DTIC FILE COPY

(1)

Technical Report 660

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1984 Fiscal Year

Newell K. Eaton, Marvin H. Goer,
James H. Harris, and Lola M. Zook

AD-A178 944

DTIC
ELECTE
APR 0 6 1987
S D
D

Selection And Classification Technical Area
Manpower and Personnel Research Laboratory



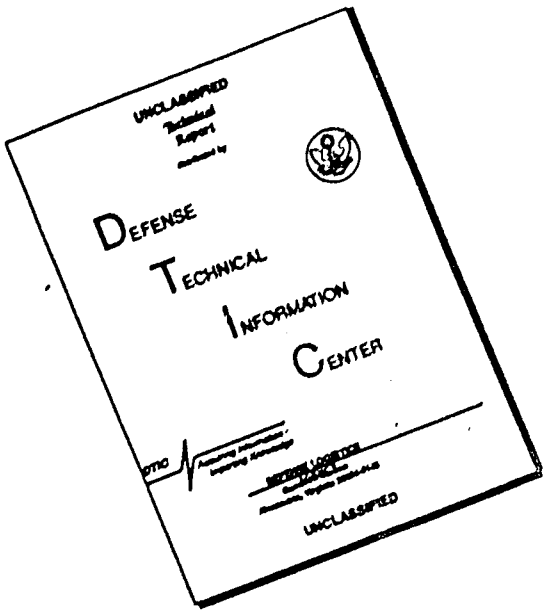
U. S. Army
Research Institute for the Behavioral and Social Sciences

July 1985

Approved for public release; distribution unlimited.

87 4 3 052

DISCLAIMER NOTICE



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

BLANK PAGES
IN THIS
DOCUMENT
WERE NOT
FILMED

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

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
COL, IN
Commanding

Research accomplished under contract
to the Department of the Army

Human Resources Research Organization

Technical review by:

Paul G. Rossmeissl
Arthur C.F. Gilbert

DTIC
COPY
INSPECTED
6

Accesion For
NTIS CRA&I
DTIC TAB
Unannounced
Justification
By
Distribution /
Availability Codes
Dist Avail and / or Special
A-1

NOTICES

DISTRIBUTION Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POT, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARI Technical Report 660	2. GOVT ACCESSION NO.: AD-A178 944
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL: ANNUAL REPORT, 1984 FISCAL YEAR	5. TYPE OF REPORT & PERIOD COVERED: Annual Report, 1 October 1983 - 30 September 1984
	6. PERFORMING ORG. REPORT NUMBER: --
7. AUTHOR(s): Newell K. Eaton, Marvin H. Goer, James H. Harris, & Lola M. Zook, Editors	8. CONTRACT OR GRANT NUMBER(s): MDA903-82-C-0531
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Human Resources Research Organization, 1100 So. Washington Street, Alexandria, VA 22314	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q263731A792
11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333-5600	12. REPORT DATE: July 1985
	13. NUMBER OF PAGES: 480
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): --	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

The Army Research Institute technical point of contact is Dr. Newell K. Eaton. His telephone number is (202) 274-8275.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Validation, Predictor measures, Construct validation, Longitudinal data base, Job knowledge tests, Validity generalization, Performance measures, Predictors, Criterion measures, Army-wide measures,

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes in detail the research performed during the second year of a project to develop a complete personnel system for selecting and classifying all entry-level enlisted personnel. Its purpose is to document, in the context of the annual report, a variety of technical papers associated with the project. In general, the second year's activities have emphasized the evaluation of the validity and fairness of existing prediction and criterion measures, and the development and initial testing of improved methods

(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ARI Technical Report 660

20. (Continued)

of predicting and measuring performance.) Research reports associated with this work are included; appendix material for certain reports is provided in ARI Research Note 85-14. A synopsis of the year's work is contained in ARI Research Report 1393.

FEB 19

Technical Report 660

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1984 Fiscal Year

Newell K. Eaton, Marvin H. Goer,
James H. Harris, and Lola M. Zook

Selection And Classification Technical Area
Newell K. Eaton, Chief

Manpower and Personnel Research Laboratory
Joyce L. Shields, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

July 1985

Army Project Number
2Q263731A792

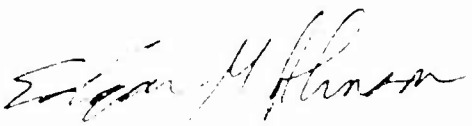
Manpower and Personnel

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

This document describes the second year of research on the Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. The thrust for the project came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training and performance. The portion of the effort described herein is devoted to the development and validation of Army Selection and Classification Measures, and referred to as "Project A." Another part of the effort is the development of a prototype Computerized Personnel Allocation System, referred to as "Project B." Together, these Army Research Institute efforts, with their in-house and contract components, comprise a major program to develop a state-of-the-art, empirically validated system of personnel selection, classification, and allocation.



EDGAR M. JOHNSON
Technical Director

PREFACE

This is a report of the second year of research conducted on Project A, "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel." The project addresses the 675,000-person enlisted personnel system of the U.S. Army, with several hundred different occupations, from infantryman to typist to medic to mechanic. The goal is a computerized personnel allocation system to match available personnel resources with Army manpower requirements, based on biographical, psychological, and performance measures, and a firm quantification of their interrelationships.

The research is being accomplished by one team of researchers addressing predictor and performance measures and their interrelationships, and by a second team using those measures to develop an allocation system (efforts in these areas have been termed "Project A" and "Project B", respectively).

The planning for this research was initiated by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) in 1980. As in-house resources were evaluated, it became apparent that the massive scope of the effort would be best met by a combination of the talents of research scientists and managers from ARI as well as contract research organizations. In 1981 ARI in-house scientists set to work developing the basic research requirements for the effort.

In 1982 a consortium, led by the Human Resources Research Organization (HumRRO), and including the American Institutes for Research (AIR) and the Personnel Decisions Research Institute (PDRI), was selected by ARI as the contract organization offering the most innovative and creative approaches to meet the objectives of Project A. Scientists from ARI and the consortium, together with a multitude of advisors, developed a research plan to guide the project (U.S. Army Research Institute Research Report 1332, May 1983). The present report describes the second year of research conducted according to that plan, with elaborations and changes outlined in the following chapters.

Each chapter of this report describes the efforts of many scientists in the consortium and ARI. Papers and reports based on their efforts are provided in this document unless they have been previously published separately. In addition to the many other scientists who have contributed to this effort, special recognition needs to be given to Dr. Joyce L. Shields. Without her vision in planning the project, ability to communicate its needs to those involved, and encouragement to all project staff, the project would not exist today.

With the conclusion of the second year of the project, we are well on our way toward meeting our goals. We are on schedule, and are prepared to meet the major challenge of the third year: a concurrent validation of our measures with 12,000 soldiers. It is our desire that the project continue to evolve and prosper over the years through continued healthy discourse among the Army's senior leadership, representatives of the Department of Defense and

the Joint Services, the scientific community, and the ARI and contractor scientists. Our aims are to provide the Army with a greatly improved, empirically based personnel system responsive to the needs of the service, while considering the unique abilities, interests, and desires of individual soldiers, and to substantially enhance scientific knowledge in applied personnel selection and classification research.

Newell Kent Eaton

NEWELL KENT EATON
ARI Principal Scientist and COR

CONTENTS

		Page
I.	Introduction to Current Army Selection and Classification Research. . . . . Newell K. Eaton, Marvin H. Goer, and Lola M. Zook	1
II.	School and Job Performance Measurement . . . . . John C. Campbell	27
III.	Predictor Measurement . . . . . Norman G. Peterson	203
IV.	Validation . . . . . Paul G. Rossmeissl and Lauress L. Wise	311
V.	Status and Future Directions of Army Selection and Classification Research . . . . . John P. Campbell and Newell K. Eaton	467

LIST OF FIGURES

1	The Army's Personnel System	2
2	Matching Personnel to Needs	3
3	Improved Personnel Management System.	4
4	The Research Flow	5
5	Project A MOS	6
6	MOS Clusters	7
7	Predictor Constructs Under Consideration for Administration to FY83/84 Cohort in FY85	9
8	30 MOS 71L Tasks Selected for Testing	11
9	Project A Performance Categories vs. EER Categories	12
10	Leading and Supporting	13
11	Predictive Validities for FY81/82 Soldiers.	15
12	Examples of MOS Using the CL or SC Composites	15

CONTENTS (Continued)

LIST OF FIGURES

		Page
13	Governance Advisory Group	20
14	Project Organization	20
15	Three Alternative Scenarios for SME Judgments of Task and Item Importance	37
16	Flow Chart of Predictor Measure Development Activities of Project A	204
17	Cognitive/Perceptual Psychomotor Measures in the Pilot Trial Battery	207
18	Non-Cognitive Measures in the Pilot Trial Battery: The Army Vocational Interest and Career Examination (AVOICE) and the Assessment of Background and Life Experiences (ABLE)	208
19	Validity Analyses Sample Sizes	314
20	Predictive Validities Systems for Nine and Four Composites	315
21	A Comparison of Current and Alternative Composites	316

LIST OF TABLES

1	Mean (SD) of Mean Estimated Validities of Predictor Factors for Criterion Factors	8
2	Mean (SD) Intercorrelations of Cognitive/Spatial Paper-and-Pencil, Non-Cognitive Paper-and-Pencil, and Perceptual/Psychomotor Computerized Measures in the Project A Pilot Trial Battery	10
3	Estimates of SD$ and Examples of Utility	16
4	Scale Values of MOS/Performance Level Hypothetical Soldiers	17
5	"Batch A" Field Test Samples	33
6	FY83/84 Soldiers with Preliminary Battery and Training Data	313

CONTENTS (Continued)

Page

Associated Reports and Papers

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report (ARI Research Report 1347); Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, Army Research Institute 23

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Technical Appendix to the Annual Report (ARI Research Note 83-37); Newell K. Eaton and Marvin H. Goer (Editors) 24

Development and Validation of Army Selection and Classification Measures--Project A: Longitudinal Research Database Plan (ARI Research Report 1356); Lauress L. Wise and Ming-mei Wang (AIR), Paul G. Rossmeissl (ARI) 25

The U.S. Army Research Project to Improve Selection and Classification Decisions; Newell K. Eaton (ARI) 26

An Analysis of SQT Scores as a Function of Aptitude Area Composite Scores for Logistics MOS; Paul G. Rossmeissl and Newell K. Eaton (ARI) 41

Administrative Records as Effectiveness Criteria: An Alternative Approach; Barry J. Riegelhaupt, Carolyn DeMeyer Harris, and Robert Sadacca (HumRRO) 49

Factors Relating to Peer and Supervisor Ratings of Job Performance; Walter C. Borman (PDRI), Leonard A. White, Ilene F. Gast (ARI) 77

Relationships Between Scales on an Army Work Environment Questionnaire and Measures of Performance; Darlene M. Olson (ARI), Walter C. Borman, Loriann Roberson, Sharon R. Rose (PDRI) 99

The Cost-Effectiveness of Hands-on and Knowledge Measures; William Osborn and R. Gene Hoffman (HumRRO) 125

Personal Constructs, Performance Schema, and "Folk Theories" of Subordinate Effectiveness: Explorations in an Army Officer Sample; Walter C. Borman (PDRI) 139

CONTENTS (Continued)

Page

Development of a Model of Soldier Effectiveness; Walter C. Borman (PDRI), Stephen J. Motowidlo (The Pennsylvania State University), Sharon R. Rose (PDRI), Lawrence M. Hanser (ARI) 167

Validity of Cognitive Tests in Predicting Army Training Success; Clessen J. Martin, Paul G. Rossmeissl, Hilda Wing (ARI) 211

Expert Judgments of Predictor-Criterion Validity Relationships; Hilda Wing (ARI), Norman G. Peterson (PDRI), R. Gene Hoffman (HumRRO) 219

Covariance Analyses of Cognitive and Noncognitive Measures of Army Recruits: An Initial Sample of Preliminary Battery Data; Leaetta Hough, Marvin D. Dunnette (PDRI), Hilda Wing (ARI), Janis Houston, Norman G. Peterson (PDRI) 271

Meta-Analysis: Procedures, Practices, Pitfalls--Introductory Remarks; Hilda Wing (ARI) 307

Verbal Information Processing Paradigms: A Review of Theory and Methods (ARI Technical Report 648); Karen J. Mitchell 310

Evaluation of the ASVAB 8/9/10 Clerical Composite for Predicting Training School Performance (ARI Technical Report 594); Mary M. Weltin, Beverly A. Popelka 319

Clustering Military Occupations in Defining Selection and Classification Composites; Lauress L. Wise, Donald H. McLaughlin (AIR), Paul G. Rossmeissl (ARI), David A. Brandt (AIR) 321

Differential Validity of ASVAB for Job Classification; Don McLaughlin (AIR) 333

Complex Cross-Validation of the Validity of a Predictor Battery; David Brandt, Don McLaughlin, Laurie Wise (AIR), Paul Rossmeissl (ARI). 347

Subgroup Variation in the Validity of Army Aptitude Area Composites; Paul G. Rossmeissl (ARI), David A. Brandt (AIR) 361

CONTENTS (Continued)

Page

Validation of Current and Alternative ASVAB Area Composites, Based on Training and SQT Information on FY1981 and FY1982 Enlisted Accessions (ARI Technical Report 651); D.H. McLaughlin, P.G. Rossmeissl, L.L. Wise, D.A. Brandt, Ming-mei Wang 413

A Data Base System for Validation Research; Paul G. Rossmeissl (ARI), Lauress L. Wise, Ming-mei Wang (AIR). . . 415

The Application of Meta-Analytic Techniques in Estimating Selection/Classification Parameters; Paul G. Rossmeissl, Brian M. Stern (ARI) 423

Adjustments for the Effects of Range Restriction on Composite Validity; David Brandt, Donald H. McLaughlin, Lauress L. Wise (AIR), Paul G. Rossmeissl (ARI) 431

Alternate Methods of Estimating the Dollar Value of Performance; Newell K. Eaton, Hilda Wing, Karen J. Mitchell (ARI) 441

I. INTRODUCTION TO CURRENT ARMY SELECTION AND CLASSIFICATION RESEARCH*

Newell K. Eaton, Marvin H. Goer, and Lola M. Zook

The purpose of this annual report is to document various aspects of the technical plans and progress during the second year (Fiscal Year 1984) of work on the U.S. Army's Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. Project A is a comprehensive, long-range research program developed by the Army Research Institute for the Behavioral and Social Sciences (ARI). Our goal is a computerized personnel allocation system to match available personnel resources with Army manpower requirements, based on biographical, psychological, and performance measures and a firm quantification of their interrelationships. Project A will develop, for first- and second-tour soldiers, new predictor tests and composites, performance measures and composites, and utility values, and an empirical description of their intercorrelations. These, along with supply and demand forecasts, will be the basis for the concurrent development by Project B of the computerized allocation system.

The second of Project A's nine years has just been completed. The project employs 40-50 researchers in a variety of specialties of industrial and organizational psychology, operations research, management science, and computer science. The project addresses the 675,000-person enlisted personnel system of the U.S. Army, with several hundred different occupations, from infantryman to typist to medic to mechanic. A schematic of the project is shown in Figure 1.

Management of the U.S. Army enlisted force is one of the most complex personnel tasks in the world. Each year over 400,000 people apply for 135,000 first-tour positions in over 250 Military Occupational Specialties (MOS), and over 80,000 soldiers reenlist in about 350 different MOS. Typically, an individual is guaranteed specific job training at the time he or she signs an enlistment contract, and a specific MOS upon reenlistment. Enlistment can be up to one year prior to entering the Army. The decision to select the individual for service/reenlistment and to allocate an MOS must be made to meet the needs of the individual as well as the near-term requirements and long-range objectives of the Army.

---

* Much of this chapter is from an invited address by the first author at the 26th Annual Conference of the Military Testing Association in Munich, Federal Republic of Germany, 5-9 November 1984. It is based in part on papers and presentations by many Project A authors, and in part on a paper previously presented at the National Security Industrial Association Fourth Annual Conference on Personnel and Training Factors in Systems Effectiveness, in Springfield, Virginia, 1-3 May 1984.

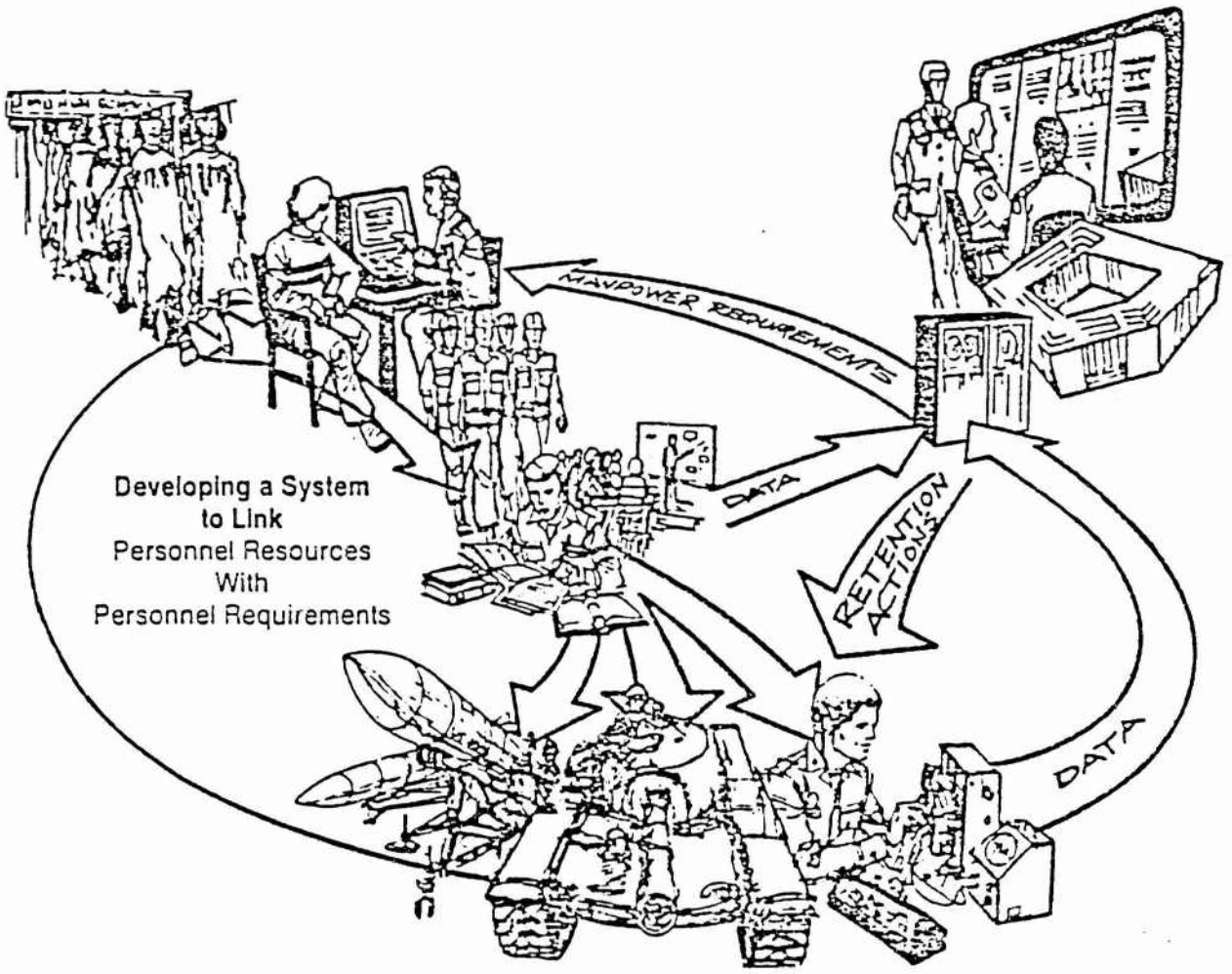


Figure 1. The Army's Personnel System

Of course, the Army is not now without tools for making such decisions. Standards are in place for initial selection and classification; they have been shown to be valid for training performance and job knowledge in many MOS. A system does exist for MOS allocation in enlistment and reenlistment. With the accomplishment of this project, however, the Army's personnel system will be far superior to existing systems, benefiting individual soldiers and the country's defense. Figure 2 shows the system as it exists, and as it will be.

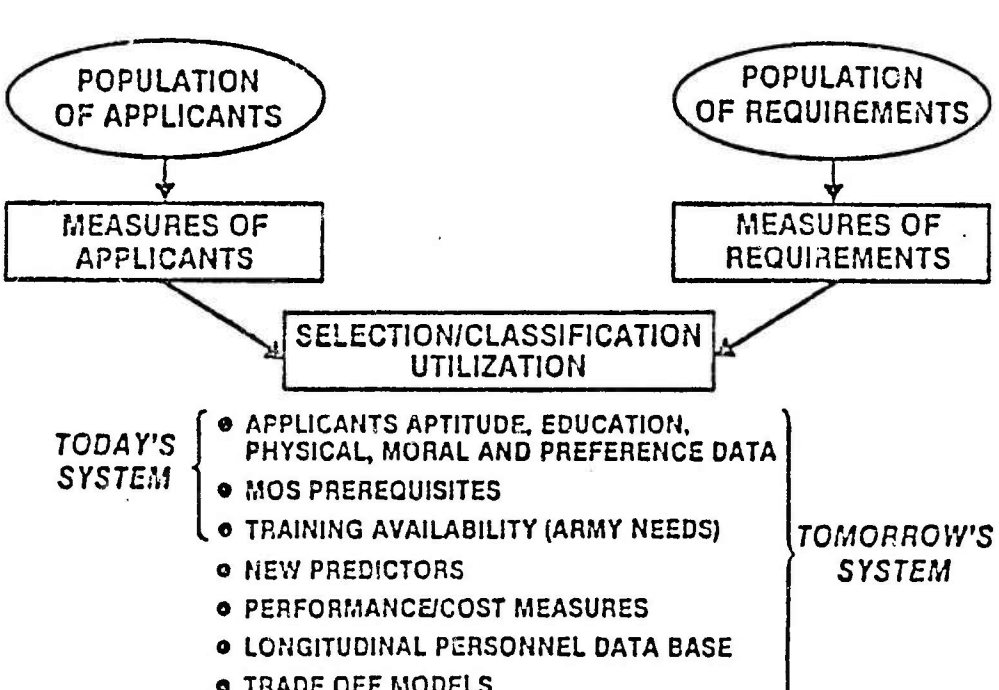


Figure 2. Matching Personnel to Needs

A major effort to develop new predictor and criterion measures is being conducted to expand the dimensionality and accuracy of measurement of the respective predictor and criterion space. At this time there appears to be a heavy general-ability (Spearman's "G") loading in both the paper-and-pencil Armed Services Vocational Aptitude Battery (ASVAB) and the current Skill Qualification Tests (SQT). This research is designed to provide measures that more completely encompass the full range of potential performance predictors and to provide criterion measures that more adequately represent actual job performance. Together, these should enable the Army to make the most valid performance predictions. Figure 3 illustrates an improved personnel management system based on a variety of better predictor and performance measures. In each MOS the most valid composite of predictors will be used as selection/classification factors to provide the best person-job match for overall soldier performance.

