, KY 40160

RICHARD CAMPBELL
TRANSPORT CANADA
PLACE DE VILLE
OTTAWA, ONTARIO
CANADA K1A ON8

DR. JEFFREY A. CANTOR
DDL OMNI ENGINEERING
1126 KERSEY ROAD
SILVER SPRING, MD 20902

EDMUND J. CARBERRY
U.S. ARMY ARMOR SCHOOL
7303 RIDAN WAY
LOUISVILLE, KY 40214

MARY LOU CARBERRY
TRAINING TECHNOLOGY AGENCY FIELD OFFICE
ATTN: ATTG-DK
GAFFEY HALL
FT. KNOX, KY 40121-5200

JAMES W. CAREY
COMMANDANT (G-PTE)
U. S. COAST GUARD
2100 SECOND STREET, S.W
WASHINGTON, DC 20593-0001

CW04 PETER F. CASE
COMMANDANT (PMR)
U. S. COAST GUARD
2100 SECOND STREET, S.W
WASHINGTON, DC 20593-000

STEVE CECIL
DEPARTMENT OF THE NAVY
5911 EDSALL ROAD #609
ALEXANDRIA, VA 22304

JEANNA F. CELESTE
WESTAT, INC.
1650 RESEARCH BLVD
ROCKVILLE, MD 20850

JAN CHERVENAK
CHIEF, ANALYSIS DIVISION
ATTN: ATSH-ES
U. S. ARMY INFANTRY SCHOOL
FT. BENNING, GA 31905-5420

LCDR RAY W. CLARK
NMPC DET - NODAC
BLDG 150
WASHINGTON NAVY YARD (ANACOSTIA)
WASHINGTON, DC 20374-1501

DR. HARRY B. CONNER
ATTN: CODE 52E
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

EDITH A. CROHN
HEALTH & PERFORMANCE DIVISION
U. S. ARMY RESEARCH INSTIT OF ENVIRON MEDICINE
KANSAS STREET
NATICK, MA 01760-5007

DR. MARK CZARNOLEWSKI
ATTN: PERI-RS
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

"D"

LT DIANNE DANIELS
NMPC DET - NODAC
BLDG 150
WASHINGTON NAVY YARD (ANACOSTIA)
WASHINGTON, DC 20374-1501

DR. DOUGLASS DAVIS
TASK ANALYSIS & DESIGN SPECIALIST
(TAP 31)
COMMANDANT OF THE MARINE CORPS
WASHINGTON, DC 20380-0001

MAJ. R. A. V. DICKENSON
RESEARCH COORDINATOR
ATTN: DPSRSC
NATIONAL DEFENCE HEADQUARTERS
OTTAWA, CANADA K1A 0K2

MARTHA W. DILG
ATTN: TDA-24
HEADQUARTERS, U. S. MARINE CORPS
WASHINGTON, DC 20380-0001

CAPT BENJAMIN L. DILLA
HQ AFMPC/DPMYAF
RANDOLPH AFB, TX 78150-6001

DENNIS A. DILLON
QUESTAR DATA SYSTEMS, INC.
7146 SHADY OAK ROAD
EDEN PRAIRIE, MN 55344

BARBARA G. DORSEY
DEFENSE LANGUAGE INSTITUTE
202 ARCADIA #101
SAN ANTONIO, TX 78209

COL. DR. SHLOMO DOVER
ISRAELI DEFENCE FORCE
MILITARY P.O. BOX 01172
ISRAEL

EARL L. DOYLE, JR.
HUMAN RESOURCES RESEARCH ORG (HumRRO)
295 WEST LINCOLN TRAIL BLVD
RADCLIFF, KY 40160-2042

RUSSELL J. DRAKELEY
DEPT OF OCCUPATIONAL PSYCHOLOGY
BIRBECK COLLEGE
UNIVERSITY OF LONDON (U.K.)
MALET STREET
LONDON, U.K. WC 1

DR. WILLIAM P. DUNLAP
DEPARTMENT OF PSYCHOLOGY
TULANE UNIVERSITY
NEW ORLEANS, LA 70118

"E"

DALE R. ECKARD
NAVAL EDUCATION & TRNG PROGRAM
MGT SUPPORT ACTIVITY (CODE 3162)
PENSACOLA, FL 32509-5000

DOROTHY S. EDWARDS
AMERICAN INSTITUTES FOR RESEARCH
1055 THOMAS JEFFERSON ST, N.W.
WASHINGTON, DC 20007

STEPHEN J. ELLIOTT
RAAF PSYCH SERVICE/DOD AIR FORCE OFFICE
RUSSELL OFFICES (E-3-33)
P.O. BOX E33 - QUEEN VICTORIA TERRACE
CANBERRA ACT 2600
AUSTRALIA

DR. JOHN A. ELLIS
ATTN: CODE 51
NAVY PERSONNEL R&D CENTER
CATALINA BLVD
SAN DIEGO, CA 92152-6800

RICHARD M. EVANS
TAEG
DEPARTMENT OF THE NAVY
NAVAL TRAINING SYSTEMS CENTER
ORLANDO, FL 32813

"F"

AMANDA J.W. FEGGETTER
HEAD, HUMAN FACTORS UNIT
HQDAAC (APRE)
MIDDLE WALLOP
N. STOCKBRIDGE, HAMPSHIRE
U.K.