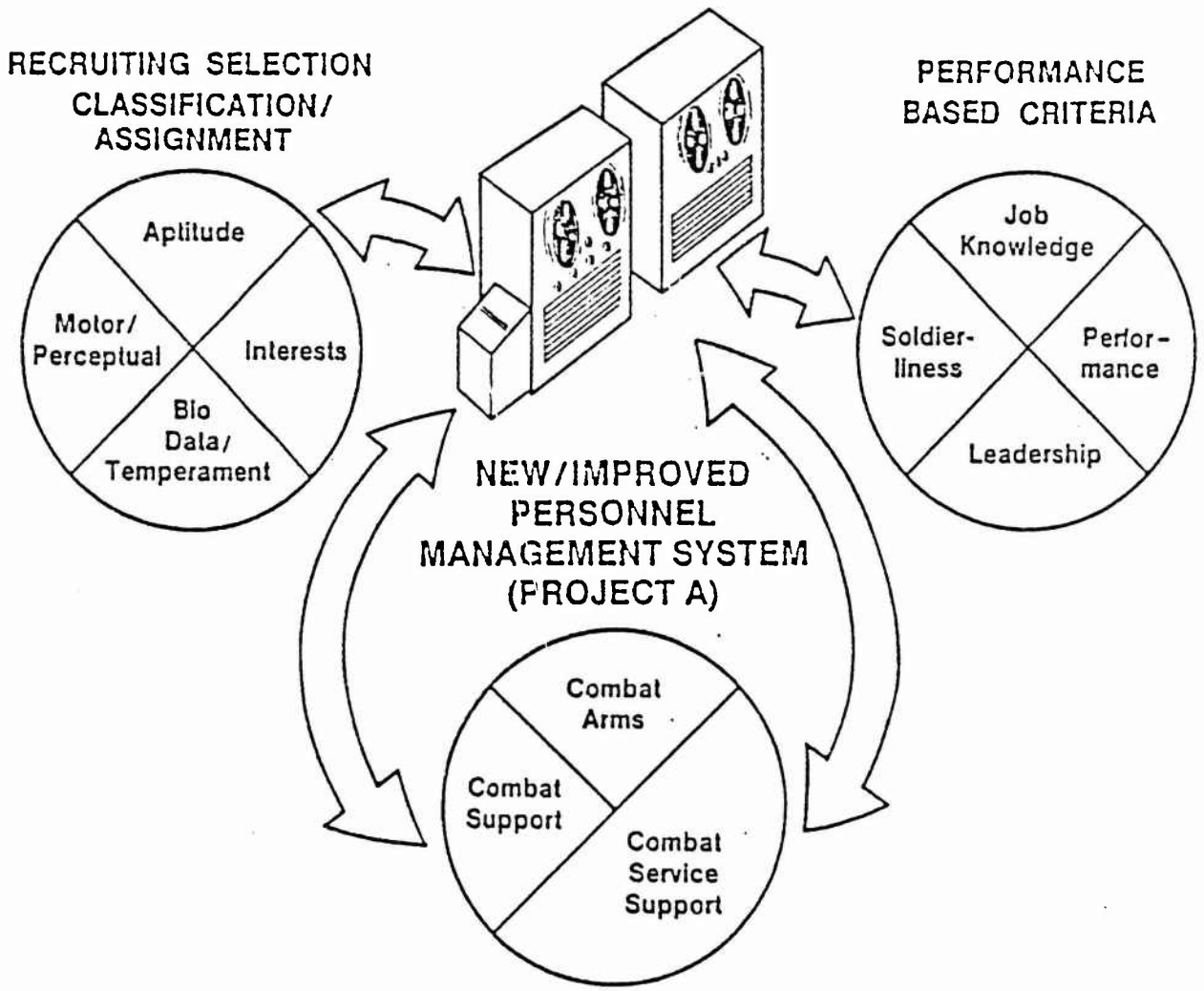


Figure 3. Improved Personnel Management System

On the basis of data from Project A as well as system design data from all services within the U.S. Department of Defense, Project B is developing a system for selecting recruits and reenlistees, determining which MOS to offer them, and providing the feedback and control system. The system will take information on the Army's requirements by MOS over the planning horizon (1 year or more), along with personnel supply forecasts, and develop an allocation plan for the current planning period (e.g., the next week). The system will support the Army guidance counselor in determining what MOS to offer

prospective recruits and will operate in near real time. Using the individual's test scores, physical profile, and preferences, the system will suggest a set of best person-job matches based on individual abilities and desires, predicted performance, and Army needs.

Research Design

The Project A research design is shown in Figure 4. A key feature of the design is its iterative nature. Data are being collected in three iterations to provide for timely and responsive results during the course of the effort, as well as to correct for errors and to take advantage of opportunities. (The research plan is described in detail in ARI Research Report 1332, May 1983.)

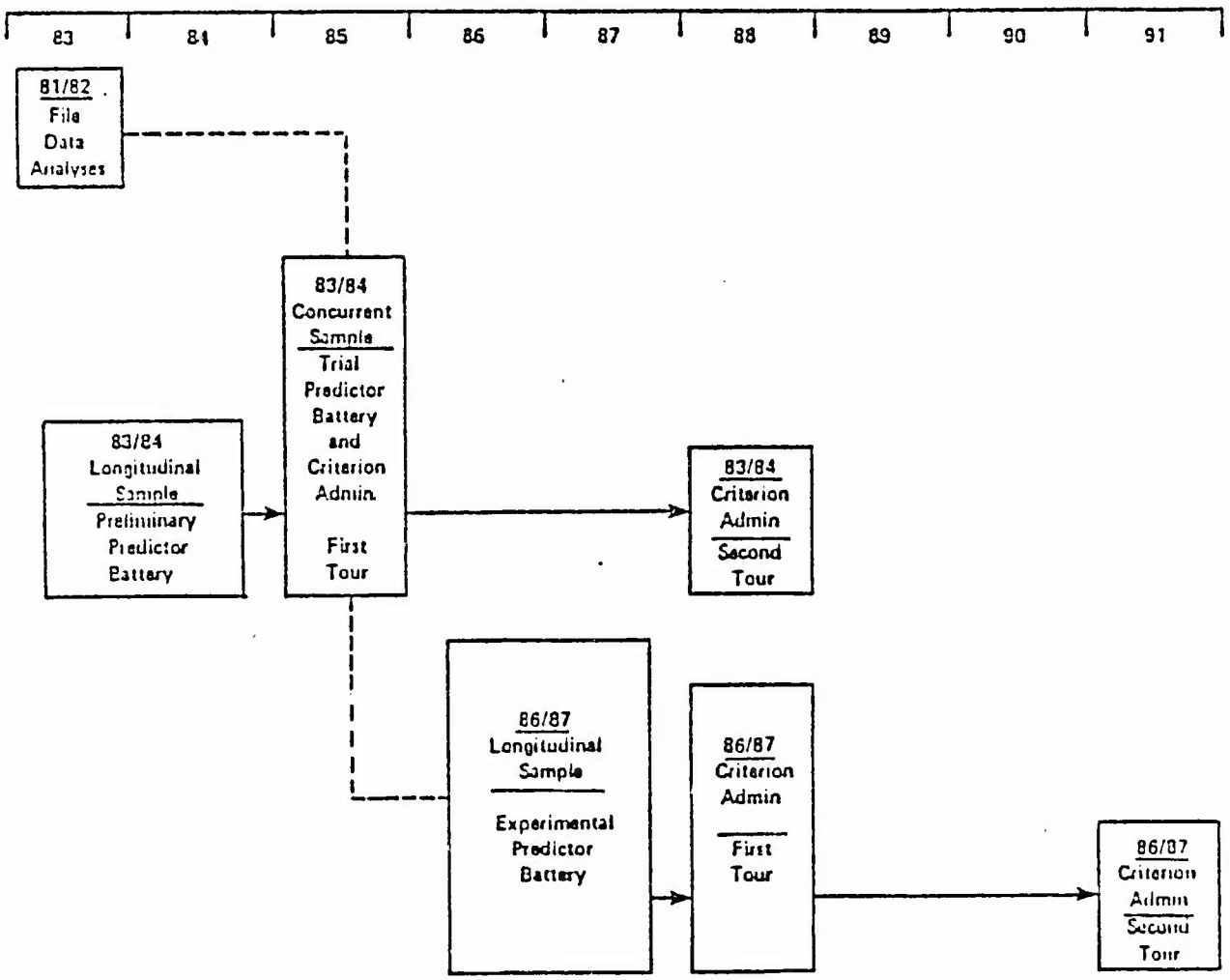


Figure 4. The Research Flow

In the first iteration, file data from accessions in fiscal year (FY) 1981 and 1982 were evaluated to verify the empirical linkage between existing ASVAB scores and subsequent training and first-tour knowledge test performance.

In the second iteration a predictive-concurrent design is being executed with FY83/84 accessions. Several thousand soldiers in four occupations have been tested at entry on a preliminary battery of spatial, perceptual, temperament/personality, interest, and biodata measures. These soldiers' data were entered into a Longitudinal Research Data Base (LRDB) containing operational ASVAB and other enlistment measures on all FY83/84 accessions.

About 600 soldiers in each of these four MOS, and in each of an additional 15 MOS, are to be tested in FY85. A revised test battery, including computer-administered perceptual and psychomotor predictor instruments, is to be concurrently administered with a set of job-specific and general performance indices based on knowledge, hands-on (for half the MOS), and rating measures. About a hundred soldiers in each MOS will be retested after three years, during their second Army tour.

The 19 MOS chosen for testing (Figure 5) comprise a specially selected representative sample of the 250 entry-level MOS. The MOS selection was based on an initial clustering of MOS, derived from rated similarities of job content. The clusters are shown in Figure 6 (see Rosse, Borman, Campbell, and Osborn, 1983). These 19 MOS account for about 45 percent of Army accessions. They permit sample sizes sufficient to empirically evaluate race and sex fairness in most MOS.

BATCH A

MOS	Title	FY83 Accessions
13B	Cannon Crewman	6431
64C	Motor Transport Oper	4282
71L	Admin Specialist	5219
95B	Military Police	5873

BATCH B

MOS	Title	FY83 Accessions
05C	Radio TT Oper	1815
11B	Infantryman	15904
19E/K	Tank Crewman	3935
63B	Vehicle & Generator Mech	4807
91B	Medical Care Specialist	4681

BATCH Z

MOS	Title	FY83 Accessions
12B	Combat Engineer	1654
16S	MANPADS Crewman	624
27E	Tow/Dragon Rpr	264
51B	Carpentry/Masonry Spec	163
54E	Chemical Operations Spec	1302
55B	Ammunition Spec	571
67N	Utility Helicopter Rpr	621
76W	Petroleum Supply Spec	1205
76Y	Unit Supply Spec	3651
94B	Food Service Spec	5375
	TOTAL	134,696

Figure 5. Project A MOS

Cluster	Title	MOS Rep
A	ELECTRONICS (NON MISSILE)	
B	MECHANICS	63B
C	WEAPONS CREWMAN	13B 16S 19E
D	RADIO OPERATOR	05C
E	SUPPLY	76W 76Y 55B
F	ELECTRONIC WARFARE	(CLUSTER A)
G	CLERICAL	71L
H	MEDICAL	91B
I	CONSTRUCTION EQUIPMENT OPERATOR	64C
J	HELICOPTER REPAIR	67N
K	MISSILE ELECTRONICS	
L	COMBAT SOLDIER	11B 12B
M	TRADES	51B
N	ARTS	
O	CSR	54E
"P"	MP	95B
"Q"	COOK	94B

Figure 6. MOS Clusters

In the third iteration all of the measures, refined by the experiences of the first and second iteration, will be collected sequentially in a true predictive validity design. About 50,000 soldiers across about 20 MOS will be included in the FY86/87 predictor battery administration. After losses from all factors, about 3500 will be included in second-tour performance measurement in FY91.

Research Activities

Predictor Development. In our predictor development the taxonomy of human abilities presented by Peterson and Bownas (1982) was used as a starting point. Based on an exhaustive literature review followed by analyses of expert judgments of predictor-criterion validity coefficients, a predictor by performance factors matrix was created. It is shown, in abbreviated form, in Table 1. In Figure 7 are shown the predictor constructs that are currently under consideration for administration to the FY 83/84 cohort in FY85. Those marked with an "A" in Figure 7 are now measured by the current ASVAB. Twelve were measured in the predictive design portion of the second design iteration, for accessions, in four MOS. Each of these 12 constructs is noted with a "P". Field tests have been completed on micro processor-based perceptual and psychomotor clusters noted with a "C". Of significant interest is the relative independence of these measures (shown in Table 2). We appear to be well on the toward extending the predictor space beyond "G". More complete reports of these data are available in Chapter III of this report (Wing, Peterson, and Hoffman, 1984; Hough, et al., 1984).

Table 1. Mean (SD) of Mean Estimated Validities of Predictor Factors for Criterion Factors

Predictor Factors	Technical Skills	Information Processing	Physical/ Combat	Personal Interaction	Commitment/ Initiative
Cognitive Abilities	.23 (.09)	.24 (.10)	.13 (.05)	.24 (.11)	.10 (.06)
Visualization/ Spatial	.24 (.08)	.13 (.03)	.14 (.04)	.14 (.05)	.07 (.03)
Information Processing	.16 (.06)	.19 (.07)	.17 (.05)	.15 (.03)	.07 (.03)
Mechanical	.21 (.12)	.10 (.06)	.18 (.07)	.10 (.07)	.10 (.04)
Psychomotor	.12 (.06)	.10 (.06)	.14 (.07)	.08 (.05)	.05 (.02)
Social Skills	.06 (.04)	.03 (.02)	.06 (.05)	.19 (.11)	.08 (.06)
Vigor	.13 (.06)	.10 (.05)	.20 (.10)	.18 (.10)	.16 (.07)
Motivation/ Stability	.15 (.07)	.16 (.07)	.15 (.07)	.18 (.09)	.28 (.10)

A		Verbal
P	C	Memory
A	C	Number Facility
A	C	Perceptual Speed and Accuracy
P		Reasoning/Induction
	C	Information Processing
P		Spatial Orientation
P		Spatial Visualization
P		Closure/Field Independence
A		Mechanical Information
	C	Multilimb Coordination
	C	Precision
	C	Movement Judgment
P		Realistic vs. Artistic Interests
P		Investigative Interests
P		Enterprising Interests
		Social Interaction
		Conventionality
P		Stress Tolerance/Adjustment
P		Dependability/Conscientiousness
P		Achievement
P		Physical Condition
P		Leadership
		Locus of Control
		Agreeableness

Note: A = Currently included in ASVAB
P = Included in predictive design portion
C = Microprocessor-based measure

Figure 7. Predictor Constructs Under Consideration for Administration to FY83/84 Cohort in 1985

Table 2. Mean (SD) Intercorrelations of Cognitive/Spatial Paper-and-Pencil, Non-Cognitive Paper-and-Pencil, and Perceptual/Psychomotor Computerized Measures in the Project A Pilot Trial Battery[a]

	Cognitive/ Spatial	Non-Cognitive	Perceptual/ Psychomotor
Cognitive/Spatial	.54 (.06)	--	--
Non-Cognitive	.09 (.07)	.35 (.18)	--
Perceptual/Psychomotor	.26 (.13)	.09 (.08)	.25 (.19)

[a] N's are approximately 110, with small variations. There are 10 cognitive measures, 37 non-cognitive measures, and 15 perceptual/psychomotor measures. Data collected at Fort Lewis, June 1984.

Performance Measurement. The work on performance measures has also developed nicely. We have prepared an extensive task inventory for the first 19 key MOS, based on Soldier's Manuals, Occupational Surveys, and data from subject matter experts. Efforts have been made to level the generality of task descriptions, and to determine the variability of performance, importance, and frequency of each task. This detailed analysis provides a firm basis for both knowledge and hands-on task sampling. Consequently, we know the degree to which our measures reflect job requirements.

Field tests have been conducted with 150 soldiers in each of the first four MOS: clerk-typist (71L), military police (95B), driver (64C), and artillery crewman (13B). Field tests for five more MOS will be completed this spring. Tests on 30 tasks representing each MOS are administered in a paper-and-pencil format; 15 are also administered in a hands-on mode. These tasks are shown in Figure 8 for MOS 71L. Ratings from peers and supervisors are also obtained on the soldier's ability to perform these tasks. Additionally, organizational variables, knowledge of information presented during training, and ratings of general soldiering behaviors are obtained during the field tests. A list of these general soldiering categories, compared to our current enlisted evaluations, is shown in Figure 9, and an example is shown in Figure 10.

HANDS-ON AND KNOWLEDGE TESTS

1. Prepare a requisition for publications
2. File documents/correspondence
3. Post regulations and directions
4. Type a joint message form
5. Type a military letter
6. Type a subsequent comment to disposition
7. Type subsequent comment to disposition
8. Type a memorandum
9. Type straight copy material
10. Type military orders
11. Receipt/transfer classified material
12. Put on M17 protective mask
13. Administer nerve agent antidote (self)
14. Put on field pressure dressing
15. Perform operator maintenance on M16A1

KNOWLEDGE TESTS ONLY

16. Establish functional files
17. Control expendable/non-expendable supplies
18. Receive, maintain, control office equipment
19. Dispatch out-going distribution
20. Assemble correspondence
21. Safeguard FOUO material
22. Load, reduce stoppage, clear M16A1
23. Perform cardiopulmonary resuscitation
24. Put on protective clothing (MOPP)
25. Determine grid coordinates on a map
26. Camouflage self and equipment
27. Determine magnetic azimuth with compass
28. Maintain M17 protective mask
29. Practice noise, light, litter discipline
30. Know rights and obligations as POW

Figure 8. 30 MOS 71L Tasks Selected for Testing

	Project A General Soldiering Performance Categories	Enlisted Evaluation Report Professionalism and Performance Categories
A.	Technical Knowledge/Skill	Demonstrates Technical Skills
B.	Initiative/Effort	Demonstrates Initiative
C.	Following Regulations/Orders	
D.	Integrity	Integrity, Loyalty, Moral Courage
E.	Leading and Supporting	Develops Subordinates, Earns Respect, Attains Results, Supports EO/EEO
F.	Maintaining Assigned Equipment	
G.	Maintaining Living/Work Areas	
H.	Military Appearance	Military Appearance
I.	Physical Fitness	Physical Fitness
J.	Self-Development	Seeks Self-Improvement
K.	Self-Control	Self-Discipline, Adapts to Changes Performs Under Pressure
		Displays Sound Judgment
		Communicates Effectively

Figure 9. Project A Performance Categories vs. EER Categories

Performing in leader role, as required, and providing support for fellow unit members.

1 2 Below Standard	3 4 5 Adequate/Mid-Range	6 7 Superior
• Performs poorly in leadership positions; is unable or unwilling to take charge when leadership is required in unit.	• Is able to step in to perform effectively in structured leadership situations where it is well known what's expected; is less able to perform well in difficult leadership situations requiring hard judgments, quick decisions, etc.	• Performs very effectively when placed in leadership position; takes charge when necessary to lead the unit and fills in effectively when NCO is absent, sick, injured, etc.
• Is ineffective at helping others get through a task, assignment, etc.; overlooks, ignores, or otherwise fails to pitch in to help unit members when they are in trouble, need encouragement, etc.	• When called upon, can instruct others effectively on a limited number of topics; in most situations is supportive of fellow unit members, although he/she will not go out of way to provide support, encouragement etc.	• Is good at teaching others when the opportunity arises, and skillfully shows unit members how to perform more effectively; looks out for and supports fellow unit members when they are in trouble, performing poorly, need encouragement, etc.

Figure 10. Leading and Supporting

Information obtained from the field tests, and during the FY85 cohort tests, will inform our decisions on the most efficient manner in which to construct comprehensive job performance measures. Preliminary information, from two of the first four MOS field tested, indicates relatively high internal consistency within measurement method, but relative independence between methods. We expect that the results of the field tests and cohort test will

provide strong evidence that will affect criterion development. Questions of "ultimate" criteria, and the parameters determining the relationship between hands-on, job knowledge, and peer or supervisory ratings, will be addressed. Because complete data will be available in nine diverse MOS, and partial data in 10 more, we expect to obtain relatively comprehensive answers to these questions.

Another question is how to determine minimum performance standards. We are beginning by presenting our quantitative performance distributions in proponent workshops. Both trainers and leaders in operational units will see how soldiers in their occupations performed or were rated on all the measures, and how the measures are intercorrelated. Through their individual judgments and consensual feedback procedures, we will attempt to elicit minimum performance standards for approval by Army policymakers. These will inform policymakers' decisions on acceptable predictor scores for entry into MOS.

Longitudinal Research Data Base. One of our major accomplishments is a longitudinal research data base, containing data on Army applicants beginning in FY81 and continuing through the present time. After exhaustive work with records, we have data on over 600,000 applicants and over 300,000 accessions.

Predictor information consists of operational accessions records data: ASVAB, the Military Applicant Profile for non-graduates, and some other biodata. Performance data consist of end-of-course training data reported by the schools (FY81 only), SQT data, and data from the Enlisted Master File (attrition, promotion, disciplinary actions, awards, etc.). The file also includes test data on every soldier to whom we administer our predictor or performance measures during pilot, field, or FY85 test administration.

The importance of the LRDB is based on the rapid, systematic access it offers to many kinds of data. It can provide, for many MOS, rapid answers to questions because new data do not have to be collected. Further, it is a prototype of the kind of data system that could be a powerful personnel management tool. A complete description of the LRDB can be found in Wise, Wang, and Rossmeissl, 1983.

First Iteration Completed. The first iteration of the data collection specified in the research design is complete. This included the analysis of the validity of the current ASVAB as a predictor of MOS training and first-tour SQT performance. The results were based on a sample in excess of 60,000 soldiers. They demonstrated the validity of the nine operational ASVAB composites, with a median validity of .48 for training and SQT combined. Further, the results showed that a change in the composition of two composites--CL (clerical) and SC (surveillance and communication)--produced an increase in predictive validity. The results are summarized in Figure 11, and described in detail in McLaughlin et al., 1984 (see Chapter IV). Some of the larger MOS selected by the CL and SC composites are shown in Figure 12. The Army scheduled operational use of these new composites for October 1984, improving the prediction of performance of 20,000 soldiers entering each year.

Cluster of MOS	Combined Criteria		Training Criteria		SQT Criteria	
	Current	Alternative	Current	Alternative	Current	Alternative
CL	48	56	40	47	49	58
CO	44	44	36	35	44	45
EL	47	48	40	41	45	46
FA	48	50	35	36	45	48
GM	47	48	52	52	40	46
MM	48	49	44	44	45	53
OF	48	49	35	36	50	53
SC	45	50	34	35	47	53
ST	58	58	54	54	55	56

Figure 11. Predictive Validities for FY81/82 Soldiers

CL Composite

- 71L Administrative Specialist
- 76C Equipment Records and Parts Specialist
- 76Y Unit Supply Specialist

SC Composite

- 05B Radio Operator
- 05C Radio Teletype Operator
- 72E Combat Telecommunications Center Operator

Figure 12. Examples of MOS Using the CL or SC Composites

The utility of any selection or classification effort is an important issue, and there has been a significant rebirth of interest in this area in the last five years. Using an estimation technique developed by Schmidt, Hunter, McKenzie, and Muldrow (1979), Rossmeissl* estimated the dollar value of the Army's change in the CL and SC composites to be $5,000,000 per year. We have also extended our effort toward better ways to evaluate the utility of selection and classification efforts. Recent work by Eaton, Wing, and Mitchell (see Chapter IV) provided an extension to the Schmidt et al. method which appears to be more appropriate in military settings, as well as an entirely new method. Our results with these two methods ("superior equivalents" and "systems effectiveness") are compared to those of the Schmidt et al. method (SDy estimation) in Table 3 (an $\underline{r}$ = .3 and selection ratio of .5 were assumed). Last, Sadacca and Campbell** are making substantial progress, with a utility effort designed to evaluate the relative worth of various levels of performance within and between MOS. Their pilot efforts have used the 50th percentile infantryman as a standard. Table 4 illustrates some of their first results.

Table 3. Estimates of SD$ and Examples of Utility

	n	SD$[a]	US or utility[a] per tank (Ws = 1)	US or utility[b] per system (Ws = 2.500)
SD$ Estimation Technique				
Group 1	48	$20,000	$ 4,800	$12,000,000
Group 2	40	$60,000	$14,400	$36,000,000
Superior Equivalents Technique				
Using Pay and Allowance Estimates of Y50				
Group 1	52	$26,700	$ 6,400	$16,000,000
Group 2	45	$26,700	$ 6,400	$16,000,000
Using SD$ Estimates of Y50				
Group 1	52	$26,700	$ 6,400	$16,000,000
Group 2	45	$31,100	$ 7,500	$18,700,000
System Effectiveness Technique	--	$60,000	$14,400	$36,000,000
Salary Percentage Technique	--	$12,000	$ 2,900	$ 7,200,000

[a] Rounded to nearest hundred dollars.
[b] Rounded to nearest hundred thousand dollars.

* Reported by P.G. Rossmeissl in ARI Research Highlights, June 1984.

** Reported by R. Sadacca and J.P. Campbell in a paper prepared for a briefing in October 1984.

Table 4. Scale Values of MOS/Performance Level Hypothetical Soldiers[a] (50th Percentile Infantrymen = 1.0; n = 8 Judges)

	Percentile			Scale Difference	
MOS	10	50	90	(90-50)	(50-10)
Administrative Specialist (71L)	.10	.23	.46	.23	.13
Ammunition Specialist (55B)	.17	.49	1.01	.52	.32
Carpentry & Masonry Specialist (51B)	.09	.21	.43	.22	.12
Chemical Operations Specialist (54E)	.26	.70	1.51	.81	.44
Food Service Specialist (94B)	.10	.23	.53	.20	.13
Light Wheel Veh./Power Gen. Mech. (63B)	.16	.43	.75	.32	.27
Medical Spcialist (91B)	.21	.58	1.29	.71	.37
Military Police (95B)	.17	.34	.66	.32	.17
Motor Transport Operator (64C)	.12	.37	.68	.31	.25
Petrol. Supply Specialist (76W)	.13	.31	.71	.40	.18
Radio Teletype Operator (05C)	.15	.41	.91	.50	.25
TOW/Dragon Repairer (27E)	.23	.64	1.25	.62	.41
Unit Supply Specialist (76Y)	.08	.23	.45	.22	.15
Util. Heli. Repairer (67N)	.17	.52	1.06	.54	.35
			Average	.42	.25
Infantryman (11B)	.34	1.00	2.01	1.01	.66
Armor Crewman (19E/K)	.42	1.28	2.71	1.43	.86
Cannon Crewman (13B)	.29	.75	1.53	.78	.46
Manpads Crewman (16S)	.27	.72	1.26	.54	.45
Combat Engineer (12B)	.26	.72	1.46	.74	.46
			Average	.90	.58

[a] Workshops 4 and 5

We expect that the research will result in a substantial savings and improved readiness. Ultimately we hope our utility efforts will converge, providing data in several forms. We would like to be able to talk about the results in terms of dollar benefits compared to research and implementation costs. Implementation of new predictor tests and evolution of the personnel system will be costly. Credible data will be needed upon which to base implementation decisions. But, more important, we wish to observe, and quantitatively describe, a significant return in terms of increased individual and system performance. We want to be able to discuss savings in terms of increased weapons systems effectiveness comparable to that obtained by adding weapons system units (tanks, howitzers, etc.) operated at current proficiency levels.

Outcome. The most important outcome from the research is increased performance. Together, better predictor and performance tests will substantially improve the performance of the Army in the field. Further, the research will better quantify the meaning of good and poor performance. It is also expected to greatly reduce personnel costs, and provide the Army's personnel managers with a powerful tool for evaluation and control. Overall, the system should improve the readiness of the Army, and the performance satisfaction and career opportunities of individual soldiers. We believe these gains are achieved most efficiently through a single integrated effort.

Project Administration

The overall organization and structure of the Project A research continued unchanged in FY84. For administrative purposes, Project A is organized into major tasks (Task 1, Validation; Task 2, Developing Predictors of Job Performance; Task 3, Measurement of School/Training Success; Task 4, Assessment of Army-wide Performance; Task 5, Develop MOS-Specific Performance Measures; Task 6, Management). The research efforts under the various tasks are interrelated and integrated through the continuous oversight of Task 6 in-house and contractor staffs as well as the regular program of Interim Progress Review (IPR) meetings and discussions.

Contract Amendment. ARI Research Report 1332, "Improving the Selection, Classification and Utilization of Army Enlisted Personnel--Project A: Research Plan" (May 1983), specified a number of changes to the original scope of work described in the RFP. These changes required that an amendment to the contract be formulated and approved to bring it into conformance with the Project A Research Plan.

The amendment provides for a shift in focus to future cohorts (from the FY81/82 and FY84/85 cohorts to the FY83/84 and FY86/87 cohorts). It also specifies the additional work entailed in:

- Acquiring school data on the FY83/84 cohort for predictor and criterion development.
- Conducting validity analyses of FY81/82 cohort data in support of mandated Aptitude Area Composite recommendations.
- Conducting job and task analyses to support new "cluster" constructs, and identifying the focal MOS.
- Preparing detailed analyses and justification to support the sampling strategy (and the resultant Troop Support Requests).
- Accomplishing a "Preliminary Battery" identification and test phase in the predictor development and test research program.
- Acquiring, using, and maintaining psychomotor/perceptual test equipment in the new predictor Trial and Experimental Battery research and development program.
- Expanding the utility research program to include the requirements for development of "monetization" metrics.

- Extending the research schedule through 1991 to retain the objective of analyzing second-term validity data on the second (FY86/87) main cohort.

In December 1983, ARI informed the consortium managers that funding plans for the second year of contract performance would have to conform to funding limitations and that the research program activities would have to be adjusted accordingly. Concurrent with accommodating to FY84 fund limitations, it was determined that the estimate of resources required for scientific quality assurance and control, interim product development and exploitation, an expanded program of communications and reporting, and maintenance of intertask coordination and interface was insufficient for a program of this scope and complexity. Accordingly, the amendment to the contract provided resources for meeting these new requirements and constraints.

An amendment proposal for the contract was provided to ARI 20 April 1984 and subjected to an intensive review and evaluation process. On 28 September 1984 the amendment was approved and was incorporated into the contract.

Psychomotor/Perceptual Test Equipment. Included in the changes noted above was a requirement for an extensive investigation of psychomotor/perceptual constructs to meet the objective of researching the broadest spectrum of potential predictors, thereby providing a better possibility of improving on the ASVAB. Implementing this decision required the acquisition, use, and maintenance of psychomotor/perceptual equipment for development work and the subsequent major data collections planned for the FY83/84 and FY86/87 main cohorts.

During FY84, all of the procedures and requirements of AR 18-1, governing the acquisition of computers, were fully complied with; this included the development and provision of a satisfactory Mission Element Need Statement (MENS), an Acquisition Plan, and an Economic Analysis supporting and justifying the requirement for the psychomotor/perceptual testing equipment. These documents were reviewed by the cognizant Army organizations, and the acquisition was approved 2 August 1984 by the Assistant Secretary of the Army (Financial Management).

Personnel Changes. During the course of the second year's work a number of personnel changes were effected in the Governance Advisory Group. BG W. C. Knudson (Office of the Deputy Chief of Staff for Operations and Plans) and BG Frederick M. Franks, Jr. (USAREUR) were designated as U.S. Army Advisors. In addition, Dr. W. S. Sellman replaced Dr. G. T. Sicilia as the DOD Interservice Advisor. These changes are reflected in Figure 13.

There were also changes in assignments for the ARI Task Monitors and consortium Task Leaders and other key personnel. The assignments for these monitor/leader positions at the end of FY84 are reflected in Figure 14. To help in providing the best advice and evaluation of task activities, members of the Scientific Advisory Group agreed to place special emphasis on specific tasks and monitor task progress at semiannual in-progress reviews. Dr. Linn is aligned with Task 1, Drs. Humphreys and Uhlaner with Task 2, Dr. Hakel with Task 3, Dr. Bobko with Task 4, and Drs. Cook and Tenopyr with Task 5.

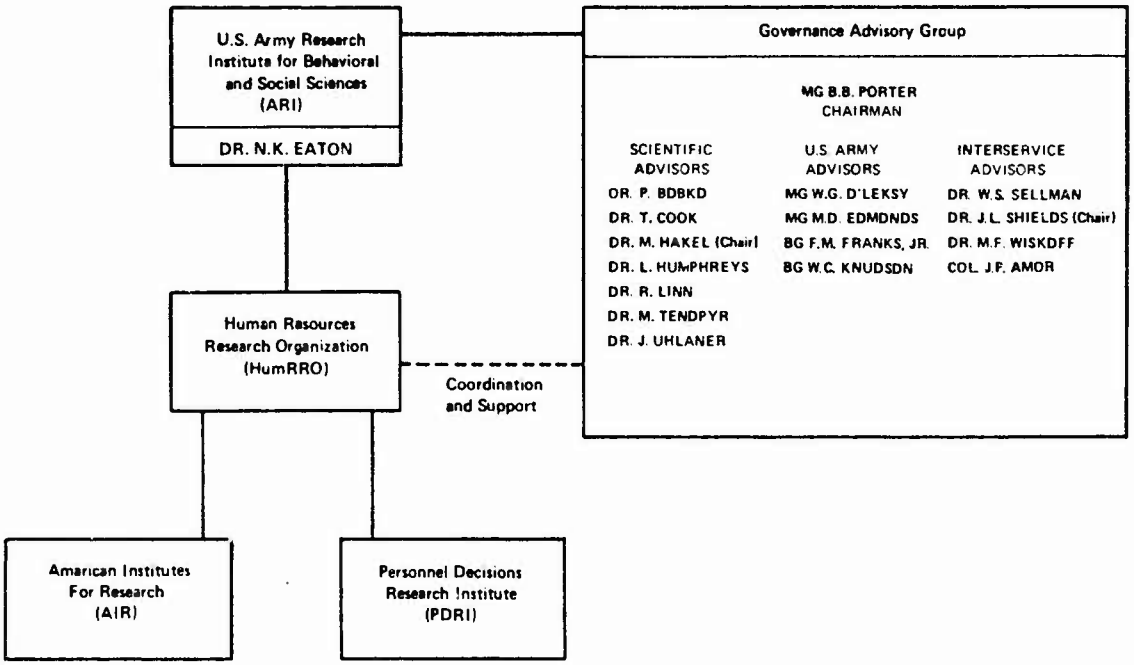


Figure 13. Governance Advisory Group

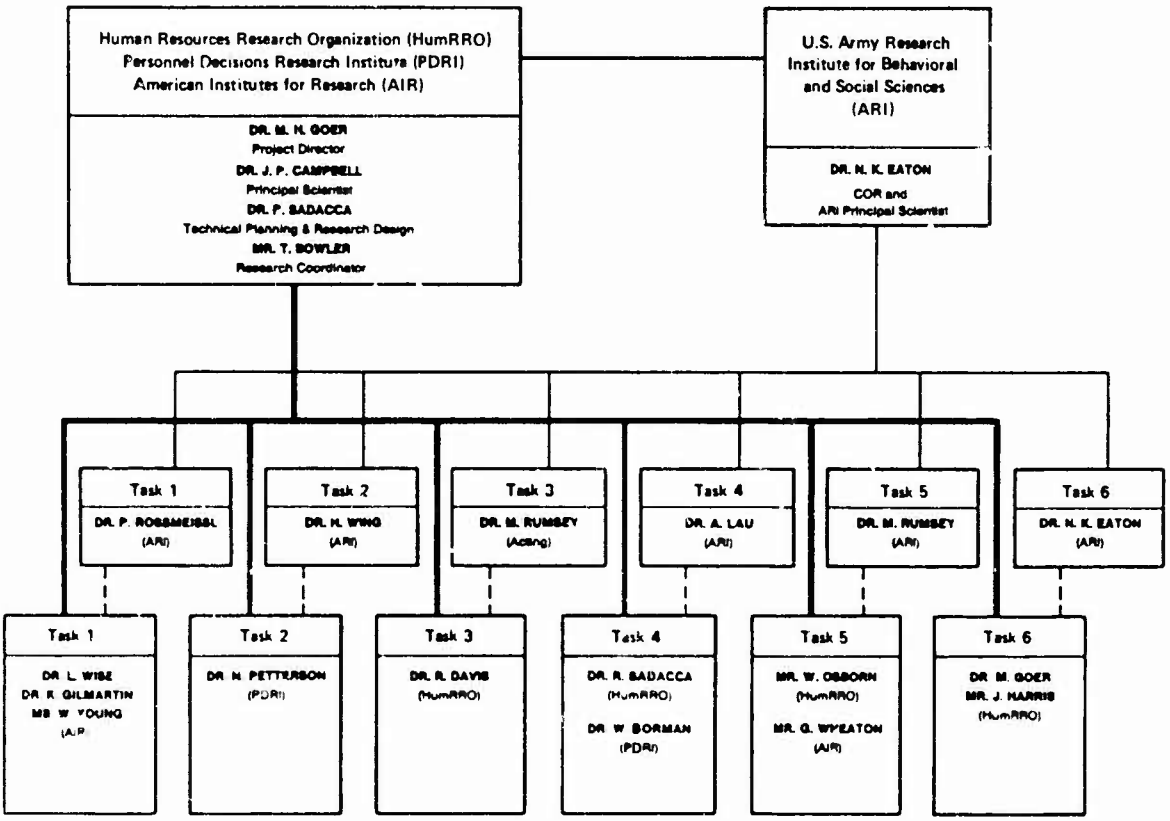


Figure 14. Project Organization

Organization of the Following Chapters

The second year's work is described in detail in the following chapters of this report: Chapter II, School and Job Performance Measurement; Chapter III, Predictor Measurement; Chapter IV, Validation.

The second year of Project A can be characterized primarily as a development and tryout phase for the new measures we are seeking to develop. It involved investigation and resolution of some troublesome methodological issues, then the subsequent conduct of both computer-oriented research on existing data and empirical studies in the field to obtain new data. It was primarily a process of trial-and-revision, trial-and-revision. It was not a period in which we expected to end up with a large number of finished products. Even so, we could document numerous units of work that had both immediate utility for the Army and broad interest for other researchers in personnel systems.

In each chapter the primary research accomplished in the area is summarized. The summary is followed by any reports or papers produced in the area during the year. Only abstracts are provided if the work has been published previously. Appendix material for certain reports included in this volume is supplied in ARI Research Note 85-14. A synopsis volume based on selected sections of this document is available as ARI Research Report 1393. It summarizes the work and contains abstracts of the documents presented in full in this volume.

Associated Reports and Papers

The annual report for the first year of Project A was published in companion volumes, and the database plan was also published.

(1) The planning, initial operations, and preliminary work on predictor criterion development during the first year of Project A were described in an ARI Research Report prepared by the ARI and consortium scientists directing the project.

(2) Supplementing the preceding report was a technical appendix published as an ARI Research Note. This volume, edited by Eaton and Goer, described the first year of Project A research in more detail and included a variety of technical papers prepared during that year.

(3) The development and long-range plans for the Project A longitudinal research database were described by Wise, Wang, and Rossmeissl in an ARI Research Report, which provided details on both content and procedures.

(4) A paper prepared by Eaton to summarize the first 18 months of work on Project A has been amplified and updated for later presentations, and as the basis for Chapter I of the present report.

References

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute. Improving the selection, classification, and utilization of Army enlisted personnel. Project A: Research plan. ARI Research Report 1332. Alexandria, VA. December 1983. (ADA 129728)

Peterson, N. G., and Bownas, D.A. Task structure and performance acquisition. In M.D. Dunnette and E.A. Fleishman (Eds.), Human capability assessment. New York: Lawrence Erlbaum & Associates, 49-105. 1982.

Rosse, R.L., Borman, W.C., Campbell, C.H., and Osborn, W.C. Grouping Army occupational specialties by judged similarity. Paper presented at the Military Testing Association in Gulf Shores, Alabama, October 1983. (See ARI Research Note 83-37.)

Schmidt, F.L., Hunter, J.E., McKenzie, R., and Muldrow, T. The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 64, 609-626. 1979.

ARI Research Report 1347*
IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL: ANNUAL REPORT

Human Resources Research Organization
American Institutes for Research
Personnel Decisions Research Institute
Army Research Institute
(October 1983)

This Research Report describes the research performed during the first year of a project to develop a complete personnel system for selecting and classifying all entry-level enlisted personnel. In general, the first year's activities have been taken up by an intensive period of detailed planning, briefing advisory groups, preparing initial troop requests, and beginning comprehensive predictor and criterion development that will be the basis for later validation work. A detailed description of the first year's work, including technical papers, is contained in the Annual Report Technical Appendix, ARI Research Note 83-37.

---

* Available from Defense Technical Information Center, 5010 Duke Street, Alexandria, VA 22314. Phone: (202) 274-7633. Order Document No. ADA141807.

ARI Research Note 83-37*
IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL: TECHNICAL APPENDIX TO THE ANNUAL REPORT

Newell K. Eaton and Marvin H. Goer (Editors)
(October 1983)

This Research Note describes in detail research performed during the first year of a project to develop a complete personnel system for selecting and classifying all entry-level personnel. Its purpose is to document, in the context of the annual report, a variety of technical papers associated with the project. In general, the first year's activities have been taken up by an intensive period of detailed planning, briefing advisory groups, preparing initial troop requests, and beginning comprehensive predictor and criterion development that will be the basis for later validation work. Research reports associated with the work reported are included.

---

* Available from Defense Technical Information Center, 5010 Duke Street, Alexandria, VA 22314. Phone: (202) 274-7633. Order Document No. ADA137117.

ARI Research Report 1356*
DEVELOPMENT AND VALIDATION OF ARMY
SELECTION AND CLASSIFICATION MEASURES
PROJECT A: LONGITUDINAL RESEARCH DATABASE PLAN

Lauress L. Wise and Ming-mei Wang
(AIR)
Paul G. Rossmeissl
(ARI)
(December 1983)

This Research Report describes plans for the development of a major longitudinal research database. The objective of this database is to support the development and validation of new predictors of Army performance and also new measures of Army performance against which the new predictors can be validated. This report describes the anticipated contents of the database, editing procedures for assuring the accuracy of the data entered, storage and access procedures, documentation and dissemination procedures, and database security procedures.