DR. DANIEL B. FELKER
AMERICAN INSTITUTES FOR RESEARCH
1055 THOMAS JEFFERSON ST, N.W.
WASHINGTON, DC 20007

JAMES M. FERSTL
COMMANDANT (G-PTE)
U. S. COAST GUARD
2100 SECOND STREET, S.W.
WASHINGTON, DC 20593-0001

BERNARD J. FINE
HEALTH & PERFORMANCE DIVISION
U.S. ARMY RESEARCH INSTIT OF ENVIRON MEDICINE
ARMY R,D & E CENTER
NATICK, MA 01760-5007

DR. MYRON A. FISCHL
ATTN: DAPE-ZBR-S
OFFICE OF DEPUTY CHIEF OF STAFF FOR PERSONNEL
U. S. ARMY HEADQUARTERS
THE PENTAGON
WASHINGTON, DC 20310-0300

DR. GERALD P. FISHER
SENIOR SCIENTIST
HUMAN RESEARCH RESOURCES ORG (HumRRO)
1100 SOUTH WASHINGTON ST
ALEXANDRIA, VA 22314

RONALD L. FLAUGHER
EDUCATIONAL TESTING SERVICE
PRINCETON, NJ 08541

PATRICK FORD
HUMAN RESEARCH RESOURCES ORG (HumRRO)
295 W. LINCOLN TRAIL BLVD
RADCLIFF, KY 40160

ROBERT L. FREY, JR.
COMMANDANT (G-P-1/2)
U. S. COAST GUARD - ROOM 4200 B
2100 SECOND STREET, S.W.
WASHINGTON, DC 20593-0001

"G"

CAPT T. J. GALLAGHER
ATTN: CODE 60A
NAVAL AIR DEVELOPMENT CENTER
WARMINSTER, PA 18974-5000

SHARON K. GARCIA
AIR FORCE HUMAN RESOURCES LAB / MODJ
BROOKS AFB, TX 78235-5601

ILENE F. GAST
ATTN: PERI-RS
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

DENNIS GAYNOR
U.S. MEPCOM
1140 INVERARY LANE
DEERFIELD, IL 60015

LTCOL FRANK C. GENTNER
MANPOWER, PERSNEL & TRNG DIRECTORATE
ASD/ALHA
WRIGHT-PATTERSON AFB, OH 45433

KENNETH C. GIVEN, SQNLDR
ROYAL AUSTRALIAN AIR FORCE EXCHG OFFCR
AIR FORCE HUMAN RESOURCES LAB / MODJ
BROOKS AFB, TX 78235-5601

DR. DWIGHT J. GOEHRING
U.S. ARMY RESEARCH INSTIT - FIELD UNIT
P.O. BOX 5787
PRESIDIO OF MONTEREY, CA 93944-5011

DR. LAWRENCE A. GOLDMAN
ATTN: ATNC-MOT-C
CDR, U.S. ARMY SOLDIER SUPPORT CTR - NCR
200 STOVALL STREET
ALEXANDRIA, VA 22332-0400"

DEBORAH L. GOODNOW
ATTN: ATSA-PD
PATRIOT DEPT
USAADASCH
FT. BLISS, TX 79916-7180

CAROL A. GRIFFITHS
USMEPCOM
BLDG 2845
FT. GEORGE G. MEADE, MD 20755-6000

"H"

HEINZ JUERGEN HAETTIG
STREITKRAEFTEAMT DEZ WEHRPSYCHOLOGIE
POSTFACH 20 30 05
5300 BONN 2
FRG

JAMES P. HANLON
SHIPPENSBURG UNIVERSITY
SHIPPENSBURG, PA 17257

JAMES H. HARRIS
HUMAN RESOURCES RESEARCH ORG (HumRRO)
1100 S. WASHINGTON ST.
ALEXANDRIA, VA 22314

LTCOL D. A. HARRIS
ATTN: OASD (FM&P)
DIRECTORATE FOR ACCESSION POLICY
THE PENTAGON, ROOM 2B271
WASHINGTON, DC 20301-4000

FREDERICK J. HAWRYSH
ATTN: DMOS
NATIONAL DEFENCE HEADQUARTERS
101 COLBY DRIVE
OTTAWA, ONTARIO
CANADA K1A 0K2

LINDA HAYS
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON, NJ 08541

ROSE HEMPSTEAD
ATTN: ATSA-DT-H
U. S. AIR DEFENSE SCHOOL
FT. BLISS, TX 79936

SUSAN M. HERNANDEZ
ASVAB TEST SPECIALISTS
BALTIMORE MILITARY ENTRANCE PROCESSING
STA
793 ELKRIDRIGE LANDING RD
LINTHICUM HEIGHTS, MD 21090-2995