---

* Available from Defense Technical Information Center, 5010 Duke Street, Alexandria, VA 22314. Phone: (202) 274-7633. Order Document No. ADA143615. The plan was included in the FY83 annual report (ARI Research Note 83-37) prior to publication as a Research Report.

THE U.S. ARMY RESEARCH PROJECT TO IMPROVE SELECTION AND CLASSIFICATION DECISIONS*

Newell K. Eaton
(ARI)

This paper provides an overview of the Army's Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel, and summarizes the results from the first 18 months of work. This major research effort will tie together the selection, classification, and job allocation of enlisted soldiers so that personnel decisions can be made to optimize performance and the utilization of individual abilities. Many activities are under way to improve predictor validity and performance measurement. Improved individual recruiting, performance, and retention are expected because the system will be designed to make the best match between the Army's needs and the individual's qualifications.

---

* Paper presented at the National Security Industrial Association Conference on Personnel and Training Factors in System Effectiveness, in Springfield, Virginia, 1-3 May 1984. It is an expansion and update of a paper written earlier by N.K. Eaton and E.J. Schmitz for presentation at the ORSA/TIMS Joint National Meeting in Orlando, Florida, 7-9 November 1983. The full text is not presented here because it was published in the proceedings of the National Security Industrial Association and is the basis of the preceding Chapter I, "Introduction to Current Army Selection and Classification Research."

II. SCHOOL AND JOB PERFORMANCE MEASUREMENT

John C. Campbell

The overall objective for criterion measurement within Project A is to develop a broad array of valid and reliable criterion measures that reflect all major factors of job performance for first-tour enlisted personnel. These should constitute state-of-the-art criteria against which selection and classification measures can be validated. Within this general objective the more specific purposes are to (a) determine the relationship of training performance to on-the-job performance, (b) measure performance "hands-on" by standardized simulations and work samples, and (c) compare rating scales, knowledge tests, and standardized work samples as alternative measures of specific task performance.

Project A is being conducted on a carefully selected sample of 19 MOS, as previously described. Using large samples of individuals from each of these 19 MOS, a major concurrent validation will be conducted in 1985 and a longitudinal validation will begin in 1986. Criterion measures that are specific to a particular MOS are being developed in "batches." The first batch (designated A or X) includes four MOS, the second batch (B/Y) five MOS, and the third batch (Z) 10 MOS.

Objectives for FY84

As described in the FY83 annual reports, Project A criterion development was at the following point at the beginning of the project's second year, in October 1983:

- The critical incident procedure had been used with two workshops of officers to develop a first set of 22 dimensions of Army-wide rating scales, as well as an overall performance scale and a scale for rating the potential of an individual to be an effective NCO.

- The critical incident procedure had also been used to develop dimensions of technical performance for each of the four MOS in Batch A (13B, cannon crewman; 64C, motor transport operator; 71L, administrative specialist; 95B, military police).

- A painstaking process had been used to select the pool of 30 tasks in each Batch A MOS that would be subjected to hands-on and/or knowledge test measurement. After preparing job task descriptions, the staff used a series of judgments by subject matter experts (SME), considering task importance, task difficulty, and intertask similarity, as the basis for selecting the final sets of tasks.

- On the way to developing norm-referenced training achievement tests for each of the 19 MOS, the staff had visited each proponent school and developed a description of the objectives and content of the training curriculum. They had also used Army Occupational Survey Program information to develop a detailed task description of job content for each MOS. After low-frequency elements were eliminated, SME judgments (N = 3-6) were used to rate the importance and error frequency for each task element. Approximately 225 tasks were then sampled proportionately from MOS duty areas. Consequently, at the end of FY83 we had a refined task sample for each MOS and systematic descriptions of the training program against which to develop a test item budget.

- A preliminary analysis had been made of the feasibility of obtaining archival performance records from the computerized Enlisted Master File (EMF) and the Official Military Personnel File (OMPF), which is centrally stored on microfiche. Because the OMPF data were incomplete, the staff decided to examine a sample of 201 Files (Military Personnel Records Jacket) to determine whether these files would be a more useful source of information.

The principal objectives for criterion development for FY84 were as follows:

(1) Use the information developed in FY83 to construct the initial version of each criterion measure.

(2) Pilot test each initial version and modify as appropriate.

(3) Evaluate the criterion measures for the four MOS in Batch A in a relatively large-scale field test (about 150 enlisted personnel in each MOS).

Construction of Initial Measures

Army-Wide Rating Scales. An additional four critical incident workshops involving 77 officers and NCOs were conducted during FY84. On the basis of the critical incidents collected in all workshops, a preliminary set of 15 Army-wide performance dimensions was identified and defined. Using a combination of workshop and mail survey participants (N = 61), the initial set of dimensions was retranslated and 11 Army-wide performance factors survived. The scaled critical incidents were used to define anchors for each scale, and directions and training materials for raters were developed and pretested.

During the same period scales were developed to rate overall performance and individual potential for success as an NCO. Finally, rating scales were constructed for each of 14 common tasks that were identified as part of the responsibility of each individual in every MOS.

MOS-Specific BARS Scales. Four critical incident workshops involving 70-75 officers and NCOs were completed for each of the MOS in Batch A and Batch B. A retranslation step similar to that for the Army-wide rating scales was carried out, and six to nine MOS-specific performance rating scales (Behaviorally Anchored Rating Scales, BARS) were developed for each MOS. Directions and training materials for scales were also developed and pretested.

Hands-On Measures (Batch A). After the 30 tasks per MOS were selected for Batch A, the two major development tasks that remained before actual preparation of tests were the review of the task lists by the proponent schools and the assignment of tasks to testing mode (i.e., hands-on job samples vs. knowledge testing).

The completeness and representativeness of the task lists were officially reviewed by the proponent school. Three of the reviews were conducted by mail and one through on-site briefing. Only slight changes were made in the task lists as a result of the reviews.

For assignment of tasks to testing mode, each task was rated by three to five project staff on three dimensions:

- The degree of physical skill required.
- The degree to which the task must be performed in a series of steps that cannot be omitted.
- The degree to which speed of performance is an important indicator of proficiency.

The extent to which a task was judged to require a high level of physical skill, a series of prescribed steps, and speed of performance determined whether it was assigned to the hands-on mode. For each MOS, 15 tasks were designated for hands-on measurement. Job knowledge test items were developed for all 30 tasks.

The pool of initial work samples for the hands-on measures was then generated from training manuals, field manuals, interviews with officers and job incumbents, and any other appropriate source. Each task "test" was designed to take from 5 to 10 minutes and was composed of a number of steps (e.g., in performing cardiopulmonary resuscitation), each of which was to be scored "go, no-go" by an incumbent NCO. A complete set of directions and training materials for scorers was developed; scorer training is thorough and is intended to take the better part of one day. The initial hands-on measures and scorer directions were then pretested on 5 to 10 job incumbents in each MOS and revised. They were ready for administration to the field test samples during the summer and fall of 1984.

MOS-Specific Job Knowledge Tests (Batch A). Concurrently, a paper-and-pencil, multiple-choice job knowledge test was developed to cover all of the 30 tasks in the MOS lists. The item content was generated on the basis of training materials, job analysis information, and interviews, with 4 to 10 items prepared for each of the 30 tasks. For the 15 tasks also measured

hands-on, the knowledge items were intended to be as parallel as possible to the steps that comprised the hands-on mode. The knowledge tests were pilot tested on approximately 10 job incumbents per MOS. After revision they were deemed ready for tryout with the field test samples.

Task Selection and Test Construction for Batch B. By the end of FY84, basic task descriptions had been developed for Batch B in a manner similar to that used for Batch A; that is, the CODAP (Comprehensive Occupational Data Analysis Program) and Soldier's Manual descriptions had been merged, edited to a uniform level of specificity, and evaluated for completeness and currency. The task descriptions have not yet been submitted to SME judgments of difficulty, importance, and similarity. The remaining steps of task selection, proponent review, assignment to testing mode, and test construction are scheduled for FY85. In addition, for Batch B a formal experimental procedure is being used to determine the effects of scenario differences on SME judgment of task importance. The design calls for 30 SMEs to be randomly assigned to one of three scenarios (garrison duty/peacetime, full readiness for a European conflict, and an outbreak of hostilities in Europe). The implications of scenario differences are discussed later in this section.

Training Achievement Tests (Batch X). During FY84 generation of refined task lists for each of the 19 MOS in the Project A sample continued. For each MOS in Batch X (same MOS as Batch A), an item budget was prepared matching job duty areas to course content modules and specifying the number of items that should be written for each combination. An item pool that reflected the item budget was then written by a team of SMEs contracted for that purpose. Next, training content SMEs and job content SMEs judged each item in terms of its importance for the job (under each of the three scenarios, in a repeated measures design), its relevance for training, and its difficulty. The items were then "retranslated" back into their respective duty areas by the job SMEs and into their respective training modules by the training SMEs. Items were designated as "job only" if they reflected task elements that were described as an important part of the job but had no match with training content; such items are intended to be a measure of incidental learning in training.

Once the sample of task elements was determined for each MOS and the items written and edited for basic clarity and relevance to the training, the job, or both, the pool was ready for tryout with the field test samples of incumbents and a sample of 50 trainers from each MOS.

Administrative (Archival) Indices. A major effort in FY84 was a systematic comparison of information found in the Enlisted Master File (EMF), the Official Military Personnel File (OMPF), and the Military Personnel Records Jacket (201 File). A sample of 750 incumbents, stratified by MOS and by location, was selected and the files searched. For the 201 Files the research team made on-site visits and used a previously developed protocol to record the relevant information. A total of 14 items of information, including awards, letters of commendation, and disciplinary actions seemed, on the basis of their base rates and judged relevance, to have at least some potential for service as criterion measures. Unfortunately the microfiche records appeared too incomplete to be useful and search of the

201 Files was cumbersome and expensive. It was decided to try out a self-report measure for the 14 administrative indices and compare it to actual 201 File information for the people in the field trials.

Batch A(X) Field Tests

The goal for the FY84 criterion field tests was to obtain enough information to permit relatively stable estimates of item and scale statistics, reliability indices, and scale/test intercorrelations. On the basis of these data, the array of criterion measures must be reduced to fit the time available (16 hours for Batch A/X and Batch B/Y MOS) for the FY83/84 concurrent validation sample which will be tested during the summer of 1985. The reduction must be accomplished by eliminating items and scales with psychometric deficiencies that cannot be fixed, redundant measures, and (if necessary) the least crucial parts of the criterion space.

Field Test Criterion Battery. The complete array of specific criterion measures that was actually used at each field test site is given below. For each rating scale every effort was made to obtain a complete set of supervisor, peer, and self ratings. This may very well be the most comprehensive array of performance measures ever used in a personnel research project.

A. MOS-Specific Performance Measures

1) Paper-and-pencil tests of knowledge of task procedures consisting of 4-10 items for each of 30 major job tasks for each MOS. Item scores can be aggregated in at least the following ways:

- Sum of item scores for each of the 30 tasks.
- Sum of item scores for common tasks.
- Sum of item scores for MOS unique tasks.
- Sum of item scores for 15 tasks also measured hands-on.

2) Hands-on measures of 15 tasks for each MOS.

- Individual task scores.
- Total score for common tasks.
- Total score for unique tasks.

3) Ratings of performance on each of the 15 tasks measured via hands-on methods by:

- Supervisors
- Peers
- Self

4) Behaviorally anchored rating scales of 5-9 performance dimensions for each MOS by:
   - Supervisors
   - Peers
   - Self

5) A general rating of overall job performance by:
   - Supervisors
   - Peers
   - Self

B. Army-Wide Measures

1) Eleven behaviorally anchored rating scales designed to assess the following dimensions. Three sets of ratings (i.e., from supervisors, peers, and self) were obtained on each scale for each individual.

   a) Technical Knowledge/Skill
   b) Initiative/Effort
   c) Following Regulations/Orders
   d) Integrity
   e) Leading and Supporting
   f) Maintaining Assigned Equipment
   g) Maintaining Living/Work Areas
   h) Military Appearance
   i) Physical Fitness
   j) Self-Development
   k) Self-Control

2) A rating of general overall effectiveness as a soldier by:
   - Supervisors
   - Peers
   - Self

3) A rating of NCO potential by:
   - Supervisors
   - Peers
   - Self

4) A rating of performance on each of 14 common tasks from the manual of common tasks by:
   - Supervisors
   - Peers
   - Self

5) A 14-item self-report measure of certain administrative indices such as awards, letters of commendation, and reenlistment eligibility.

6) The same administrative indices taken from 201 Files.

7) Attrit/not attrit during the first 180 days.

The Field Test Samples. The field test data were collected at different sites over a period of four months. Data for administrative specialists and military police were collected in U.S. installations during May, July, and August of 1984. Data on cannon crewmen and motor transport operators were obtained from two sites in Germany during August and September of 1984. The breakdown of subjects by MOS and by location is shown in Table 5. All subjects were incumbent enlisted personnel who had been in the Army 12 to 24 months.

Table 5. "BATCH A" FIELD TEST SAMPLES

MOS		N
Administrative Specialists (71L)		129
Fort Polk	60	
Fort Hood	48	
Fort Riley	21	
Military Police (95B)		113
Fort Polk	42	
Fort Hood	42	
Fort Riley	29	
Cannon Crewmen (13B)		
Herzobase		150
Motor Transport Operators (64C)		
Mannheim		155
Total		547

Procedure. Staff members worked closely with the point of contact to secure testing sites, assemble equipment, and gain the cooperation of support personnel. The week before data collection, a project team visited the site to make sure everything was ready and to train the scorers of the hands-on measures. The tests and rating scales were administered by project personnel. Each participant was tested on each measure during a 2-day testing period. Approximately half the participants returned 6-12 days later and were retested on the hands-on measures. Every effort was made to obtain at least two supervisors and two peers to serve as raters for each incumbent on the rating scale measures. However, only one scorer was used for each hands-on task and scorers differed across tasks.

Analyses: Field Test Data. By the end of FY84, the field tests had been completed but the analyses of the data had not yet begun. To proceed from the current array of criterion measures to the set of measures to be used in the FY83/84 concurrent validation during 1985, a "Criterion Measures

Task Force" composed of appropriate consortium and ARI scientists and outside scientific advisers is being assembled. Their assignment is to systematically review the field test data and, through a series of decision meetings, eliminate poor quality or redundant measures, authorize revisions, and eventually make the reductions necessary to meet the concurrent validation time constraints. The first major meeting to review the field test data analysis was scheduled for November 1984.

Arriving at the criterion composites for the FY83/84 cohort validation is not the goal at this stage; those decisions will be a function of the FY83/84 concurrent validation data. The overall analysis objective is to reduce the amount of criterion measurement to fit the available time and at the same time maintain as broad a coverage of the criterion space as possible.

The specific objectives for the Criterion Measures Task Force are:

- Identify criterion measures that can be eliminated on the basis of poor psychometric quality or redundancy.
- Specify a prioritized list of options for reducing the Batch A criterion measures to fit the time constraints of the 1985 concurrent validation.

Confirmatory Analysis: A Beginning

After all analyses of the field test data are complete, Project A can take another step toward one of its major criterion development goals, the further refinement of the working model of soldier effectiveness. This could be done by first presenting the complete results of the field tests at a meeting of key task scientists and discussing them thoroughly. Next, task scientists would generate their own model of the criterion space. This would consist of naming and offering a definition for the latent variables, specifying how they are best measured by the available criteria, and describing any important features of the criterion space that he or she thinks are worth noting (e.g., "it is hierarchical in the following way ..."). Then a Delphi procedure could be used to show each model to everyone else and have each task leader produce a revised model. The revised models could be discussed at another group meeting to find out where there is agreement and disagreement about what the criterion space looks like. On the basis of that meeting, one or more alternative structural models that could be put to a confirmatory analysis in the FY83/84 cohort sample would be produced.

Discussion and Conclusions

As has been noted, the major accomplishments in criterion development for FY84 were:

(1) Construction, for four military jobs, of the initial operational versions of the largest and most comprehensive array of job performance criterion measures in the history of personnel selection/classification research.

(2) Revision and refinement of each measure through pilot testing.

(3) Development and pilot testing of training materials for raters and test administrators.

(4) Completion of a comprehensive field test of all criterion measures for four MOS, which involved two days of testing for approximately 600 job incumbents in several locations in the continental United States and in Europe.

(5) Preparation of the field test data for analysis.

Consequently, we now have the information necessary for making final revisions and for creating the final array of operational criterion measures for use for four MOS in the FY83/84 cohort concurrent validation during the summer of 1985. There is also an operational plan for how to analyze the field test data and an operational decisionmaking procedure for the final selection of criterion measures to be used in the concurrent validation.

During the past year a number of special issues have arisen that bear on criterion development in Project A. Some have been resolved and some are still under discussion. None have precise answers or are completely scientific in nature.

Scenario Effects. At several points in Project A, raters or SMEs are being asked to make judgments about such things as (a) the relative importance of specific job tasks to an MOS, (b) the relative importance of a knowledge test item for the objectives of a particular AIT program, (c) the degree of effective job performance reflected in a particular critical incident, (d) the job proficiency of a ratee on specific performance factors, and (e) the relative value (i.e., utility) of different job performance levels across MOS (e.g., How much more or less valuable to the Army is high performance for administrative specialists vs. low performance for motor transport operators?). It is often asserted that such judgments can be made meaningfully only when the context for the judgment (i.e., the scenario) is specified for the judge. For example, the relative importance of a specific task in the array of tasks that comprise an MOS can be judged only when the SME knows the context in which the task is to be performed (e.g., peacetime, wartime, field exercises).

There are two major reasons why differential scenario effects, if they exist, would be important for Project A.

First, they would influence the selection of content for all the criterion measures that we are using. For example, if job tasks vary in importance depending on the scenario, and hands-on or knowledge tests of task proficiency are to be constructed, then a wider variety of tasks may have to be included in the hands-on measure or knowledge test. That is, more items would be needed to cover all the important tasks if the subset of important tasks is not the same under each scenario.

Second, if the relative importance weights (i.e., utilities) for different MOS and for different performance levels within MOS vary substantially as a function of major scenario changes, then the selection/classification algorithm must incorporate different sets of utility weights which can be changed as the mission needs of the Army change.

To account for scenario differences in the selection of content for the MOS-specific job performance measures and the MOS-specific training performance measures, the following steps are currently being undertaken. For the five MOS in Batch B (same MOS as Batch Y), scenario effects on SME judgments of task importance are being studied experimentally. A total of 30 SMEs will be randomly assigned to one of three different scenarios, which are shown in Figure 15. Mean differences in importance ratings (by task and task cluster) will then be compared across scenarios. The same three scenarios are being used in a repeated measures design to study scenario effects on judgments of item relevance for the knowledge tests to be used in Batch Y and Batch Z; SMEs are being asked to judge the relative importance of each knowledge test item for the content of the job. Each SME makes three importance judgments for each item corresponding to the three scenarios.

Results from the above steps will be used to determine whether scenario effects do in fact exist, and if so, for what types of tasks they are largest (e.g., common vs. MOS-specific). Preliminary results indicate that scenario effects on importance judgments are significant for certain kinds of tasks within some MOS. In particular, for non-combat support MOS the common tasks become more important and the MOS-specific tasks somewhat less important under a conflict rather than peacetime scenario.

Since some scenario effects do exist, the resolution has been to select tasks and test items that accommodate the differences. The preliminary data suggest that this should be possible within the constraints imposed by the FY83/84 concurrent validation design.

Multi-Method Measurement. In virtually any research project it is very desirable if the major variables can be measured by more than one method. In Project A, MOS-specific task performance is being assessed by three different methods (i.e., ratings, hands-on tests, and knowledge tests). Since testing time is not unlimited, a relevant issue is whether multiple measures should be retained for the concurrent validation at the expense of breadth of coverage, or vice versa. The relevant analyses that will inform this decision are not yet available, but the prevailing strategy is to do everything possible to preserve multiple measurement.

Weighting Criterion Components. Several measures in the criterion array are made up of component scores in the form of subtests on performance on complete tasks, as in the hands-on measures. A general issue concerns whether such components (e.g., the 15 separate hands-on tasks) should be differentially weighted before being combined into a total score. The same question arises when the aim is to combine specific criterion measures (e.g., ratings, knowledge tests, hands-on tests) into an overall composite for test validation.

1) Your unit is assigned to a U.S. Corps in Europe. Hostilities have broken out and the Corps' combat units are engaged. The Corps' mission is to defend, then reestablish, the host country's border. Pockets of enemy airborne/heliborne and guerilla elements are operating throughout the Corps sector area. The Corps maneuver terrain is rugged, hilly, and wooded, and weather is expected to be wet and cold. Limited initial and reactive chemical strikes have been employed but nuclear strikes have not been initiated. Air parity does exist.

2) Your unit is deployed to Europe as part of a U.S. Corps. The Corps' mission is to defend and maintain the host country's border during a period of escalating hostilities. The Corps maneuver terrain is inhibiting, weather is expected to be inclement. The enemy approximates a combined arms army and has nuclear and chemical capability. Air parity does exist. Enemy adheres to same environmental and tactical constraints as does U.S. Corps.

3) Your unit is a TO&E Field Artillery Battalion stationed on a military post in the Continental United States. The unit has personnel and equipment sufficient to make it mission capable for training and evaluation. The training cycle includes periodic field exercises, command and maintenance inspections, ARTEP evaluations, and individual soldier training/SQT testing. The unit participates in post installation responsibilities such as guard duty and grounds maintenance and provides personnel for ceremonies, burial details, and training support to other units.

Figure 15. Three Alternative Scenarios for SME Judgments Of Task and Item Importance

Two principal considerations govern the weighting of criterion components. First, the relative weight given to a particular component of job performance is a value judgment. Such judgments are part of the overall question of what an organization wants its people to be able to do. Weighting on other grounds, such as the relative reliability of measurement or degree of predictability, might produce composites in which the least important components are given the greatest weight. Second, the literature on differential weighting strongly suggests that if the number of components is very large (i.e., more than 4-6), then differential weighting makes very little difference in the psychometric properties of the total score. Consequently, a reasonable strategy for Project A would be to compare weighted vs. unweighted criterion composites to determine whether differential weighting produces an advantage. The issue is scheduled to be considered during FY85.

Criterion Differences Across MOS. In Project A's validation of predictor measures for each of 19 jobs, the extent to which the same array of criterion measures will be used for the criterion composite in each MOS is a relevant question. For example, would job knowledge tests be used as a component of job performance in some MOS but not in others? This issue is being addressed directly by the continuing effort in Project A to develop an overall model of the effective soldier. Within its current form, the model specifies the same set of constructs, or basic performance factors, for each MOS. In general, this means that very much the same measures would be used across MOS; however, their relative weights could vary considerably depending on the results of the MOS-specific development work and the criterion importance judgments. For example, the criterion factors assessed by the Army-wide rating scales could receive a much greater weight for combat MOS than for support MOS. Again, however, the most relevant data for informing this issue are not scheduled to be collected until FY85.

Potential Applications of FY84 Criterion Development Products

Since Project A is an R&D project designed to produce an improved selection and classification system for U.S. Army enlisted personnel, the purpose of criterion development is to produce optimal performance measures against which to validate new and improved selection and classification tests, rather than to produce new methods for operational performance appraisal. However, much of Project A's R&D work has operational implications. The major items that flow from the work during FY84 are as follows:

(1) The extensive work on the development of Army-wide performance factors via the critical incident workshops will provide a means both to confirm the validity of the current EER factors and to refine and extend the content of the EER if the Army so desires.

(2) The results of the 201 File analysis would be a valuable aid in any future attempts to refine the use of 201 File information in making future promotion or reenlistment decisions.

Associated Reports and Papers

We have divided Project A reports and papers associated with performance measurement into two categories. Those dealing with operational research activities are presented first, while those addressing methodological considerations follow.

Reports and papers dealing with operational research activities

(1) SQT scores were analyzed as a function of aptitude area composite scores in four logistics MOS, in a report by Rossmeissl and Eaton. Particular attention was paid to the SQT scores of soldiers whose earlier ASVAB aptitude area scores had been close to the minimum score for eligibility to enter the MOS. This analysis made it possible to explore the potential effects, on both numbers of eligible recruits and subsequent probable performance levels in the MOS, of changing the minimum cutoff score for MOS eligibility.

(2) The advantages and the difficulties of attempting to use administrative records as a measure of a soldier's general effectiveness in the Army were analyzed in a paper by Riegelhaupt, Harris, and Sadacca. The record used was the Military Personnel Record Jacket (MPRJ); 38 variables were studied to determine the amount of information that could be compiled from these records, and the information's usefulness in establishing criteria for effectiveness.

(3) In view of the emphasis being placed on developing ratings as a criterion of an individual's general Army effectiveness, factors that affect peer and supervisor ratings were studied by Borman, White, and Gast. How such ratings are made and how they relate to other means of measuring performance are important topics on which more information is needed.

(4) Another approach to measuring an individual's general Army effectiveness is discussed in a paper by Olson, Borman, Roberson, and Rose. Scales to show environmental and situational influences that affect job performance were developed, and the ratings from these scales were compared with the results of direct measures to job performance.

BLANK PAGE

Working Paper

RS-WP-84-12

AN ANALYSIS OF SQT SCORES AS A FUNCTION OF APTITUDE AREA COMPOSITE SCORES FOR LOGISTICS MOS

Paul G. Rossmeissl and Newell K. Eaton
SELECTION AND CLASSIFICATION TECHNICAL AREA

April 1984

Reviewed by
Lawrence M. Hanser
Selection and Classification
Technical Area

Approved by
Newell K. Eaton
Chief
Selection and Classification
Technical Area

U.S. Army Research Institute
for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria VA 22333

This working paper is an unofficial document intended for limited distribution to obtain comments. The views, opinions, and/or findings contained in this document are those of the author(s) and should not be construed as the official position of ARI or as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

An Analysis of SQT Scores as a Function of Aptitude Area Composite Scores for Logistics MOS

Paul G. Rossmeissl
and
Newell K. Eaton

US Army Research Institute
for the Behavioral and Social Sciences

The purpose of this research was to provide information on the relationship between Armed Services Vocational Aptitude Battery (ASVAB) aptitude area (AA) scores and soldier performance on the Skill Qualification Test (SQT) in selected Quartermaster Military Occupational Specialties (MOS). We hoped that such information could prove useful to the U. S. Army Quartermaster School in recommending the AA minimum scores used to determine enlistment eligibility in their MOS.

In order for a prospective soldier to be eligible to enlist in an Army MOS he or she must first obtain a score equal or greater than the minimum on the AA composite for that MOS. The reason for setting minimum eligibility scores is to increase the likelihood that individuals selected for entry into an MOS will be good performers. The basic procedure in determining the minimum eligibility score is first to set a minimum performance standard, or range of acceptable performance, and then to tradeoff improved performance predicted by increased scores on a valid predictor against the reduced supply of applicants with higher predictor scores. For these purposes a valid predictor is one for which higher predictor scores are associated with higher performance scores.

At this time the best predictor of Army enlisted performance is the ASVAB, a cognitive test battery used by all the services as a basis for enlistment decisions. It has been shown to be valid across a large number of MOS over many years. Each of the nine AA composites now being used by the Army to select/classify enlisted personnel is some combination of the subtests of the ASVAB.

Currently, the best available measure of a soldier's on-the-job performance within the Army is that soldier's SQT. The Army has administered SQTs to enlisted soldiers since 1977. The SQTs were initially developed to assess a soldier's qualifications for promotion and to evaluate the effectiveness of Army training programs, however, they are no longer formally used for promotion decisions . Originally the SQTs were composed of three parts: a written MOS specific test, a MOS specific hands on test , and a commander's evaluation. In 1983 the SQT became a component of the Army Individual Training Evaluation Program (ITEP) and now contains only written MOS-specific measures or items.

Method

Sample. Four quartermaster school MOS were selected for this research. These MOS were chosen because they were the only ones from this school with a sufficient number of soldiers for which we had both ASVAB and SQT data to permit meaningful analyses. The four MOS were 76C (Equipment Records and Parts Specialist, n = 154), 76V (Material Storage and Handling Specialist, n = 167), 76W (Petroleum Supply Specialist, n = 427) and 94B (Food Services Specialist, n = 3536). All of the data for this research came from soldiers in these MOS who entered the Army between October 1980 and September 1982.

Predictors. All of the MOS in the research sample use AA composites from the ASVAB as the standard for enlistment eligibility. For three of the MOS (76C, 76V, and 76W) the appropriate AA standard was the clerical (CL) composite. This composite is currently formed by combining three ASVAB subtests; verbal ability (VE), numerical operations (NO), and coding speed (CS). The operators and food (OF) composite, formed by combining the verbal ability (VE), numerical operations (NO), mechanical comprehension (MC), and auto/shop information (AS) subtests, is used for enlistment into MOS 94B. These two composites, therefore, were used as the predictor variables in this research. (If a new CL composition, currently under discussion, is implemented, tables prepared by ARI can be used to identify equivalent AA scores between the old and new CL composites).

Criteria. The criterion measures for this research were the SQT scores obtained by the soldiers in the four MOS for whom ASVAB data were also available. As is the case with all recent SQT tests, the SQTs providing data for this research were written tests. The SQT scores used in this research were obtained during the first two quarters of the 1983 testing year.

Analyses. Two way distribution tables were calculated for AA composite and SQT scores. The composite score range was broken down into intervals of five points in length, starting at the current cutoff score for each MOS. Five point intervals were chosen because in the Army's existing classification system all cutoff scores for the AA composites are in 5 point increments. The SQT scores were broken into four categories: scores less than 60, scores greater than or equal to 60, scores greater than or equal to 70, and scores greater than or equal to 80. An SQT score of 60 is considered to be passing by the Army. If the total n for any column (ASVAB category) was less than 25 data were not entered for that column. Such data was considered, however, in calculating row (SQT) totals for each MOS.

Results and Discussion

The data from each of the four MOS are given in Tables 1 through 4. For example, Table 1 indicates there were 63 soldiers in the 76C sample with AA scores between 95 and 99. Of these 63 soldiers, 46% had SQT scores below 60% and 54% had scores at or above 60%. Of these 63 soldiers, 22% had SQT scores at or above 70% and 2% at or above 80%. In the entire sample there were 154

Table 1
Percentage of Soldiers Obtaining Given SQT Scores
By AA Composite Category

MOS 76C

Percentage SQT	Composite Category 95-99	100-104	105-109	Total 95-109
>= 80	2	10	6	6
>= 70	22	31	36	29
>= 60	54	72	64	63
< 60	46	28	36	37
Sample Size (n)	63	58	33	154

soldiers with AA scores between 95 and 109. Of this total sample, 37% obtained SQT scores below 60% and 63% scores at or above 60% on the SQT. Of the total sample of 154 soldiers 29% had SQT scores at or above 70% and 6% at or above 80%. Tables 2 through 4 can be interpreted similarly. Table 5 presents a summary of the data for the total sample from each of the four MOS.

Three things should be noted from these tables. First, performance on SQT in general is higher for soldiers with higher AA scores. In each MOS, a 5 point increase in the minimum AA score was associated with higher SQT performance. Second, SQT performance is already quite high, with 80% or more of the soldiers passing in 3 of the 4 MOS. Third, in these MOS one third or more of the soldiers had AA scores within 5 points of the minimum score for entry into these MOS.

The policy decision regarding any potential increase in AA score must weigh the relatively modest anticipated increase in SQT performance against the relatively major exclusion of previously qualified applicants. Among these MOS, only in 76C does it seem to us that these data suggest an increase in the AA cutoff merits further consideration. However, such decisions must be made based on more complete information on the MOS than these data provide. The current and future structure of the MOS, judged performance , and anticipated demands must also be weighed against the overall needs of the Army and anticipated number and qualifications of new enlistees.

Table 2
Percentage of Soldiers Obtaining Given SQT Scores
By AA Composite Category

MOS 76V

Percentage SQT	Composite Category 90-99	95-99	100-104	105-109	Total 90-109
>= 80	14	17	.	.	17
>= 70	45	55	.	.	49
>= 60	76	86	.	.	80
< 60	24	14	.	.	20
Sample size (n)	107	29	18	13	167

Table 3
Percentage of Soldiers Obtaining Given SQT Scores
By AA Composite Category

MOS 76W

Percentage SQT	Composite Category 90-94	95-99	100-104	105-109	Total 90-109
>= 80	23	27	23	37	25
>= 70	50	55	61	70	54
>= 60	79	86	89	80	82
< 60	21	14	11	20	18
Sample Size (n)	251	90	56	30	427

Table 4
Percentage of Soldiers Obtaining Given SQT Scores
By AA Composite Category

MOS 94B

Percentage SQT	Composite Category 85-89	90-94	95-99	100-104	105-109	Total 85-109
>= 80	47	52	63	69	70	54
>= 70	77	81	91	92	97	83
>= 60	94	97	98	98	98	96
< 60	6	3	2	2	2	4
Sample Size (n)	1549	903	576	294	214	3536

Table 5
Summary Values for the Four MOS

Percentage SQT	MOS 76C (95-109)	76V (90-109)	76W (90-109)	94B (85-109)
>= 80	6	17	25	54
>= 70	29	49	54	83
>= 60	63	80	82	96
< 60	37	20	18	4
Sample Size (n)	154	167	427	3536

Administrative Records as Effectiveness Criteria: An Alternative Approach

Barry J. Riegelhaupt
Carolyn DeMeyer Harris
Robert Sadacca

Human Resources Research Organization

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine-year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Paper presented at the 92nd Annual Convention of the American Psychological Association in Toronto, Canada, August 1984.

Administrative Records as Effectiveness Criteria: An Alternative Approach[1]

The accurate measurement of individual job performance is critical in personnel selection research (Dunnette, 1966; Guion, 1965). Considerable time and energy are often spent in developing predictor tests and measures at the expense of: (a) identifying performance constructs that should be the targets of the predictor measures, and (b) actually measuring, in a reliable and valid manner, the effectiveness of individuals on those performance constructs. Test validation results, however, can be meaningful only if proper attention is paid to the criterion side, so that an accurate depiction of job performance effectiveness is provided.

Performance measures can be classified into two general types: objective indexes and performance ratings. Examples of objective measures, for a clerical position, would be the number of pages typed per eight-hour day and the number of typing errors made per page. Performance ratings rely on the human judgment of an individual's job performance. Because of the subjective nature of performance ratings, objective indexes of a worker's performance are, in certain cases, preferable to ratings. Good objective measures, however, are difficult to acquire (Guion, 1965; Landy & Trumbo, 1980).

[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 necessary express the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

The difficulty with the vast majority of objective measures of performance is that they are almost invariably deficient and contaminated (Guion, 1965; Smith, 1976). By deficient, it is meant that the measure provides only a partial picture of the worker's effectiveness on the job; that is, there are important aspects of the job left untapped by the objective measure. Referring to the clerical example above, typing speed and accuracy may be an important index of effectiveness in this job, but if helping break-in inexperienced typists and willingness to work very hard during heavy production periods are also important for job success, then the former two measures, individually or together, do not adequately measure effectiveness on the job. They are deficient.

The administrative indexes that appear in Army personnel records are certainly no exception. When viewed separately, reports of AWOL, nonjudicial punishment of a serious nature (Articles 15), Certificates of Commendation, etc., tap only a part of the soldier effectiveness criterion domain and are probably deficient as indicators of effectiveness (Borman, Johnson, Motowidlo, & Dunnette, 1975; Shields, Hanser, Williams, & Popelka, 1981).

Contamination in objective measures occurs when factors that affect how well individuals do with respect to the measure are beyond their control. Referring again to the example above, suppose that the number of pages typed in a day depends to some extent on the kind of text that the typist is to work on, and the soldier has no control over those assignments. The "number of pages" measure provides an impure index of effectiveness; it is contaminated.

The most prevalent type of contamination is opportunity bias. The administrative indexes that appear in Army personnel records are possibly contaminated by opportunity bias. The number of reports of AWOL, violations of an article of the Uniform Code (Articles 15), awards, letters of commendation, etc., that appear in a soldier's record, may in part be influenced by such factors as the Military Occupational Specialty (MOS), post, organizational unit, and commanding officer (CO) to which the soldier is assigned. Therefore, comparing the effectiveness of soldiers in different MOS, assigned to different locations on the basis of administrative indexes, without information concerning differential opportunities, may be misleading. The most important question, however, is the degree to which opportunity bias, if it exists, is predictor correlated or predictor free. Predictor correlated contamination refers to a situation where the opportunity to receive letters, awards, Articles 15, etc., is influenced by a predictor score. Thus, if knowledge of a soldier's Armed Forces Qualification Test (AFQT) score impacted on the opportunity to receive awards, that would be an example of predictor correlated contamination. Brogden and Taylor (1950) have noted, that in general, opportunity bias is predictor free and while it may attenuate validity coefficients, it will not seriously distort their relative magnitude.

There exists an additional potential difficulty in using administrative records as soldier effectiveness criteria. Previous research, which has used objective performance indexes extracted from personnel files, often reports low correlations with predictors or other criteria, e.g., performance ratings. This has been found in both military (Allen & Bell, 1980; Drucker & Schwartz, 1973; Shields, et al., 1981) and non-military settings (Cascio &

Valenzi, 1978; Landy & Farr, 1975). This is often in part because administrative records reflect only exceptionally good or exceptionally poor performance. In Army personnel records, for example, consider reports of AWOL and Articles 15 on the poor performance side and awards and certificates or letters of commendation on the good performance side. A small percentage of first-tour soldiers is likely to have these performance indicators in their personnel folders. Thus, the skewed distributions found for individual, separate indexes based on administrative actions seriously constrain the usefulness of the indexes as criteria of soldier effectiveness (Hammer & Landau, 1981).

Construct Validation Approach

One strategy for dealing with these issues is to view the content of administrative indexes as critical incidents and form composites on the basis of conceptual similarities. For example, several different kinds of awards, letters, and certificates could be combined into one index if they reflect performance in some psychologically homogeneous behavioral domain. A soldier's "score" would then be the total number of such indexes received in that particular category. If measures are combined that reflect the same underlying construct, base rates might improve to a level where significant correlations with other variables would be more possible.

An indication of how the combining of individual administrative indexes might constitute a beneficial approach can be seen using data presented by Shields, et al, (1981). The researchers gathered information on soldier effectiveness in the 193rd Infantry Brigade, Panama. Data were collected on such variables

as Skill Qualification Test (SQT) scores, number of awards, number of military courses completed, number of times honor graduate status was attained in training courses, number of Articles 15, and number of letters of appreciation.

One result of the research was that positive correlations emerged between some criterion pairs--for example, SQT scores and number of awards (r=.43); number of awards and number of military courses completed (r=.63); etc. This indicates that these different indexes may indeed reflect to some extent an underlying effectiveness construct. Relationships between other pairs of indexes were low, but low base rates may have been a contributor to the low correlations in some cases. For example, less than four percent of the 125 soldiers examined had attained honor graduate status. This low base rate, in part, reduces the likelihood of significant correlations between this variable and other variables.

The above findings suggest that composites of administrative indexes formed within a soldier effectiveness conceptual framework would not only produce administrative measures with improved base rates and more variance, they would also provide an approach for managing the deficiency inherent in individual objective measures. Since, as part of the construct validity framework adopted by this project, individual administrative indexes will be used as one of several methods to index a soldier's effectiveness on one or more performance constructs, the issue of these measures being deficient as criteria when used separately would be less critical. Additionally, since the conceptual framework within which composites were formed is comprised of a set of dimensions from which rating scales were also developed, the use of

administrative indexes in this fashion is consonant with the rating scale rationale.

This paper describes the steps that were taken: (1) to determine which administrative indexes have sufficient variance and acceptable base rates to warrant consideration in the formation of criteria of soldier effectiveness, and (2) to combine these indexes within a model of soldier effectiveness, into psychometrically sound and conceptually meaningful variables. These composite indexes were expected to result in improved base rates and have more variance than the individual administrative measures.

Method

Sample Selection

In addition to examining the benefits of forming administrative index composites, as part of the larger selection and classification project, we were interested in determining whether significant differences in the frequency of administrative actions existed across MOS and posts. Accordingly, the plan was to collect records data from the Military Personnel Records Jackets (MPRJ) for a random sample of 750 soldiers, 150 in each of five MOS at five Army posts. Selected soldiers were in their first tour, and at the time of data collection had been in the Army between 10-1/2 and 27-1/2 months.

In order to strengthen the case for the generalizability of the records collection findings, the selected MOS were chosen based on their diversity. As can be seen in Table 1, each MOS represented a different Career Management Field (CMF), a different ASVAB area composite, and a different cluster. Prior to this effort, Military Occupational Specialties (MOS) had been clustered into homogeneous groups according to rated job content (Rosse, Borman, Campbell, & Osborn, 1983). Additionally, each of the five MOS has a relatively large population in the Army and is well represented by blacks. Females are also well represented with the exception of Infantryman (11B).

Table 1

MOS Selected for Records Collection

MOS	Title	CMF	Aptitude Composite	Cluster	FY81 Accessions		
					Total	Women	Blacks
05C	Radio TT Operator	31	SC	H	3176	535	898
11B	Infantryman	11	CO	G	7028	0	1128
64C	Motor Transport Operator	64	OF	P	5440	774	1279
71L	Admin. Specialist	71	CL	N	4484	2744	1967
91B	Medical Care Specialist	91	ST	O	3074	924	876

Identification of Administrative Indexes

A list of administrative measures indicative of soldier effectiveness was developed from a review of relevant Army Regulations, previous research efforts in military settings, and interviews with knowledgeable Army personnel. The list is presented in Table 2.

Table 2

Candidate List of Administrative Measures
Indicative of Soldier Effectiveness

- Reenlistment Eligibility
- Reenlistment Eligibility Bar
- Enlisted Evaluation Report (EER)
- Promotion Rate
- Number and Duration of AWOL/Desertions
- Number and Content of Articles 15
- Number and Content of Courts-Martial
- Number and Type of Awards/Badges
- Number and Content of Letters of Appreciation/Commendation
- Number and Content of Letters of Reprimand/Admonition
- Number and Content of Certificates of Achievement/Commendation
- Number and Type of Civilian Courses Attended/Completed
- Number and Type of Service Courses Attended/Completed
- Performance in Service Courses

Development Data Collection Instrument

In order to develop a data collection form that could be used for the recording of administrative measures extracted from personnel files it was necessary to conduct a detailed examination of the make-up of the MPRJ via reviews of relevant Army Regulations and interviews with knowledgeable Army personnel. Two preliminary versions of the data collection form were field tested before the final Records Collection Form was developed. A complete

guideline to accompany the form was also developed to ensure that the form could be used efficiently, unambiguously, and with consistency by each team which would be at different sites during the field data collection period.

Data Collection

The examination of Military Personnel Records Jackets was conducted by teams of two research staff members who conducted 2-day site visits to each of the five posts. At each post, MPRJ are located at the Military Personnel Office (MILPO) that serves the soldier's unit. Larger posts typically have more than one MILPO. Where this was the case, each MILPO was represented in the sample. Using the Records Collection Form and accompanying Guidelines the two days were spent extracting records data from 747 MPRJ.