DR. ALLYN HERTZBACH
ATTN: PERI-RG
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

REBECCA D. HETTER
ATTN: CODE 63
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

D. G. HIESTER
VICE PRESIDENT
PERFORMANCE METRICS, INC.
825 CALLAGHAN, SUITE 225
SAN ANTONIO, TX 78228

MAJ. G. J. HIGGS
CANADIAN FORCES
135 BENLEA DRIVE
NEPEAN, ONTARIO
CANADA K2G 3V6

DR. EDWARD N. HOBSON
MANTECH MATHETICS CORP.
5904 OLD RICHMOND HWY, SUITE 301
ALEXANDRIA, VA 22303

DR. WALTER J. HOFFELT
GAF INST. OF AVIATION MEDICINE
8080 FURSTENFELDBRUCK
FLIEGERHORST
FRG

DR. R. GENE HOFFMAN
HUMAN RESOURCES RESRCH ORG (HumRRO)
295 W. LINCOLN TRAIL BLVD
RADCLIFF, KY 40160

LT F. D. HOLCOMBE
ATTN: CODE 112
NAVAL AEROSPACE MEDICAL INSTITUTE
NAVAL AIR STATION
PENSACOLA, FL 32508-5600

DR. FRED D. HOLT
ATTN: ATZN-CM-ES
COMMANDANT
U. S. ARMY CHEMICAL SCHOOL, DOES
FT. MCCLELLAN, AL 36205-5020

ROBERT F. HOLZ
ATTN: PERI-RL
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

DAVID K. HORNE
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

DR. GLORIA B. HOUSTON
PERSONNEL QUALIF STANDARDS (PQS)
NAVAL ED TRNG & SUPP CTR, PACIFIC (N-7)
SAN DIEGO, CA 92132

DR. FUMIYO T. HUNTER
ATTN: PERI-RL
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

W. KARL HUNTSINGER
ATTN: CODE 30
NAVAL OCCUPATIONAL DEVLPMT & ANALY CTR
WASHINGTON NAVY YARD, BLDG 150
WASHINGTON, DC 20374-1501

PATRICIA M. HURST
U. S. MARINE CORPS
5991 ANCHUCA COVE
MILLINGTON, TN 38053

"I"

JOSEPH W. ILLES
ATTN: ATSK-E
U.S. ARMY ORDNANCE MSC &
MUN CTR & SCHOOL
REDSTONE ARSENAL, AL 35897-6000

TIMOTHY A. ISENHART
U.S. ARMY ACADEMIC TRAINING CENTER
AJGH-PT-ED
APO SAN FRANCISCO 96343

"J"

IAN L. JACKSON
STAFF DEVELOPMENT BRANCH
PUBLIC SERVICE COMMISSION
L'ESPLANADE LAURIER, WEST TOWER
300 LAURIER AVENUE WEST
OTTAWA, ONTARIO
CANADA K1A OM7

DR. DAVID M. JOHNSON
2990 TRAWOOD DRIVE, 8F
EL PASO, TX 79936

ALAN JONES
MINISTRY OF DEFENCE (NAVY), U.K.
SP(N), ROOM 426 ARCHWAY
BLOCK SOUTH, SPRING GARDENS
LONDON SWIA 2BE
U. K.

KAREN N. JONES
U.S. COAST GUARD INSTITUTE
P.O. SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

GLORIA JONES-JAMES
ATTN: CODE 63
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

"K"

DR. ROBERT S. KENNEDY
ESSEX CORPORATION
1040 WOODCOCK ROAD, SUITE 227
ORLANDO, FL 32803

CDR ROBERT H. KERR, CF
CF FLEET SCHOOL HALIFAX
323 COLBY DRIVE
DARTMOUTH, NOVA SCOTIA
CANADA 82V 2B7

WILLIAM F. KIECKHAEFER
RGI, INCORPORATED
1360 ROSECRANS, SUITE F
SAN DIEGO, CA 92106

DR. MELVIN J. KIMMEL
ATTN: PERI-RP
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

ROBERT W. KING
NAVAL ED & TRNG PROG MGT SUPP ACTIVITY
PENSACOLA, FL 32509

DEIRDRE J. KNAPP
ATTN PERI-RS
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

DR. ROBERT E. KRUG
AMERICAN INSTITUTES FOR RESEARCH
1100 MASSACHUSETTS AVE
CAMBRIDGE, MA 02138

ELIZABETH KUNZ
DYNAMIC RESEARCH CORP
60 FRONTAGE ROAD
ANDOVER, MA 01810

JAMES H. KVICALA
U.S. ARMY INFANTRY SCHOOL (DOES)
FT. BENNING, GA 31905

"L"

DAVID R. LAMPE, SR. EDITOR
DEFENSE MGT JRNL, DEPT OF DEFENSE
716-R CHURCH STREET
ALEXANDRIA, VA 22314

FAY J. LANDRUM
NAVAL ED & TRNG PROG MGT SUPP ACTIVITY
SAUFLEY FIELD
PENSACOLA, FL 32509-5000

RICHARD S. LANTERMAN
COMMANDANT (G-P-1/2)
U. S. COAST GUARD, ROOM 4200
2100 SECOND STREET, S.W.
WASHINGTON, DC 20593-0001