Data Reduction

Of the 747 completed forms, 37 were usable, but represented MOS other than the five MOS selected for investigation. Five forms were not usable owing to incorrect entries that could not be rectified. The 742 usable forms were divided into four Batches as follows:

Training Batch	145 = 51 (FS17)* + 57 (FS29)* + 37 (Other MOS)
Batch A	200 = 153 (FS24)* + 47 (FS33)*
Batch B	199 = 125 (FS13)* + 47 (FS18)* + 27 (FS27)*
Batch C	198 = 137 (FS11)* + 61 (FS23)*

*MILPO Codes

Table 3

Dimensions Defining Overall Soldier Effectiveness

A - Controlling own behavior related to personal finances, drugs/alcohol, and aggressive acts.

B - Adhering to regulations, orders, SOP and displaying respect for authority.

C - Displaying honesty and integrity

D - Maintaining proper military appearance

E - Maintaining proper physical fitness

F - Maintaining own equipment

G - Maintaining living and work areas to Army/unit standards

H - Exhibiting technical knowledge and skill

I - Showing initiative and extra effort on the job

J - Attending to detail on jobs/assignments/equipment checks

K - Developing own job and soldiering skills

L - Effectively leading and providing instruction to other soldiers

M - Supporting other unit members

Table 4

Examples of how the Content of Letters, Certificates, and Articles 15 were Coded Within the Soldier Effectiveness Dimensions

- Maintaining Proper Physical Fitness
  - Achieving Maximum Army Physical Readiness Test score of 300
  - Finishing 2nd place on boxing team
  - Being overweight
- Maintaining Living and Work Areas to Army/Unit Standards
  - Outstanding job on Post HQ clean-up detail
  - Faiure to pass morning room inspection
- Exhibiting Technical Knowledge and Skill
  - 100% on hands-on component of SQT
  - High score in weapons qualification
  - Duty performance not such to warrant promotion consideration
  - Accidental discharge of weapon

Preliminary Work File Creation. Upon completion of the coding, the OPSCAN sheets were read, fields were edited, and frequency distributions were generated for each field. Based upon these frequencies, a set of 38 variables was created. The variables appear in Table 5.

Table 5

List of Created Variables

Variable Number	Description
G2Y4001	Has SQI, ASI, or Language Identifier
G2Y4002	Is working at skill level DMOS higher/lower than PMOS
G2Y4003	Is eligible to reenlist
G2Y4004	Highest grade attained
G2Y4005	Current grade
G2Y4006	Never demoted
G2Y4007	Number of awards
G2Y4008	M16 rating
G2Y4009	Has EXP grenade rating
G2Y4010	Number of letters/certificates
G2Y4011	Cited for exhibiting technical knowledge and skill (Construct H & J)
G2Y4012	Cited for physical and mental self development (Construct E & K)
G2Y4013	Cited for constructs other than E, H, J, and K
G2Y4014	Has had special military education
G2Y4015	Number of military training courses
G2Y4016	Years of civilian education
G2Y4017	Has high school diploma
G2Y4018	Has earned civilian education credits
G2Y4019	Number of Articles 15/FLAG actions
G2Y4020	Has been AWOL
G2Y4021	Cited for failure to adhere to rules and regulations and disrespect for authority (Construct B)
G2Y4022	Cited for failure to control own behavior (Construct A)
G2Y4023	Cited for construct violations other than constructs A and B
G2Y4024	Number of times cited for construct violations (G2Y4021 + G2Y4022 + G2Y4023)
G2Y4025	Number of times assigned extra duty
G2Y4026	Has had punishment suspended
G2Y4027	Has forfeited pay
G2Y4028	Has been restricted
G2Y4029	Has been confined
G2Y4030	Initial grade
G2Y4031	Change in grade (G2Y4005 - G2Y4030)
G2Y4032	Time period in years between first and last grade change
G2Y4033	Promotion rate (number of grades advanced per year -- G2Y4031/G2Y4032)
G2Y4034	Has received punishment
G2Y4035	Has received AAM
G2Y4036	Has received air assault badge
G2Y4037	Has received parachute badge
G2Y4038	Has received other award

Having created these variables for each case, at this point, the 597 records that were independently coded by each of three coders contained three values for each of the 38 variables. Thus, the next steps were to examine coder agreement and create a final work file which contained one value per variable per case.

Coder agreement was assessed by two methods. Table 6 presents the correlations between coders and the average intercoder correlation for each of the 38 variables. As can be seen, the product moment correlations are, for the most part, consistently high, and generally above .90. For the six variables where average intercoder correlations were lower than .90, four of the variables dealt with the assignment of the content of a letter, certificate, or Article 15 to a construct (G2V4011, G2V4012, G2V4013, G2V4023). In making

Table 6

Correlations Between Coders for Created Variables

Variable No.	Variable	C_1C_2	C_1C_3	C_2C_3	Average Intercoder r
G2V4001	Has SQI/ASI/LI	.95	.97	.96	.96
G2V4002	Has Different Skill Level -- DMOS/PMOS	.98	.91	.92	.94
G2V4003	Is Eligible to Reenlist	1.00	1.00	1.00	1.00
G2V4004	Highest Grade Attained	.97	.98	.98	.98
G2V4005	Current Grade	.97	.97	.98	.97
G2V4006	Never Demoted	.89	.87	.98	.91
G2V4007	Number of Awards	1.00	1.00	1.00	1.00
G2V4008	M16 Rating	.97	.99	.97	.93
G2V4009	Has EXP Grenade Rating	.99	.99	1.00	.99
G2V4010	Number of Letters/Certificates	.97	.98	.99	.98
G2V4011	Number of Times Cited for Technical Knowledge and Skill	.89	.86	.87	.87
G2V4012	Number of Times Cited for Physical and Mental Self Development	.77	.76	.87	.80
G2V4013	Number of Times Cited for Other Constructs	.78	.70	.72	.73
G2V4014	Has Had Special Military Education	.81	.80	.93	.85
G2V4015	Number of Military Training Courses	.91	.95	.92	.93
G2V4016	Number of Years of Civilian Education	1.00	1.00	1.00	1.00
G2V4017	Has High School Diploma	.90	.91	.96	.92
G2V4018	Has Earned Civilian Education Credits	.75	.71	.89	.78
G2V4019	Has Received Article 15/FLAG	.99	.98	.98	.98
G2V4020	Has Been AWOL	.88	.84	.97	.90
G2V4021	Cited for Failure to Adhere to Regulations/ Disrespectful	.87	.89	.94	.90
G2V4022	Cited for Failure to Control Own Behavior	.92	.92	.93	.92
G2V4023	Cited for Other Construct Violation	.80	.78	.89	.84
G2V4024	Number of Times Cited for Construct Violations	.97	.97	.99	.98
G2V4025	Has Received Extra Duty	.99	.99	1.00	.99
G2V4026	Has Had Punishment Suspended	.94	.93	.93	.93
G2V4027	Has Forfeited Pay	.99	.99	1.00	.99
G2V4028	Has Been Restricted	.99	.99	1.00	.99
G2V4029	Has Been Confined	.90	.95	.95	.93
G2V4030	Initial Grade	.99	.99	1.00	.99
G2V4031	Change in Grade	.96	.97	.98	.97
G2V4032	Number of Years First to Last Grade Change	.99	.99	.99	.99
G2V4033	Promotion Rate (Grades Advanced/Year)	.93	.94	.97	.95
G2V4034	Has Received Punishment	.98	.98	.99	.98
G2V4035	Has Received AAM	1.00	1.00	1.00	1.00
G2V4036	Has Received Air Assault Badge	1.00	.99	.99	.99
G2V4037	Has Received Parachute Badge	1.00	1.00	1.00	1.00
G2V4038	Has Received Other Award	1.00	1.00	1.00	1.00

n = 508.

these assignments, coders had only a preliminary definition for each construct. It is anticipated that when definitions are refined, and rating scale points, anchored with behavioral examples of each construct are available, correlations would improve to levels above .90. For the remaining two variables (G2V4014 and G2V4018), the distinction between Special Military Education and Civilian Credits was complicated by the fact that certain military courses were taken at or through civilian colleges and universities. In future data collections, military education will be counted as such, regardless of where courses were actually taken.

In Table 7, the results of a one-way analysis of variance performed on each of the 38 variables are presented. Once again the findings reflect high coder agreement. For the nine variables for which statistically significant coder differences were found, inspection of the means revealed differences among coders that are not at all alarming in size. For example, mean differences among coders of only .034, .018, .033, and .018 were found for variables G2V4011, G2V4012, G2V4013, and G2V4014 respectively. Not only are these differences relatively unimportant but as just mentioned, the circumstances that produced the significant differences are not expected to influence future data collections. Taken together, the results of the correlational analyses and the analyses of variance provide sufficient support for the conclusion that only one researcher will be required to collect administrative measures from each Military Personnel Records Jacket in future large-scale data collection efforts.

Table 7

Results of One-Way ANOVA for Created Variables

Variable No.	Variable	F
G2V4001	Has SQI/ASI/LI	1.13
G2V4002	Has Different Skill Level -- OMOS/PMOS	1.41
G2V4003	Is Eligible to Reenlist	-
G2V4004	Highest Grade Attained	<1
G2V4005	Current Grade	1.12
G2V4006	Never Demoted	2.05
G2V4007	Number of Awards	1.00
G2V4008	M16 Rating	4.13**
G2V4009	Has EXP Grenade Rating	2.34
G2V4010	Number of Letters/Certificates	<1
G2V4011	Number of Times Cited for Technical Knowledge and Skill	4.39**
G2V4012	Number of Times Cited for Physical and Mental Self Development	4.36**
G2V4013	Number of Times Cited for Other Constructs	7.44**
G2V4014	Has Had Special Military Education	7.48**
G2V4015	Number of Military Training Courses	12.96**
G2V4016	Number of Years of Civilian Education	-
G2V4017	Has High School Diploma	3.01*
G2V4018	Has Earned Civilian Education Credits	2.16
G2V4019	Has Received Article 15/FLAG	1.00
G2V4020	Has Been AWOL	<1
G2V4021	Cited for Failure to Adhere to Regulations/ Disrespectful	<1
G2V4022	Cited for Failure to Control Own Behavior	3.01*
G2V4023	Cited for Other Construct Violation	<1
G2V4024	Number of Times Cited for Construct Violations	1.58
G2V4025	Has Received Extra Duty	1.00
G2V4026	Has Had Punishment Suspended	<1
G2V4027	Has Forfeited Pay	1.00
G2V4028	Has Been Restricted	1.00
G2V4029	Has Been Confined	<1
G2V4030	Initial Grade	1.51
G2V4031	Change in Grade	2.97*
G2V4032	Number of Years First to Last Grade Change	1.55
G2V4033	Promotion Rate (Grades Advanced/Year)	1.06
G2V4034	Has Received Punishment	1.00
G2V4035	Has Received AAM	-
G2V4036	Has Received Air Assault Badge	1.00
G2V4037	Has Recieved Parachute Badge	-
G2V4038	Has Received Other Award	-

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

Final Work File Creation. Two decision rules were used to obtain the desired one value per variable per case. For the dichotomous variables, a coder agreement rule was employed where majority ruled. For example, if all three coders had assigned a value of 1 for a variable, or if two out of the three coders had assigned a 1, a value of 1 was given to that variable. For the continuous variables, the assigned value was the average of the three coders rounded to the nearest whole number.

At this point, the 17-month time band was reduced to 13 months to more accurately reflect the time that soldiers in the actual FY83/84 first tour data collection will be in the Service. Only those soldiers who entered the Army between 1 July 81 - 31 July 82 at an initial grade of PFC or less were retained. This reduced the sample from 597 to 553. Additionally, 97 of the 145 records used in the training session were those of soldiers in the five MOS, and were added to the sample. The result was a sample of 650 soldiers in the 11B, 05C, 64C, 71L, or 91B MOS who had been in the Army between 14 and 27 months.

Results

An important issue in the determination of the usefulness of criterion and predictor measures is the capability of discriminating between levels of effectiveness of job performance among personnel. If everyone gets about the same score on some measure of job performance, there is practically no variance on that measure, and it is therefore incapable of discriminating levels of job performance. Thus, a first step in determining the usefulness of the administrative variables collected from personnel files, was to select those measures with an acceptable amount of variance.

Since many of the variables are components of larger summary measures, the correlations among variables were also an important criteria for selecting useful administrative measures. The product moment correlations among the administrative variables are presented in Table 8.

Based upon the frequencies and correlations and the regulations governing reenlistment and promotion criteria, six variables were selected as potentially useful criteria of soldier effectiveness. The six measures were:

- Eligible to Reenlist
- Number of Letters/Certificates
- Number of Awards
- Number of Military Training Courses
- Has Received Article 15/FLAG Action
- Promotion Rate (Grades Advanced/Year)

TABLE 8

Correlation Coefficients of Administrative Variables[ab]

Var. No.	Variables	V01	V03	V06	V07	V08	V09	V10	V11	V12	V13	V15	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V33	V34	V35	V36	V37	V38
V01	Yrs SQI/ASI/CI	-																											
V03	Eligible to Reenlist		-																										
V06	Career Denoted Indicator		.30	-																									
V07	Number of Awards	.45	.08	.65	-																								
V08	PIG Rating	.23		.30	.20	-																							
V09	EAP Overall Rating	.31			.25	.08	-																						
V10	No. of Letters/Certificates				.18			-																					
V11	Cited: Tech Knowl. & Skill				.17			.80	-																				
V12	Cited: Phys & Ment Self Deve				.09			.44	.13	-																			
V13	Cited: Other Constructs							.58	.18	.20	-																		
V15	Military Training Courses	.49			.51	.18	.22	.11	.08	.10		-																	
V19	Received Article 15/FLAG		-.45	-.46	-.10								-																
V20	Has Been AWOL		-.32	-.10									.46	-															
V21	Cited: Failure Adhere to Regs		-.27	-.31									.71	.18	-														
V22	Cited: Failure to Control Behavior		-.17	-.15									.38		.15	-													
V23	Cited: Other Construct Violation		-.30	-.08			-						.23				-												
V24	No. Times Cited, Construct Violation		-.50	-.39	-.09								.84	.39	.71	.54	.45	-												
V25	Received Extra Duty		-.39	-.37									.80	.41	.55	.28	.20	.65	-											
V26	Had Punishment Suspended		-.30	-.22									.64	.23	.54	.38	.15	.59	.55	-										
V27	Forfeited Pay		-.42	-.38	-.08			-.08			-.08		.87	.45	.65	.38	.20	.78	.77	.66	-									
V28	Been Restricted		-.31	-.35							-.09		.69	.37	.51	.25	.18	.60	.75	.42	.67	-								
V29	Been Confined		-.34	-.21									.31	.18	.20	.19	.15	.39		.25	.37		-							
V33	Promotion Rate		.16	.36									-.22	-.14	.20			-.19	.18		-.19	-.20	-.12	-						
V34	Received Punishment		-.46	-.45									.94	.45	.70	.40	.23	.64	.64	.69	.93	.70	.38	-.22	-					
V35	Received AAM				.53			.10	.10																	-				
V36	Received Air Assault Badge				.25							.23	-.08													-.08	-			
V37	Received Parachute Badge	.74			.61	.22	.33	.10		.08		.61																-		
V38	Received Other Award	.15	.08		.59	.11		.17	.15		.12	.17									-.08			.08		.09		.20	-	

[a] G2V4012, G2V4013, G2V4015, G2V4014, G2V4016, G2V4017, G2V4018, G2V4030, G2V4031, and G2V4032 served only as interim variables and as such are not shown in this table.

[b] Only correlations significant at the .05 level appear in this table.

Eligible to Reenlist. Ninety percent of the sample was eligible for reenlistment at the time of data collection. In addition to the acceptable amount of variance found for this measure, the factors considered in determining a soldier's reenlistment eligibility make this index a potentially excellent summary variable that can serve as both a useful criterion and an in-service predictor.

Number of Letters/Certificates. Of the soldiers sampled, 17 percent had one letter or certificate, and almost 12 percent had two or more. Although the original plan had been to relate each letter or certificate to one or more constructs to create construct variables, base rates were too low. Additionally, as expected, the product moment correlations presented in Table 8 between the variables which reflected the content of letters and certificates (G2Y4011-G2Y4013) and the Number of Letters/Certificates Received variable by a soldier were quite high.

Since knowing whether a soldier had ever been recognized for outstanding performance was viewed as more meaningful than knowing if recognition had occurred once or twice, a dichotomous variable, Has Received Letter/Certificate, was created. The likely impact that letters and certificates have on reenlistment and promotion decisions further establishes this variable as a potentially useful indicator of soldier effectiveness.

Number of Awards (e.g., Army Achievement Medal). Similar to the Number of Letters/Certificates variable, this summary variable also exhibited greater variance than its components viewed individually (G2V4035-G2V038). Again, as expected from the part/whole relationships involved, the correlations between Number of Awards and the variables representing each type of award were quite high. Since awards and decorations are used formally for promotion decisions to E5 and above, and likely are considered for promotions from E1 to E4, the index was transformed into a dichotomous variable, Has Received Award, and selected for further analyses.

Number of Military Training Courses (e.g., Drill Corporal Program, Patient Care Procedures). The weight given to training courses in promotion decisions and the finding that 20 percent of the sample had one training course and 6 percent had 2 or more courses, made this a variable worthy of further examination. As before, it was viewed as more meaningful to know whether a soldier had or had not completed military training courses than knowing whether one or two courses had been completed. Thus, a dichotomous variable Has Had Military Training Courses was created.

Has Received Article 15/FLAG Action. In addition to finding that 11 percent of the soldiers sampled had been

either cited for a violation of the Uniform Code or had a personnel action pending, this measure, as expected, was negatively correlated with positive indicators of performance. For example, correlations of -.45 and -.46 were found between this variable and Reenlistment Eligibility and the Never Demoted Indicator, respectively.

Promotion Rate. A relatively normal distribution of promotion rates was obtained with a mean/median of about two grades per year. In addition, this variable's relationship with other measures was generally as expected. Positive relationships were found between Promotion Rate and Reenlistment Eligibility ($r = .16$) and the Never Demoted Indicator ($r = .36$); whereas negative correlations were found with Number of Articles 15/FLAG Action ($r = -.22$) and Has Been AWOL ($r = -.16$).

As was the case with Reenlistment Eligiblity, in addition to finding an acceptable amount of variance and expected relationships with other variables, the factors considered in making promotion decisions make this index a potentially excellent summary variable for distinguishing levels of effectiveness among soldiers.

Discussion

An often cited shortcoming of using performance measures obtained from personnel records is the skewed distributions which result from measures that typically reflect only very good or very bad performance. This was found to be the case in this investigation as well. For example, when viewed individually, Army Achievement Medals, Air Assault Badges, etc. have very low base rates, and thus skewed distributions. However, when combined into the dichotomous variable, Has Received Award, the base rate improved to a level where significant and meaningful relationships with other variables might be possible. Similar results were found for Has Received Letter/Certificate, and Has Had Military Training Courses. When letters or certificates of appreciation, achievement, or commendation were viewed independently, base rates were very low. However, when combined into one composite index, base rates improved considerably.

The original strategy to combat low base rates had been to consider the content of letters, certificates, Articles 15, etc., as critical incidents and to combine indexes that reflected the same underlying constructs. Analyses would then proceed on the constructs, rather than the index. When this was done, however, base rates did not show enough improvement to warrant further analysis at the level of constructs. The decision to create variables comprised of administrative indexes instead of performance examples, however, followed the same general strategy, and produced the desired result. Composite index measures were created, base rates were improved, and the potential for detecting significant and meaningful relationships with other variables is more likely.

While the attempt to create variables within the Model of Soldier Effectiveness (Borman, et al., 1983) by collapsing across indexes met with less than optimal success, considerable merit exists in knowing the content of a letter, certificate or Article 15. Knowledge of the content or "why" a soldier's performance received recognition or resulted in a disciplinary action will permit an evaluation of the convergent validity of other measures. For example, if a soldier received a Letter of Commendation for exhibiting outstanding technical skills, one would expect that soldier to receive a positive rating on that dimension. Similarly, if a soldier received an Article 15 for possession of marijuana, one would expect convergence between that information and the evaluation on the corresponding dimension. Finally, convergence would be expected between letters or certificates that recognized technical knowledge and scores on paper and pencil knowledge tests. Evaluations of this type, however, must await future data collections.

References

Allen, J.P., & Bell, B. (1980, July). Correlates of military satisfaction and attrition among Army personnel. Alexandria, VA: U.S. Army Research Institute.

Borman, W.C., Johnson, P.D., Motowidlo, S.J., & Dunnette, M.D. (1976). Measuring motivation, morale and job satisfaction in Army careers. Minneapolis: Personnel Decisions, Inc.

Borman, W.C., Motowidlo, S.J., & Hanser, L.M. (1983, August). Developing a model of soldier effectiveness: A strategy and preliminary results. Paper presented at the meeting of the American Psychological Association, Anaheim, CA.

Brogden, H.E., & Taylor, E.K. (1950). The theory and classification of criterion bias. Educational and Psychological Measurement, 10, 159-186.

Cascio, W.F., & Valenzi, E.R. (1978). Relations among criteria of police performance. Journal of Applied Psychology, 63, 22-28.

Drucker, E.H., & Schwartz, S. (1973, January). The prediction of AWOL, military skills, and leadership potential (HumRRO TR-73-1). Alexandria, VA: Human Resources Research Organization.

Dunnette, M.D. (1966). Personnel selection and placement. Belmont, CA: Wadsworth.

Guion, R.M. (1965). Personnel testing. New York: McGraw-Hill.

Hammer, T.H., & Landau, J. (1981). Methodological issues in the use of absence data. Journal of Applied Psychology, 66, 574-581.