SHANE LEON LATIMER
ROYAL NAVAL SCHOOL OF EDUCATION &
TRAINING TECHNOLOGY
H.M.S. NELSON
PORTSMOUTH, HAMPSHIRE
U.K. P01 3HH

LCDR GEOFFREY W. LEACH
ROYAL AUSTRALIAN NAVY
c/o KILLEARN
3813 LEANE DRIVE
TALLAHASSEE, FL 32308

PEG LEAVITT
INDIANAPOLIS MEPS
141 S. MERIDIAN - 5TH FLOOR
INDIANAPOLIS, IN 46225-1088

LISA L. LEFFLER
MILITARY ENTRANCE PROCESSING STATION
7070 SPRING STREET
OMAHA, NE 68106

HARRIS R. LIEBERMAN
E2O-138 DEPT OF BRAIN & COGNITIVE SERVCS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MA 02139

DR. RICHARD A. LILIENTHAL
ATTN: PECC-CMP
U. S. ARMY CIVILIAN PERSONNEL CENTER
200 STOVALL STREET
ALEXANDRIA, VA 22332-0300

DAVID K. LIPHAM
NAVAL ED & TRNG PROG MGT SUPP ACTIVITY
SAUFLEY FIELD
PENSACOLA, FL 32509-5000

DR. STEPHEN G. LISTER
ARMY PERSONNEL RESEARCH ESTAB
c/o ROYAL AIRCRAFT ESTABLISHMENT
FARNBOROUGH, HANTS
U.K.

GEN. HOMER S. LONG, JR. (RET.)
EDUCATIONAL TESTING SERVICE
1825 EYE STREET, N.W., SUITE 475
WASHINGTON, DC 20006

"M"

MURRAY J. MACK
ATTN: PECC-CMP
U. S. ARMY CIVILIAN PERSONNEL CENTER
200 STOVALL STREET
ALEXANDRIA, VA 22332-0300

PETER W. MADEMANN, DIPL.-PSYCH.
FEDERAL ARMED FORCES RECRUITING OFFICE
KREISWEHRERSATZAMT HAMBURG
SOPHIENTERRASSE 1 A,
2000 HAMBURG 13
FRG

DR. RAY E. MAIN
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

WILLIAM P. MCALEER
U. S. ARMY
HQ, STCFA (ROK/US)
CAMP RED CLOUD
APO SAN FRANCISCO 96358-0198

CHRISTINA M. MCBRIDE
ATTN: M324
NATIONAL SECURITY AGENCY
FT. GEORGE G. MEADE, MD 20755-6000

DR. BARBARA L. MCCOMBS
UNIVERSITY OF DENVER
DENVER RESEARCH INST/SSRE DIV
2135 E. WESLEY
DENVER, CO 80208

DR. CLARENCE C. MCCORMICK
HQ US MEPCOM/MEPCT
2500 GREEN BAY ROAD
N. CHICAGO, IL 60064-3094

JEFFREY J. MCHENRY
AMERICAN INSTITUTES FOR RESEARCH
1055 THOMAS JEFFERSON ST., N.W.
WASHINGTON, DC 20007

HEATHER MCINTYRE
HUMAN FACTORS UNIT, HQ DAAC
MIDDLE WALLOP
N. STOCKBRIDGE, HAMPSHIRE
U.K.

DR. DONALD H. MCLAUGHLIN
AMERICAN INSTITUTES FOR RESEARCH
1791 ARASTRADERO ROAD
PALO ALTO, CA 94302

KAROL MCMILLAN
DIRECTORATE OF TRAINING & DOCTRINE
STAFF & FACULTY DEVELOPMENT DIVISION
U. S. ARMY AIR DEFENSE ARTILLERY SCHOOL
FT. BLISS, TX 79912

DR. ALBERT MELTER
PERSONALSTAMMAMT DER BUNDESWEHR
MUDRA-KASERNE
KOLNER STR 262
D-5000 KOELN 90
FRG

DR. JAMES W. MILLER
QUESTAR DATA SYSTEMS, INC.
7146 SHADY OAK ROAD
EDEN PRAIRIE, MN 55344

LCDR ANTHONY E. MIZEN, R.N.
ROYAL NAVAL SCHOOL OF EDUC & TRNG TECH
H.M.S. NELSON
PORTSMOUTH, HAMPSHIRE PO1 3HH
U.K.

AMELIA E. MOBLEY
COMMANDANT (G-PMR-5)
U. S. COAST GUARD
2100 SECOND STREET, S.W.
WASHINGTON, DC 20593-0001

DR. VALERIE MORRIS
ARMY PERSONNEL PSYCHOLOGY DIVISION
c/o RAE
FARNBOROUGH, HANTS
U.K.

JOHN D. MORROW
AIR UNIV CENTER FOR PROF DEV
AIS/EDM
MAXWELL AFB, AL 36112-5553

C. JILL MULLINS
CHIEF OF NAVAL EDUCATION & TRNG
NROTC RESEARCH & EVALUATION
CHET CODE N-313
NAS PENSACOLA, FL 32508

ILSE MUNRO
U.S. ARMY RESEARCH INSTIT OF ENV MEDICINE
NATICK, MA 01760-5007

"N"

DR. PETER F. NEWTON
RESEARCH DIRECTOR
NATIONAL SECURITY AGENCY
BOX 1, FANX3
FT. GEORGE G. MEADE, MD 20755-6000

"O"

DR. LAUREL W. OLIVER
ATTN: PERI-RL
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA VA 22333-5600

DR. DARLENE M. OLSON
ATTN: PERI-RS
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

CDR JERROLD E. OLSON, USN
ADVANCEMENT IN RATING DEPT HEAD
ATTN: CODE 03
NAVAL ED & TRNG PROGRAM MGT SUPP ACTIVITY
PENSACOLA, FL 32509-5000

DR. PHILIP K. OLTMAN
RESIDENT PSYCHOLOGIST
MAIL STOP: 17R
EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON, NJ 08541

DAVID J. OWEN
NATIONAL DEFENCE HQ
101 COLONEL BY DRIVE
OTTAWA, ONTARIO
CANADA K1A OK2

"P"

GLENN E. PETERS
ATTN: M324
NATIONAL SECURITY AGENCY
FT. GEORGE G. MEADE, MD 20755-6000

DR. NORMAN G. PETERSON
PERSONNEL DECISIONS RESEARCH INSTIT
43 MAIN STREET, S.E. - SUITE 405
MINNEAPOLIS, MN 55414

LTCOL FRANKLIN C. PINCH
ATTN: DIR PERSNL SELEC RESRCH & 2ND CAREER
NATIONAL DEFENCE HQ
101 COLONEL BY DRIVE
OTTAWA, ONTARIO K1A OK2
CANADA

DR. REBECCA M. PLISKE
ATTN: PERI-RP
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVE
ALEXANDRIA, VA 22333-5600

BRIAN PRITCHARD
TRC OXFORD (FOR APRE)
HEATH FARMHOUSE
COTTISFORD, N. BRACKLEY
NORTHANTS
U.K.

LCOL TERRY J. PROCIUK
APPLIED RESEARCH UNIT
CANADIAN FORCES PERSONNEL
4900 YONGE STREET, SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

DR. ALBERT P. PRUETT
NAVAL AIR MAINTENANCE TRAINING GROUP
NAS MEMPHIS
MILLINGTON, TN 38054

SQN LDR BRIAN N. PURRY, RAF (Ret)
c/o BRIAN N. PURRY & ASSOCIATES
51 THRAPSTON ROAD
BRAMPTON, HUNTINGDONSHIRE
U.K. PE18 8TB

DR. KLAUS J. PUZICHA c/o
BUNDESWEHRVERWALT
NGSAMT
BONNER TALWEG 177
5300 BONN 1
FRG

"Q"

RONALD J. QUAYLE
NATIONAL COMPUTER SYSTEMS
1101 30TH ST. N.W., SUITE 500
WASHINGTON, DC 20007

"R"

DR. DOUGLAS L. RACHFORD
ATTN: PERI-RL
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

DR. J. CLARK RAMSAUR
SHREVEPORT MILITARY ENTRANCE PROC STA
200 N. THOMAS DRIVE
SHREVEPORT, LA 71137-7623

DR JOSEPHINE M. RANDEL
MAN.TECH MATHETICS CORPORATION
8525 GIBBS DRIVE, SUITE 300
SAN DIEGO, CA 92123-1737

DR. BARRY J. RIEGELHAUPT
HUMAN RESOURCES RESRCH ORG (HumRRO)
1100 SOUTH WASHINGTON STREET
ALEXANDRIA, VA 22314

DR. MYRON A. ROBINSON
DEPT. 443 - ELECTRIC BOAT DIVISION
GENERAL DYNAMICS CORPORATION
EASTERN POINT ROAD
GROTON, CT 06340

GWYN N. ROBSON
SPECIAL PROGRAMS DEPT
MARINE CORPS INSTITUTE
ARLINGTON, VA 22222-0001

COL. LEWIS F. ROGERS, USMC
BOQ, ROOM 128, BLDG. S-599
NAS, MFS
MILLINGTON, TN 38054

KENDALL L. ROOSE
ACADEMIC TRAINING DEPARTMENT
TRAINING AIRWING FIVE
NAVAL AIR STATION WHITING FIELD
MILTON, FL 32570

SHIRLEY G. ROSCOE
6 CAMBRIDGE CT. WEST
OLD SAYBROOK, CT 06475

PAUL ROSENFELD
ATTN: CODE 623
NAVY PERSONNEL R&D CENTER
POINT LOMA
SAN DIEGO, CA 92152-6800

DR. HENDRICK W. RUCK
AIR FORCE HUMAN RESOURCES LAB / IDT
BROOKS AFB, TX 78235

MICHAEL G. RUMSEY
ATTN: PERI-RS
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