Kavanaugh, M.J., MacKinney, A.C., & Wolins, L. (1971). Issues of managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 75, 34-49.

Landy, F.J., & Farr, J.F. (1975). Police performance appraisal, Technical Report, LEAA.

Landy, F.J., & Trumbo, D.A. (1980). Psychology of work behavior. Homewood, IL: Dorsey.

Lawler, E.E., III. (1967). The multi-trait-multi-rater approach to content validity. Journal of Applied Psychology, 51, 369-381.

Rosse, R.L., Borman, W.C., Campbell, C.H., & Osborn, W.C. (1983, October). Grouping Army occupational specialties by judged similarity. Paper given at the Military Testing Association.

Shields, J.L., Hanser, L.M., Williams, E.W., & Popelka, B.A. (1981, October). Pilot research for validation of ASVAB and enlistment standards against performance on the job. Paper presented at the Military Testing Association.

Smith, P.C. (1976). Behaviors, results, and organization effectiveness: The problem of criteria. In MD. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally.

Factors Relating to Peer and Supervisor Ratings of Job Performance

Walter C. Borman
Personnel Decisions Research Institute

Leonard A. White and Ilene F. Gast
Army Research Institute

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide information and procedures required to meet military manpower challenges of the future by enabling the Army to enlist, allocate, and retain the most qualified soldiers. This research is funded primarily by Army Project Number 20263731A791 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Paper presented at the 92nd annual meeting of the American Psychological Association in Toronto, Canada, August 1984.

All statements expressed in this paper are those of the authors and do not represent the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Factors Relating to Peer and Supervisor Ratings of Job Performance

The measurement of job performance has long been an important issue to industrial psychologists. Performance measures serve as a basis for personnel decisions ranging from disciplinary actions to promotions, are used as criteria in personnel research, and provide feedback to employees and to the organization on past accomplishments and training needs. A recent review of test validation efforts (Landy & Trumbo, 1980) revealed that ratings were used as criteria in 75% of the research reviewed. Yet, despite the widespread use of ratings, much remains to be learned about what appraisals are based on and the relationship of ratings to other methods of performance measurement.

For years performance appraisal research focused on psychometric considerations. In particular, efforts to improve the psychometric properties of ratings often centered on format characteristics and rater training procedures (e.g., Landy & Farr, 1980; Schwab, Henman, & Decotiis, 1975).

Recently, attention has turned to the performance appraisal process (e.g., Borman, 1983; Feldman, 1981; Ilgen & Feldman, 1983). Questions are being asked regarding the mechanisms of halo (Cooper, 1981), attributions raters make in judging others' performance (Feldman 1981), rater individual differences associated with rating accuracy (Borman, 1979a), and aspects of rating "style" (Banks, 1979). The notion is that by learning about the judgment process raters employ in making performance evaluations, we will achieve a better understanding of the variables that account for variance

in performance ratings and gain enough understanding of the process to improve the accuracy of performance appraisals.

A Causal Analysis of Supervisor Ratings

The present research follows in this trend. More specifically, it can be viewed as addressing a challenge made by Guion (1983) in his comments on Hunter (1983). In a meta-analysis of 14 studies, Hunter (1983) used causal analysis techniques to identify relationships among three variables relevant to work performance: cognitive ability, job knowledge and job performance as measured by job sample tests and supervisor ratings. The analysis showed supervisor ratings to be related to both ability to do the job under a standard set of conditions (i.e., work sample tests) and job knowledges required for effective performance. From a job performance measurement perspective, this pattern of findings is encouraging because some convergence was obtained among measures relevant to job performance. However, in the model, the multiple correlation for the prediction of supervisor ratings of overall job effectiveness from job knowledge and demonstrated task proficiency was only 0.42. Thus, factors other than those examined by Hunter (1983) would appear to account for a large portion of the variance in ratings. To uncover these "other influences", Guion (1983) suggested that the Hunter model be enlarged to examine the possible impact of a wide range of interpersonal and rater-ratee relationship factors on ratings. He stated that attention should also be given to factors that are presumably not job-related, but that might have an impact on ratings.

A Research Opportunity

Typically, validation efforts are unlikely to include all of the measures or sufficient Ns needed to extend the Hunter (1983) research. Fortunately, a large scale US Army research project now underway has provided the authors with an opportunity to take on the challenge by Guion (1983). In general, this project is concerned with improving the selection, classification, and utilization of Army enlisted personnel. A major focus of this effort is the development and administration of new, comprehensive measures of soldier performance to include (a) job knowledge tests, (b) work samples, and (c) supervisor and peer ratings of Army-wide and MOS-specific performance based on newly developed rating scales. Within the context of this larger project, the present research examines relationships among ratings, job knowledges, and work sample tests in two Army jobs and will extend the Hunter (1983) research in two ways: first, by investigating correlates of both supervisor and peer ratings of the performance of first term Army enlisted; and second, by enlarging the Hunter (1983) model to examine how supervisor and peer evaluations are influenced by variables such as personal characteristics of the ratee (e.g., social skill) and rater-ratee relationships (e.g., friendship).

Interviews With Army Raters

As a part of this research, 25 non-commissioned officers (NCOs) and officers were interviewed to elicit ideas about factors that influence Army job performance ratings and to assess the importance of each factor. As we expected, many component job performance factors such as technical competence and consistent performance of assigned duties were mentioned as important,

along with "good soldier" factors, such as discipline (e.g., following orders), effort/initiative and physical fitness. Job knowledge was reported to be relevant to the extent that soldiers were willing to apply it to performance on-the-job. In addition, on the basis of the interviews and the research literature (e.g., Guion, 1983; Landy & Farr, 1980) several interpersonal/relationship factors were identified that might influence ratings. These factors included (a) friendship/knowing between rater and ratee (Hollander, 1956; Love, 1981), (b) mutual trust and support between rater and ratee (Graen & Cashman, 1975), (c) ratee social skill, and (d) characteristics of ratees (e.g., moodiness) which may influence evaluation by affecting the image raters have of ratees.

The Research Approach

To summarize, the main objective of this work was to explore relationships between the overall job performance and ratings on various factors identified as potential influences on job performance ratings. Peer and supervisor raters were considered separately as were the two Army jobs since the rating source and nature of the job could potentially affect the obtained relationships.

Figure 1 presents the relationships we explore in this research within each job and rating source. Briefly, it is hypothesized that overall job performance ratings are a function of component job performance factors, interpersonal relationship factors, "good soldier" factors, and job knowledge and skill factors. Based on the results of research (e.g., Zammuto, London, & Rowland, 1982), and interviews with Army raters, it was anticipated that both component job performance factors and "good soldier"

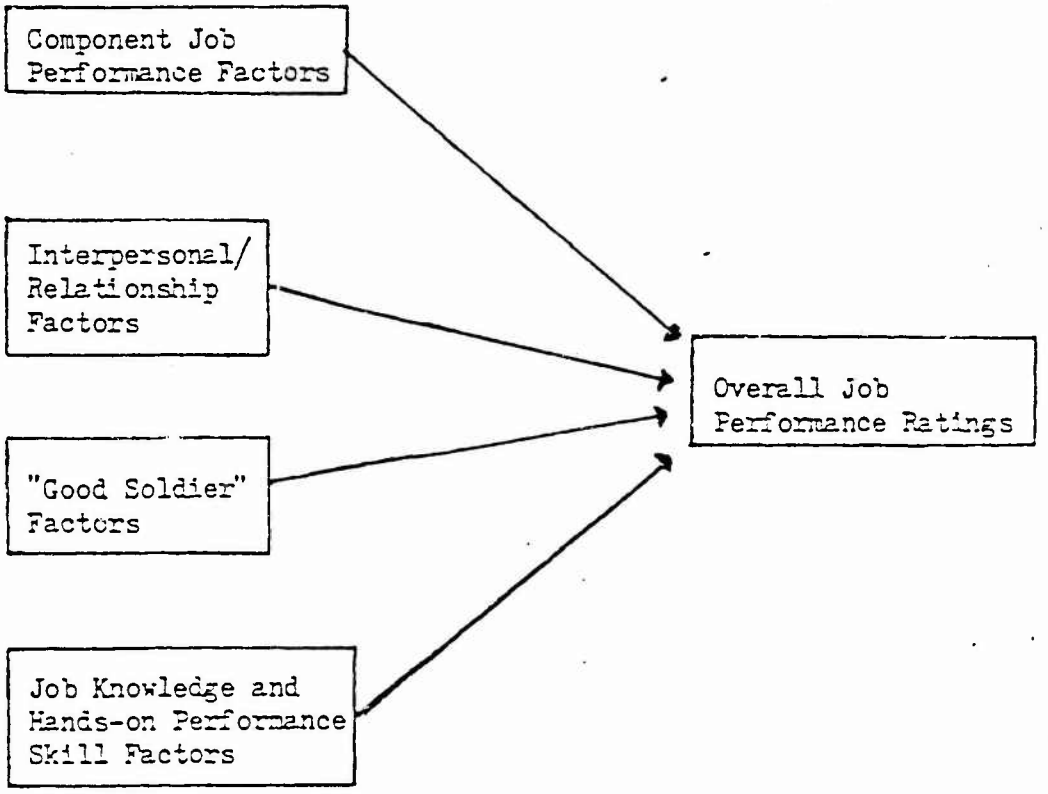


Figure 1: Possible Factors Affecting Ratings of Overall Job Performance

factors would be highly related to ratings. Relative to job component factors, correlations of performance ratings with job knowledges and specific task skills were expected to be somewhat lower (cf. Hunter, 1983). With respect to interpersonal relationship factors, ratees perceived to be loyal and willing to back up the rater were expected to receive much higher ratings than those who were viewed as less supportive (e.g., Graen & Cashman, 1975). Correlations of relatively lower magnitude (r = .30-.40) were anticipated between ratings of ratee-rater friendship/knowing and performance appraisals (Love, 1981).

Again, this work is exploratory largely because of the relatively small Ns. Subsequent investigations in this project will provide substantially larger numbers of subjects. Nonetheless, results should suggest the most important factors contributing to peer and supervisor ratings.

Method

Instrument development

The first step in this research was to develop rating scales to measure (a) performance on all relevant job factors and overall job performance; (b) effectiveness as a soldier on "Army-wide" dimensions, factors presumably relevant to a broader effectiveness construct, including "good soldiering" and overall contribution to unit effectiveness; and (c) personal characteristics and perceptions of the rater-ratee relationship. Also developed for the research were hands-on job sample tests for 15 critical tasks (for each job) and a job knowledge test for each of the two jobs.

Job performance rating scales. Critical incidents workshops were conducted with NCO, first-line supervisors for each of the target jobs. For the administrative specialist job, 65 NCO generated a total of 989 examples of effective, mid-level, and ineffective job performance. For the military police job, 84 NCO wrote a total of 1,183 behavioral examples reflecting all different levels of job performance. We then employed a variant of the behaviorally anchored rating scale procedure (Smith & Kendall, 1963) to develop behavior-based scales for each job. These procedures resulted in behavior summary scales (Borman, 1979b) for the administrative specialist job (e.g., Establishing/Maintaining Files) and behavior summary scales for the military police job (e.g., Traffic Control/Enforcement). In addition, an overall job performance dimension was developed for each of the two jobs.

Army-wide performance rating scales. To develop these scales, 77 NCO and junior officers working in a wide variety of Army jobs generated 1,215 behavioral examples. The examples represent those aspects of soldier effectiveness that contribute broadly speaking to organizational effectiveness, such as following orders and regulations. The target criterion space for these scales went beyond job performance to include aspects of socialization and commitment to the organization. Eleven behavior-based rating scales emerged from this effort (e.g., Leading/Supporting).

Hands-on-job-sample tests. For each of the two jobs, 15 critical tasks representative of the entire task domain were the target for test development work. Job sample tests were prepared for each of the tasks. Examples for the administrative specialist job are typing a memorandum or

filing documents and correspondence. Establishing and operating a road-block and checkpoint and performing preventive maintenance on a jeep are two examples for the military police job.

Each task has several procedure and performance steps, and each step is scored pass or fail. A proportion-passed score was derived for a testee on each task and the proportions averaged across tasks to yield an overall hands-on test score.

Job knowledge test. Important knowledge areas for each of the two jobs were carefully identified in job analysis work, and items intended to tap those knowledges were written with the help of subject matter experts. For each soldier, the overall job knowledge test score was the percentage of correct answers on the test.

Interpersonal/Relationship factors (Rating Questionnaire). This instrument was used to measure rater perceptions of ratee characteristics (e.g., social skill) and rater-ratee relationship factors.

The measure of mutual support/loyality was a 4-item scale based on the work of Graen and his associates (e.g., Dansereau, Graen, & Haga, 1975). Examples of items included, "This soldier is willing to back me up if I need it" (1 = Definitely yes/5 = Definitely not) and "I can trust and depend on this soldier" (1 = Strongly agree/5 = Strongly disagree). Mean scale interitem correlations were, $\bar{r}$ = .62 for peers, and $\bar{r}$ = .66 for supervisors.

Ratee social skills were assessed using a 5-item scale (Lewinshon, Mischel, Chaplin, & Barton, 1980). Items on the scale were attributes, such as, "good sense of humor" and "assertiveness". For each item, raters indicated how accurately the attribute described the ratee. Ratings were made on a 5-point scale (1 = Extremely accurate/5 = Extremely inaccurate). Mean interitem correlations were, $\bar{r}$ = .45 for peer raters, and $\bar{r}$ = .52 for supervisors.

The measure of rater-ratee friendship/knowing was based on a scale developed by Love (1981). Examples of items included, "How well do you know this soldier" (1= Extremely well/5= Not at all) and for peers only, "How much do you like this soldier" (1= Extremely well/5= Not at all). Mean interitem correlations for this measure were, $\underline{r}$ =.55, for peers, and $\underline{r}$ =.52, for supervisors.

Other ratee attributes assessed were (a) "bootlicker" (for peers only), (b) know-it-all (for peers only), (c) outgoing, (d) athletic, (e) complains a lot, (f) even-tempered, (g) moody. Raters used a 5-point scale (1 =Extremely accurate/5 =Extremely inaccurate) to report how accurately each adjective described the ratee. Responses to the last three items above were combined to form a measure of emotional stability. Mean interitem correlations for these attributes were, $\underline{r}$ =.28, for peers, and $\underline{r}$ =.26, for supervisors.

Finally, rater expectations of ratee combat performance were assessed by responses to the item, "Compared to other soldiers in his/her MOS, how would you expect this soldier to perform in a combat situation" (1=Top 10%, 2=Upper 33%, but not in top 10%, 3=Average, 4=Lower third, but not in bottom 10%, Lowest 10%).

Subjects

Participants in the research included 102 first-term soldiers in two jobs, 60 administrative specialists and 42 military police men and women. In the total sample, 69 were male and 33 were female; 34 blacks, 4 hispanics, 1 native American, and 63 whites participated. Importantly,

each of the two samples was randomly selected from units assigned to participate in the research. Specifically, 140 administrative specialists formed the population in the units assigned, and 60 were randomly selected from the 140. All but five participated in the research and five others were substituted. Forty-five military police men and women were selected randomly from the population of 95 available in the units assigned, and all but three participated.

Regarding peer ratings, few of the administrative specialists worked together, and therefore, 45 of that group received a total of 92 sets of peer ratings (1 - 5 for each subject or a mean of approximately two per ratee for those rated). Military police participants generally worked more closely together in three different company-sized units. Thus, we first determined all possible peer rater assignments to members of the sample and then randomly made peer rating assignments such that each rater had roughly four sets of ratings to make. Each member of the sample received approximately four sets of ratings. In all, 160 peer ratings were provided for the 42 military police subjects.

For the supervisor ratings, 55 of the administrative specialists were evaluated by 1 supervisor and the remaining 5 received ratings from 2 of their supervisors. Raters were all first-line, immediate supervisors, NCO or junior officers. With the military police group, all but 4 of the 412 soldiers were rated by 2 supervisors, generally, their first-line NCO supervisor and an NCO or junior officer one level higher. However, both sets of supervisors work closely with the first-termers on this job, and all supervisors expressed confidence that they could make fair and accurate performance appraisals of their subordinates.

Table 1

Correlations Between Ratings on Component Job Performance Factors and Overall Job Performance

	Administrative Specialist	
Job Factors	Peer (n=45)	Supervisor (n=60)
Preparing, Typing, Proofreading	.69	.80
Distributing/Dispatching Documents	.58	.85
Maintaining Office Resources	.59	.74
Posting Regulations	.51	.80
Establishing/Maintaining Files	.30	.69
Keeping Records	.55	.78
Security of Classified Documents	.54	.49
Customer Service	.75	.70

	Military Police	
	Peer (n=42)	Supervisor (n=60)
Traffic Control/Enforcement	.78	.66
Providing Security	.67	.62
Investigating Crimes/Making Arrests	.77	.73
Patrolling	.74	.49
Providing Good Public Image	.78	.73
Interpersonal Communications	.76	.58
Medical Emergencies	.75	.62

job, with one exception. Peer ratings on Establishing/Maintaining Files do not appear to correlate as highly with the overall job performance ratings as do peer ratings on many of the other factors.

Table 2 depicts correlations between the overall performance ratings and ratings on the interpersonal/relationship factors. Note that perceived support from the ratee is related quite highly with job performance ratings (administrative specialist peers are something of an exception). Rated more highly is the performance of those soldiers who are perceived as backing up the rater, being someone he/she can trust and depend on, and for supervisor raters as someone who supports his/her decisions. With respect to interpersonal competence, ratees viewed as socially skilled, assertive, likeable, and as having a good sense of humor, were rated more highly by their supervisors, particularly for the administrative specialist job.

Also of interest, knowing and being friends with the ratee is not very highly related with perceptions of overall job performance, with $r < .40$, in all cases. This is consistent with previous findings on peer evaluations (e.g., Hollander, 1956; Love, 1981) and may extend to supervisor ratings as well. Likewise, perceptions of emotional stability (e.g., moody) are not correlated highly with job performance ratings. Perceptions of being even-tempered or moody apparently have little to do with how job performance is evaluated. Finally, as shown in Table 2, correlations between ratings of job performance and ratee attributes of athletic, outgoing, a "bootlicker", and a know-it-all, were less than 0.35 in all cases.

Table 2

Correlations Between Ratings on Interpersonal/Relationship Factors and Overall Job Performance

	Administrative Specialist		Military Police	
Rating Factors	Peer	Supervisor	Peer	Supervisor
Mutual Support/Loyality [a]	.35	.54	.58	.68
Knowing/Friendship[a]	.18	.35	.38	.28
Ratee Characteristics				
Social Skill[a]	.19	.54	.42	.49
Emotional Stability[a]	.13	.06	.35	.28
Athletic	.29	.19	.04	.16
Outgoing	.35	.30	.19	.22
Bootlicker	-.19	-	-.35	-
Know-it-all	.07	-	-.32	-

[a]These are based on data from responses to 2-5 items.

Table 3 contains correlations between ratings on the Army-wide, "good soldier" factors and overall job performance ratings. The highest correlates are technical competence, leading/supporting, and expected combat performance, although all of the factors except physical fitness are quite highly related to performance ratings.

Table 4 presents correlations between the job knowledge and job sample test scores and the rating measures. In general, these correlations are low. This is somewhat in contrast to Hunter (1983) who found mean correlations (uncorrected) of job knowledge and job sample test scores with supervisor ratings to be .25. The correlation of job knowledge with peer ratings for the administrative specialist job was somewhat higher, and similar in magnitude to those reported by Hunter. Of course, the small number of Ns here rules out any definitive statement about these results. In addition, the job knowledge tests and work samples are in the preliminary stage of development. Thus, scores on these measures must be interpreted with extreme caution.

Discussion

Relationship Between Rating Factors and Overall Job Performance

This research explored relationships between overall job performance and various factors thought to influence these ratings. Of the factor sets studied here, the component job performance factors have in general the highest correlations with overall job performance (Mdn r_s = .57, .76, .76, and .62 respectively, for peer and supervisor, administrative specialist, peer and supervisor, military police). Also, of the "good soldier" factors, technical competence was the highest or almost the highest correlate with overall job performance. Conceptually, that factor is definitely the most

Table 3

Correlations Between Ratings on Army-Wide, Good Soldier Factors and Overall Job Performance

	Administrative Specialist		Military Police	
Army-Wide Factors	Peer	Supervisor	Peer	Supervisor
Technical Competence	.55	.82	.75	.76
Compliance with Rules and Regulations	.41	.59	.54	.69
Motivation/Effort	.43	.64	.52	.68
Maintaining Activities (Self, Equipment, Quarters)	.30	.62	.52	.66
Leading/Supporting	.45	.64	.75	.78
Physical Fitness	.20	.27	.29	.11
Expected Combat Performance	.56	.54	.66	.70

Table 4

Correlations Between Job Knowledge and Job Sample Test Scores and Overall Job Performance

	Administrative Specialist		Military Police	
Measure	Peer (n=45)	Supervisor (n=60)	Peer (n=42)	Supervisor (n=42)
Job Knowledge Test	.24	.13	-.11	.10
Job Sample Test	.08	-.11	.17	.02

job relevant in the "good soldier" set. In comparison, the median correlations between "good soldier" factors and overall job performance (with technical competence and motivation/effort removed because they are directly job related) are r = .49, .59, .54, and .69, for the same rating source/job combinations.