"S"

WILLIAM A. SANDS
ATTN: CODE 63
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

DR. AMIEL T. SHARON
ATTN. SEA 0721C
NAVAL SEA SYSTEMS COMMAND
WASHINGTON, DC 20362

DR. NORMAN P. SHERWOOD
ATTN: RC-PP
ARMY RECRUITING CMD
HQ, USAREC
FT. SHERIDAN, IL 60037

BARBARA SHUKITT
U.S. ARMY RESEARCH INSTIT OF ENV MEDICINE
KANSAS STREET
NATICK, MA 01760-5007

DR. NORMAN M. SHUMATE
ATTN: ATSB-DOTD
DIRECTORATE OF TRAINING & DOCTRINE
U. S. ARMY ARMOR SCHOOL
FT. KNOX, KY 40121-5200

MAJ. LEVON SIMMONS, Ph.D.
MANPOWER, PERSONNEL & TRNG (MPT)
ASD/ALH
WRIGHT-PATTERSON AFB
DAYTON, OH 45433

CAPT MELINDA D. SIMS
AIR FORCE MILITARY PERSONNEL CTR
HQ AFMPC / DPMYOT
RANDOLPH AFB, TX 78150-6001

LOYD D. SINGLETARY
ATTN: CODE N-3121
CHIEF OF NAVAL EDUCATION & TRNG
NAS PENSACOLA, FL 32508-5100

ELIZABETH P. SMITH
ATTN: PERI-RS
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

DR. MARGARET J. SMITH
NETPMSA - CODE 301
PENSACOLA, FL 32509

KAREN SPATH
ATTN: ATIC-ITT
U. S. ARMY TRAINING SUPPORT CENTER
FT. EUSTIS, VA 23604-5206

MORRIS S. SPIER
SCHOOL OF HUMAN BEHAVIOR
U. S. INTERNATIONAL UNIVERSITY
10455 POMERADO ROAD
SAN DIEGO, CA 92064

MICHAEL R. STALEY
MAXIMA CORP.
8301 BROADWAY, SUITE 212
SAN ANTONIO, TX 78247

PAUL P. STANLEY II
USAFOMC / OMD
RANDOLPH AFB, TX 78150

DR. FRIEDRICH W. STEEGE
FEDERAL MINISTRY OF DEFENSE
BMVG - P II 4
POSTFACH 13 28
D - 5300 BONN 1
FRG

DR. ALMA STEINBERG
ATTN: PERI-RL
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

CHARLES R. STEWART III
ATTN: ATSI-ES-A
INTELLIGENCE CENTER & SCHOOL
FT. HUACHUCA, AZ 85613-7000

DR. LAWRENCE J. STRICKER
EDUCATIONAL TESTING SERVICE
PRINCETON, NJ 08541

LTCOL WILLIAM J. SUBLETTE
ATTN: CODE TDA-20
COMMANDANT OF THE MARINE CORPS
ARLINGTON, VA 20380-0001

WARREN SWANSON
ATTN: CODE 019
U. S. NAVAL SUBMARINE SCHOOL
BOX 700
GROTON, CT 06349

"T"

JOESPH S. TARTELL
U. S. AIR FORCE OCCUPATIONAL MEASMT CTR
USAFOMC / OMY
RANDOLPH AFB, TX 78150-5000

CAPT R. D. TETZ
COMBAT TRAINING CENTRE
CANADIAN FORCES BASE GAGETOWN
OROMMOCTO, NEW BRUNSWICK
CANADA EOG 2PO

NORMAN E. THRASH
ATTN: CODE N-3123
CHIEF OF NAVAL EDUCATION & TRNG
NAS PENSACOLA, FL 32508-5100

DR. PAUL TWOHIG
ATTN: PERI-RL
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22310-5600

"U"

DR. GEORGE M. USOVA
110 NELSON DRIVE
NEWPORT NEWS, VA 23601

"V"

CAPT KENNETH W. VAIL
COMBAT TRAINING CENTRE
CANADIAN FORCES BASE GAGETOWN
OROMOCTO, NEW BRUNSWICK
CANADA EOG 2PO

DR. PAUL P. VAN RIJN
ATTN: PERI-RL
U. S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333-5600

PAUL R. VAUGHAN
ATTN: ATIC-ITP
INDIVIDUAL TRNG EVAL DIRECTORATE
FT. EUSTIS, VA 23604

LAWRENCE A. VEACH
ATTN: ATSK-TC
U.S. ARMY ORD MISL & MUN CTR &
SCHOOL
REDSTONE ARSENAL, AL 35897-6000

SUSAN VOLDAHL
CENTRAL SECTION
MILITARY ENTRANCE PROCESSING
COMMAND
28 GREEN BAY ROAD
N. CHICAGO, IL 60064

"W"

DR. LLOYD W. WADE
SPECIAL PROGRAMS DEPT
MARINE CORPS INSTITUTE
ARLINGTON, VA 22222-0001

DR. HOWARD WAINER
RESEARCH AND STATISTICS GROUP
EDUCATIONAL TESTING SERVICE (21T)
PRINCETON, NJ 08541

DR. RAYMOND O. WALDKOETTER
ATTN: ATSG-DSS
BUILDING 1
U. S. ARMY SOLDIER SUPPORT CENTER
FT. BENJAMIN HARRISON, IN 46216-5060

LEE WALKER
DDL OMNI ENGINEERING
RT. 1, BOX 593
PURCELLVILLE, VA 22132

GWENN E. WALTZ
ATTN: CODE N1
NAVAL ED & TRNG SUPPORT CTR, PACIFIC
SAN DIEGO, CA 92132

DR. THOMAS A. WARM
U. S. COAST GUARD INSTITUTE
P.O. SUBSTATION 18
OKLAHOMA CITY, OK 73169-6999

DR. BRIAN K. WATERS
HUMAN RESOURCES RESEARCH ORG (HumRRO)
1100 SOUTH WASHINGTON ST
ALEXANDRIA, VA 22314

STEPHEN B. WEHRENBERG
COMMANDANT (G-P-1/2)
U. S. COAST GUARD
2100 SECOND STREET, S.W.
WASHINGTON, DC 20593-0001

MAJ. KAROL W. WENEK
CANADIAN FORCES PERSNL APPL RESRCH UNT
4900 YONGE STREET, SUITE 600
WILLOWDALE, ONTARIO
CANADA M2N 6B7

ELIZABETH R. WILBUR
ATTN: CODE 63
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

ROBERT WILKES
CASPER COLLEGE
CASPER, WY 82601

CAPT EARL WILKINS
CANADIAN FORCES TRNG DEVLPMT CENTRE
CANADIAN FORCES BASE BORDEN
BORDEN, ONTARIO
CANADA L0M 1C0

CAPT MICHAEL S. WILLIAMS
USAFA/DFBL
U. S. AIR FORCE ACADEMY
COLORADO SPRINGS, CO 80840

DR. MICHAEL J. WILSON
WESTAT, INC.
1650 RESEARCH BLVD
ROCKVILLE, MD 20850

ROBERT J. WILSON
NMPC DET NODAC
WASHINGTON NAVY YARD, BLDG 150
ANACOSTIA
WASHINGTON, DC 20374-1501

DR. WINFORD D. WIMMER
U. S. ARMY CHEMICAL SCHOOL
638 BRENTWOOD DRIVE
ANNISTON, AL 36206

LAURESS WISE
AMERICAN INSTITUTES FOR RESEARCH
1055 THOMAS JEFFERSON ST, N.W. - SUITE 200
WASHINGTON, DC 20007

DR. MARTIN F. WISKOFF
ATTN: CODE 06
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152-6800

DR. BOB G. WITMER
U. S. ARMY RESEARCH INSTITUTE
FT. KNOX FIELD UNIT
STEELE HALL
FT. KNOX, KY 40121-5620

GREGG J. WRIGHT
BOOZ, ALLEN & HAMILTON INC.
7315 WISCONSIN AVE - 1100 W
BETHESDA, MD 20814

"Y"

OTHALENE T. YOUNG
HQMC - TRAINING DEPT (TPI)
WASHINGTON, DC 20380

"Z"

TIMOTHY C. ZELLO
ATTN: ATSL-DES-E
U.S. ARMY ORDNANCE CENTER & SCHOOL
ABERDEEN PROVING GROUND, MD 21005

DR. RAY A. ZIMMERMAN
THE BDM CORPORATION
2600 GARDEN ROAD, NORTH BLDG
MONTEREY, CA 93940

COL. JERRY A. ZYPCHEN
ATTN: DMOS
CANADIAN FORCES - NDHQ
2781 SPRINGLAND DRIVE
OTTAWA, ONTARIO
CANADA K1V 9X2

AUTHOR LIST

AUTHOR	PAGE
ALBA, P.A.	443
ARABIAN, J.M.	348
ARTH, T.O.	301
ASHWORTH, S.	556
ATWOOD, N.K.	242
AUSTIN, J.S.	516
BAKER, H.G.	381,443,448,470
BANDERET, L.E.	425,568,586,592
BARNETT, E.G.	101
BART, W.M.	396
BITTNER, A.C. Jr.	218
BLACKHURST, J.L.	448
BORMAN, W.C.	419
BOTHWELL, C.	369
BRITTAIN, C.V.	66
BROSVIC, G.M.	89,95
BROWN, G.	337
BURKE, E.F.	320
BURSE, R.L.	425,592
CAMPBELL, C.H.	231,454,544
CAMPBELL, J.P.	544,550
CAMPBELL, R.C.	454
CANTOR, J.A.	119
CARROLL, L.	498
CELESTE, J.F.	54
CONNER, H.B.	307
CORY, C.H.	295
CROHN, E.A.	425,592
CYMERMAN, A.	425,592
CZARNOLEWSKI, M.Y.	19
DAVIS, D.	178
DAY, L.E.	167
DICKSON, A.M.	277
DILG, M.	194
DILLA, B.L.	522
DILLON, R.F.	365
DOHERTY, L.	498
DONOFRIO, R.M.	84
DOYLE, E.L. Jr.	454
DRAKELEY, R.J.	267
DUNLAP, W.P.	218,225
EDWARDS, D.S.	459
ELIG, T.W.	25,624
ELLIOTT, S.J.	72,354
ELLIS, J.A.	143
EVANS, R.M.	504
FEGGETTER, A.J.W.	437
FINE, B.J.	401
FISHER, G.P.	254
FOLCHI, J.S.	618

AUTHOR	PAGE
FORD, P.	314,465
GARCIA, S.K.	343
GAST, I.F.	200
GILBERT, A.C.F.	89
GILROY, C.L.	48
GIVEN, K.C.	387
GOEHRING, D.J.	248
GOLDMAN, L.A.	375
GOUGE, J.A.	522
GRISSMER, D.W.	540
HAETTIG, J.H.	431
HANSER, L.M.	550
HARRIS, J.H.	42,544
HAWRYSH, F.J.	331
HERMAN, J.	242
HERTZBACH, A.	624
HETTER, R.D.	13
HILLER, J.H.	242
HOFFMAN, R.G.	237,314,465
HOLLAND, J.L.	381
HOLZ, R.F.	413
HORNE, D.K.	48
HOUGH, L.	254,556
HUNTER, F.T.	212,630
ILLES, J.W.	273
JOHNSON, D.M.	131
JONES, A.	486
JONES, K.	369
JONES, M.B.	218,225
JONES-JAMES, G.	612
KANTOR, J.	498
KEMERY, E.R.	225
KENNEDY, R.S.	218,225,568
KERR, R.H.	101
KIECKHAEFER, W.F.	167
KIMMEL, M.J.	108
KNAPP, D.J.	1
KNIRK, F.G.	143
KOBRICK, J.L.	401
LANE, N.E.	568
LATIMER, S.L.	137
LIEBERMAN, H.R.	574
LILIENTHAL, R.A.	254
LISTER, S.G.	172
LOCKHART, J.M.	131
LONGMIRE, K.M.	313
MADEMANN, P.W.	492
MAIN, R.E.	188
MARTIN, C.J.	19
MASON, J.K.	348

AUTHOR LIST

AUTHOR	PAGE
MATTSON, J.D.	480
McCOMBS, B.L.	160
McDONALD, B.	143
McHENRY, J.J.	42,562
McINTYRE, H.M.	437
McLAUGHLIN, D.H.	325
McMENEMY, J.P.	331
MELTER, A.H.	528
MORRIS, V.	113
MORTENSON, L.	437
MUNRO, I.	580
NOGAMI, G.Y.	108,540
NORRIS, L.	474
OLIVER, L.W.	639
OLSON, D.M.	419
OPPLER, S.M.	42,562
OWEN, D.J.	260
PETERS, G.E.	393
PETERSON, N.G.	556
PHALEN, W.J.	313
PLISKE, R.M.	1,25
PRITCHARD, B.	437
PUZICHA, K.J.	492
RACHFORD, D.L.	206
RADTKE, P.	459
RAFACZ, B.A.	606
RANDEL, J.M.	359
RAUCH, T.M.	580
REZNICK, R.K.	365
ROBERTS, D.E.	425,592
ROGERS, L.F.	182
ROMANCZUK, A.P.	36
ROSENFELD P.	498
ROSSMEISSL, P.G.	289,562
RUMSEY, M.G.	231
SANDS, W.A.	598
SEGALL, D.G.	13
SHORT, L.O.	60
SHUKITT, B.L.	425,586,592
SIEBOLD, G.L.	89
SKINNER, M. J.	301
SMITH, A.L. Jr.	155
SMITH, E.P.	289,534
STEEGE, F.W.	7
STEINBERG, A.G.	212
TAYLOR, B.	143
THOMAS, M.	498
THOMSON, M.W.	84
TOQUAM, J.	556

AUTHOR	PAGE
TREMBLE, T.R.	89,95
TWOHIG, P.	636
USOVA, G.M.	150,283
VAIL, K.W.	78
VAN RIJN, P.P.	212,510
VAUGHAN, P.R.	66
WALDKOETTER, R.O.	407
WALKER, C.B.	534
WALKER, L.	119
WANOUS, J.P.	470
WEGNER, T.G.	60
WEISSMULLER, J.J.	313
WHITE, L.A.	200
WILBUR, E.R.	601
WILKES, R.L.	568
WILLIAMS, M.S.	516
WILLIAMS-MORRIS, R.	396
WILSON, M.J.	30,36,54
WISE, L.	550,562
WITMER, B.G.	125
WOOD, F.R.	516
WORSTINE, D.A.	375
ZIMMERMAN, R.A.	206