In general, the interpersonal/relationship factors generally correlate lower with overall job performance than do the directly job-related component job performance factors, with the exception of the dyadic loyality factor, especially for the military police job, as was noted previously. Also, the perception of social skills was associated with higher ratings from supervisors, particularly for the administrative specialist job. The successful performance of Customer Service (for the administrative specialist) and Interpersonal Communication (for military police) seems likely to require social skills and the ability to conduct smooth, effective relationships with the public (Hogan, Hogan, Busch, 1984). Thus, this finding may simply reflect raters awareness that ratees who are socially skilled are in fact better performers. Of course, it is also possible that ratees' interpersonal charm independent of their performance contributed to higher ratings. Of interest here, perceptions of knowing and friendship do not correlate very highly with the job performance ratings, relative to other factor sets. Similarly, whether ratees are seen as moody, outgoing, or athletic, or by peers as a know-it-all or "bootlicker", appears to have little impact on performance ratings. Taken collectively, the higher correlations for the directly job-related factors (for supervisors and peers) suggests that the overall job performance rating does reflect more attention paid to individuals' performance on the job than to their standing on other factors less directly relevant to performance.

One concern is that almost all of the ratings correlate substantially with overall job performance ratings. We should say that the differentiation the does occur (e.g., differences between factor sets in correlations with overall job performance, low correlations for friendship/knowing factor) is reassuring. Our point is, however, that halo appears to be significant in the ratings. It is true, of course, that halo in the ratings cannot be distinguished from the actual correlations between these underlying performance and personal characteristics constructs. Cooper (1981) makes this point very well in the context of expecting positive correlations between performance constructs because of "natural selection" factors in personnel selection programs and in employee turnover. Still, we believe that many of the correlations between these rating factors and overall job performance are higher than would be the case if we could measure the underlying constructs in a "true score" sense.

Also, the correlations between job knowledge and hands-on test scores and job performance ratings are low. We did not expect high correlations because overall job performance reflects much more than knowing how to do the job and having proficiency in accomplishing specific tasks. The knowledge/proficiency criteria appear to represent the "can-do" portion of job performance. The "will-do" part of job performance is not necessarily tapped by these measures (Guion, 1983). Nonetheless, higher correlations than those depicted in Table 4 obviously were expected.

Limitations of the Research Approach

We would like to interpret the correlations presented in Tables 1-4 as shedding light on the factors that "cause" the ratings. Clearly, with a correlational design this is difficult. It may be, for example, that once peers or supervisors perceive that a person is doing a good job, they begin

to see them positively on other factors, rather than the other way around. Other interpretations are possible as well.

There are other limitations to this kind of research. Comparisons between correlations of overall job performance and different other rating factors seem legitimate if these comparisons are made within rating source (e.g., peer or supervisor) and job. Halo is in a sense partialed out with these comparisons, because it should apply essentially equally to all of these relationships. However, across rating source or across job comparisons between these correlations are more difficult to interpret because of likely different magnitudes of halo with different source/job rating data. This would certainly restrict the investigation. Perhaps, all that can be done is to make comparisons <u>within</u> source and job, and then evaluate the stability of the findings across the different data sets.

<u>Summary</u>

Of the rating factors considered in this research, the highest correlations were obtained between directly job related factors and overall job performance ratings. This finding was obtained for peer assessments and supervisor ratings. Relationships of overall performance ratings with "good soldier" factors and interpersonal/relationship factors were of relatively lower magnitude. This pattern of results was interpreted as suggesting that the overall performance rating does reflect <u>more</u> attention paid to ratees' performance on the job than to factors less directly relevant to performance.

Again, the research reported here is exploratory. Future research in the Project A program will provide larger Ns to allow more stable estimates of the relationships presented in this paper. The project plan also calls for longitudinal data collection and analysis. This is an opportunity to obtain more definitive information on links between rating factors and job performance ratings.

References

Banks, C. G. (1979, August). Analyzing the rating process: A content analysis approach. Paper presented at the meeting of the American Psychological Association, New York, New York.

Bernardin, H. J. & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory (pp. 127-165). New Jersey: Lawrence Earlbaum Associates.

Borman, W. C. (1979a) Individual differences correlates of accuracy in evaluating others' performance effectiveness. Applied Psychological Measurement, 3, 103-115.

Borman, W. C. (1979b). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Cooper, W. H. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218-244.

Dansereau, F., Graen, G., & Haga, W. J. (1975). A vertical dyad linkage approach to leadership within formal organizations: A longitudinal investigation of the role making process. Organizational Behavior and Human Performance, 13, 46-78.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Graen, G., & Cashman, J. (1975). A role-making model of leadership in formal organizations: A developmental approach. In G. Hunt & L. L. Larson (Eds.), Leadership frontiers (pp. 143-165). Kent, OH: Kent State University Press.

Guion, R. M. (1983). Comments on Hunter. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory (pp. 267-276). New Jersey: Lawrence Earlbaum Associates.

Hogan, J., Hogan, R., & Busch, C. (1984). How to measure service orientation. Journal of Applied Psychology, 69, 167-173.

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

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory (pp. 257-266). New Jersey: Lawrence Earlbaum Associates.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In L. Cummings & B. Staw (Eds.), Research in organizational behavior, Vol. 3. Greenwich, CN: JAI Press.

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

Lewinsohn, P. M., Mischel, W., Chaplin, W., & Barton, R. (1980). Social competence and depression: The role of illusory self perceptions. Journal of Abnormal Psychology, 89, 203-212.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. Journal of Applied Psychology, 66, 451-457.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose. Journal of Applied Psychology, 69, 147-156.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-562.

Smith, P. C. & Kendall, J. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Zammuto, R., London, M., & Rowland, K. (1982). Organization and Rater Differences in Performance Appraisals. Personnel Psychology, 35, 643-658.

Relationships Between Scales on an Army Work Environment Questionnaire and Measures of Performance

Darlene M. Olson
Army Research Institute

Walter C. Borman
Loriann Roberson
Sharon R. Rose
Personnel Decisions Research Institute

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide information and procedures required to meet military manpower challenges of the future by enabling the Army to enlist, allocate, and retain the most qualified soldiers. This research is funded primarily by Army Project Number 20263731A791 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

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.

Presented at the Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

RELATIONSHIPS BETWEEN SCALES ON AN ARMY WORK ENVIRONMENT QUESTIONNAIRE AND MEASURES OF PERFORMANCE

Darlene M. Olson
U.S. Army Research Institute

Walter C. Borman
Personnel Decisions Research Institute

Loriann Roberson
Personnel Decisions Research Institute

Sharon R. Rose
Personnel Decisions Research Institute

This paper discusses the development of an Army Work Environment Questionnaire (AWEQ) and some preliminary results from administering it to 102 first-term Army enlisted personnel. The major purposes of this research were: 1) to identify environmental and situational influences that impact on job performance through application of a critical incident methodology, 2) to develop questionnaire items which assess these positive and negative environmental factors encountered by soldiers during their first-tour of duty, and 3) to examine preliminary relationships between these environmental factors and Army-wide ratings (i.e., supervisory and peer) of overall soldier effectiveness.

Current Army selection and classification measures [e.g., Armed Services Vocational Aptitude Battery (ASVAB)] have only been linked to performance on measures of training success. Congress (House Committee on Armed Services, 1982) and the Department of Defense have mandated that the services must pursue "a long-range systematic program of validating ASVAB and enlistment standards against performance on the job" (Robert B. Pirie, Assistant Secretary of Defense, MRA&L, September, 1980). In response to this mandate the Army initiated Project A, a comprehensive research program directed at improving the selection, classification, and utilization of Army enlisted personnel. This research effort is needed to help the Army deal with declining manpower availability and the complexity of modern military operations and equipment.

Paper presented at the 92nd Annual Convention of the American Psychological Association in Toronto, Canada, August 1984. All statements expressed in this paper are those of the authors and do not represent the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Conceptualization of the Relationship Between Performance and Environmental Influences

Job performance is a product of individual attributes, abilities, and skills which are measurable at the time a soldier first enters the Army, of events and situational factors which impact on the individual after job entry, and of a soldier's motivation to perform (Wetrogan, Olson, & Sperling, 1983). Hence, in order to adequately describe the linkages/interrelationships among human attributes, enlistment standards/selection criteria, and job performance, research should investigate all potential influences on performance, not exclusively job entry predictors.

In general, previous research has concentrated on explaining work performance in terms of human abilities (e.g., Dunnette, 1976) and motivation (e.g., Campbell & Pritchard, 1976). Although these approaches have explained some of the variability in performance across workers and job settings, other variables may enhance this prediction, or even better describe performance. One class of variables which could have an effect on performance, but has only recently received systematic investigation is the environment or situation. In a broad sense, the environment functions as the context in which performance occurs (Magnusson, 1981). Specifically, situational or environmental factors have been defined as a set of circumstances that are likely to influence the behavior of at least some individuals and are likely to reoccur repeatedly in essentially the same form (Frederiksen, Jensen, & Beaton, 1977).

This conceptualization of the domain of situational/environmental factors implies that work settings can be effectively described and that characteristics of these environments can be empirically identified through examination of their impact on workers at homogeneous (comparable) levels of an organization.

The environment/situation is known to influence behavior (e.g., performance) in two ways. First, it can influence behavior through constraint (Naylor, Pritchard, & Ilgen, 1980; Peters & O'Connor, 1983). The environment can interfere with or set limits on the range of behaviors that are displayed, which can have a potential effect on task performance and the relationship between ability and performance.

Second, the environment can influence behavior through affect (Naylor et al., 1980). The environment signals the kinds of available reinforcers, which in turn arouses motives, affect, and expectations for certain consequences/outcomes. As a result of these contingencies, behaviors are established and their direction, magnitude, and duration are modified.

Although the environment provides the context and opportunity for behavior and sets the limits for behavior, individuals are not simply passive recipients of environmental stimulation. Rather, individuals are active (i.e., in terms of the cognitive processing of information) and goal-directed participants in a continuously ongoing reciprocal person by situation interaction process (Bandura, 1978; Magnusson, 1981). Hence, in order to effectively describe performance in work environments, it is necessary to identify and measure the relative contributions of situational variables versus individual differences in accounting for variance in performance.

Theoretical and Empirical Research on Environmental Influences on Performance

Recently, Peters and O'Connor (1980) initiated a program of research to conceptualize and study the effects of constraints, which are one type of situational influence, on performance. Specifically, environmental/situational constraints have been defined as aspects of the immediate work situation that act, in some fashion, to interfere with the use of individual abilities and motivation in performing various jobs (Peters & O'Connor, 1980).

A theoretical model which describes the impact of situational constraints on performance and affective reactions of workers to their jobs has been developed by Peters and O'Connor (1980) and their colleagues. When work conditions are highly constraining, it is hypothesized that there will be a corresponding decrement in performance. Further, these researchers proposed that the presence of constraints will have a differential impact on individual performance based on the adequacy of task-relevant abilities and level of motivation. Specifically, it is assumed that constraints will have the most severe impact on the performance of highly capable and well motivated workers, and that these individuals will experience more frustration and dissatisfaction with their jobs than their counterparts with lower levels of ability and motivation.

On the basis of preliminary theoretical and empirical research, Peters, O'Connor, and Eulberg (1984) proposed a domain of situational constraints which is applicable across work environments and can be classified according to the following 11 general factors:

1. Job-Related Information
2. Tools and Equipment
3. Materials, Supplies, and Parts
4. Budgetary Support
5. Required Services and Help from Others
6. Task Preparation
7. Time Availability
8. Work Environment

9. Scheduling (e.g., coordination of work activities)
10. Transportation
11. Job-Relevant Authority

This proposed taxonomy does not necessarily represent an exhaustive list of situational influences (e.g., constraints), nor are all the dimensions mutually exclusive. In particular, since these situational variables are defined as "constraints" and emphaize deficiencies within work environments, potentially positive/facilitating attributes of these same dimensions may not receive adequate research investigation.

On the basis of content, this proposed taxonomy represents a broad combination of environmental influences, which have been traditionally investigated under the guise of job/task characteristics (Hackman & Oldham, 1974; Sims, Szilagy, & Keller, 1976), organizational variables (Payne & Pugh, 1976; Porter, Lawler, & Hackman, 1975), and climate factors (James & Jones, 1974 ; Schneider, 1975; Schneider & Reichers, 1983). Research which has examined the relationships between job and organizational variables and outcome criteria such as performance, motivation, and job satisfaction has been heavily criticized. This has occurred because of serious limitations in theoretical models, ambiguous definitions of relevant constructs, reliance on inadequate measures of the variables (e.g., exclusive use of questionnaires), and difficulties in analyses/ interpretation of perceptual data (Roberts & Glick, 1981). Consequently, any research conducted on the role of situational constraint factors should be cognizant of these existing problems and endeavor to improve the conceptual and methodological adequacy of research in this area.

Although this conceptual model of situational constraints has received support in laboratory studies, the results have not been as consistent or encouraging in applied settings. Data from analog laboratory research (Peters, O'Connor, & Rudolf, 1980; Peters, Chassie, Lindholm, O'Connor, & Rudolf, 1981; Peters, Fisher, & O'Connor, 1982; O'Connor, Peters, & Segovis, 1980) have demonstrated the negative impact of situational constraints on performance and affective reactions to the job (e.g., frustration and dissatisfaction). For example, in the Peters et al. (1980) investigation, four of the eight situational constraint factors identified by Peters & O'Connor (1980) (i.e., Job-Related Information, Tools and Equipment, Materials and Supplies, and Task Preparation) were manipulated to create either facilitating or inhibiting conditions. Findings showed that significantly lower performance and higher levels of frustration and dissatisfaction were associated with the presence of inhibiting conditions during experimental task performance.

Further, in the Peters et al. (1982) laboratory study, the contribution of the individual (i.e. ability and experience) versus the situation (i.e., constraints) to variance in

performance was examined. As hypothesized, findings indicated that individual differences in ability and experience predicted performance better when the variance in performance was not strongly related to situational constraints. In addition, O'Connor et al. (1980) in a reanalysis of earlier laboratory data found that both frustration and performance could be predicted better by differential abilities in low constraint rather than high constraint conditions.

Hence, the results from analog experiments suggest that the presence of inhibiting performance constraints is related to lower experimental task performance and can generate negative affective reactions to constraining task conditions. Also, these laboratory findings tentatively suggest that situational constraints may moderate known predictor (e.g., ability) and criterion (e.g., performance) relationships.

Several correlational field studies have been conducted, which examined the effects of environmental variables on various work outcomes. In research which used measures of satisfaction and frustration as outcome criteria, O'Connor, Peters, Rudolf and Pooyan (1982) found that situational constraints were significantly associated with negative affective responses to the job (e.g., frustration and job dissatisfaction). These findings were consistently observed across samples of employees from different jobs and occupational levels in private sector organizations. These main effects of situational constraints on affective reactions were replicated in a bank environment, where employees who depicted their jobs as high in constraining conditions were less satisfied and more frustrated with their job environment (Pooyan, O'Connor, Peters, Quick, Jones, & Kulisch, 1982). However, correlations were near zero between environmental constraints and job performance.

Although the previous field investigations were conducted in civilian work environments and did not find significant correlations between constraint conditions (factors) and performance criteria, other applied research has examined negative environmental influences on performance in military settings (e.g., Kane, 1981; White, Atwater, & Mohr, 1981).

A recent comprehensive field study sponsored by the Air Force Human Resources Laboratory (AFHRL) measured situational constraint dimensions for a sample (n=1352) of enlisted personnel in multiple Air Force Job Specialties (AFS) (Watson, O'Connor, Eulberg, & Peters, 1983). In this research 14 environmental constraint dimensions (e.g., Job-Related Information, Time Availability, Tools and Equipment, Communication, and Authority to accomplish work goals) were identified through a critical incident approach (Flanagan, 1954). A 57-item multiple-choice questionnaire was constructed to assess these constraint dimensions. The environmental measure was validated not only against performance criteria, but also in relation to other

outcome variables such as satisfaction, locus of control, supervisory culpability, and reenlistment intentions for the entire Air Force sample.

Findings demonstrated that total scores on the environmental constraint questionnaire correlated significantly ($p < .001$) with measures of frustration (.44), general satisfaction (-.28), supervisor satisfaction (-.43), pay satisfaction (-.28), locus of control (.14), and supervisory culpability (.37). Reenlistment intentions (-.07) were also significantly ($p < .05$) associated with total scale score on the environmental questionnaire. Further, significant correlations which were theoretically appropriate were obtained between scores on the above outcome measures and scores on the 14 separate constraint dimensions. No correlations between Air Force-wide performance measures and scale/total scores on the environmental constraint questionnaire were reported.

In order to examine the generalizability of these findings across Air Force jobs, additional data were collected with this environmental constraint questionnaire for several AFS (e.g., Fire Protection Specialist, Aircraft Systems Mechanic, and Security Specialist). The sample size ranged from 59 to 100 in the various AFS. Besides the previous measures of affective reactions to the job, performance measures (e.g., scores on specific and general Air Force occupational performance scales) were examined as outcome criteria. Contrary to the Peters and O'Connor (1980) conceptual model of situational influences, constraints tended not to significantly influence performance outcomes, and did not interact with ability or motivation in the prediction of performance. However, results did corroborate previous findings that constraints did decrease satisfaction, and increase frustration and thoughts of leaving the Air Force.

Although the previously described research has shown that these environmental constraint factors impact on laboratory task performance and can result in negative affective reactions to the job in applied settings, research data have only begun to accumulate for situational variables and their relationship to performance criteria across multiple job environments.

Further studies should be conducted which 1) attempt to develop taxonomies and to measure situational variables that operate in specific occupational settings (e.g., the Army environment); 2) examine the assumptions of the Peters and O'Connor (1980) model of situational constraints for other work settings; and 3) assess whether positive and negative environmental factors act as moderators of the relationships between task-relevant abilities, motivation, and affective reactions to the job and performance. The present investigation represents an important first step towards addressing each of these issues. Further, it may provide a basis for addressing such other

important research needs as: 1) examining the homogeneity of environmental perceptions within different organizational units (e.g., squad, platoons, companies); 2) exploring the possibility of correcting scores on performance measures for situational influences; and 3) developing research interventions designed to reduce or alleviate known constraints in work environments.

METHOD

The research described in this paper was conducted in two stages. The first stage involved identification of environmental/situational influences that impact on Army-wide performance. The second stage focused on the development of an Army Work Environment Questionnaire (AWEQ) to measure these environmental influences, and subsequent exploration of relationships between AWEQ scale scores and performance criteria (i.e., overall supervisory and peer ratings of soldier effectiveness).

Stage I: Identification of Environmental Influences on Army Performance

A taxonony of first-tour environmental influences on Army performance was derived through application of a critical incident methodology. An open-ended narrative questionnaire was used to generate behavioral examples in which environmental factors were reported as responsible for either effective or ineffective soldier performance. This critical incident approach to the identification of environmental factors is consistent with and parallels the work of other researchers (e.g., Peters, O'Connor, & Eulberg, 1984; Schneider, 1978).

Specifically, a series of six workshops was held at Forts Benning, Riley, and Carson over a nine month time period. A combined sample of 67 Commissioned Officers (e.g. Majors and Captains) and NCO, who were incumbents from a wide array of Army military occupational specialties (MOS), participated in the development of the taxonomy. During this research phase, these Army experts provided examples of environmental/situational factors that influenced performance positively and examples that impacted negatively on performance.

In order to generate examples of situational influences on performance, research participants were instructed to focus on incidents involving individual soldiers where environmental factors beyond the control of the soldier made a significant difference in his or her performance, either inhibiting or facilitating that performance. Although the creation of this response set should help control the influence of individual difference factors, generally the critical incident method may be vulnerable to other errors associated with social desirability or poor selective memory.

The 282 critical incidents collected from these various workshops were independently content-analyzed by a group of six judges, psychologists from the Army Research Institute and Personnel Decisions Research Institute. Each judge independently developed a category system and sorted the critical incidents on the basis of perceived similarity of content into these dimensions. After the critical incidents were categorized, judges discussed and reconciled any differences in the dimensions. Each environmental dimension was then defined jointly by the group of judges according to the critical incidents which were representative of the specific dimension.

While defining the environmental dimensions, the judges tried to maintain a close correspondence between the actual content of the critical incidents and resulting definitions. We also expanded the taxonomic work of Peters and O'Connor (1980) by identifying facilitating as well as constraining aspects of environmental variables. Table 1 presents a taxonomy of the 14 environmental factors which resulted from the content-analysis of the critical incidents. These dimensions are bipolar and tend to be similar to others identified in the civilian and military literature (Sulberg et al., 1984).

In terms of classification, the first nine environmental factors are "job-content related", whereas the remaining five dimensions are more indicative of climate variables. The definitions of these situational factors and the corresponding items on the AWEQ attempt to focus on the more observable attributes or descriptive qualities of the environment. This perspective on the development of the taxonomy and questionnaire items was taken in order to minimize the errors associated with purely perceptual data.

After the environmental dimensions were defined, a retranslation procedure was conducted where the entire group of critical incidents was sorted back into the 14 dimensions by two of the previous judges. The critical incidents were sorted into their respective dimensions about 72% of the time.

Stage II: Development of the Army Work Environment Questionnaire (AWEQ) and Preliminary Analysis of the Measure Against Performance

In order to measure the environmental dimensions identified in Stage I of the research, a 110-item multiple choice questionnaire was developed. Once the environmental factors were defined, the 14 dimensions were divided among four psychologists for the construction of questionnaire items. Each of the 14 environmental dimensions was treated as a scale on the Army Work Environment Questionnaire and items were written to cover the content of the separate dimensions. The number of items used to measure the environmental factors ranged from a low of 6 for such factors as Physical Working Conditions and Job-Relevant Authority to a high of 11 items for the Training dimension.

The items on the AWEQ 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 the questionnaire items occurs on their present job. For example, items consisted of statements such as "In your job, changes in equipment are introduced with little or no explanation" (Changes in Job Procedures or Equipment), or "If you needed help, you could depend on your co-workers to help you perform your required job tasks," (Job Related Support/Guidance). An effort was made to balance the number of positively and negatively worded items.

The Army Work Environment Questionnaire was administered on a pilot basis to 102 first-term Army enlisted personnel at Ft. Polk, Louisiana. The total sample contained 42 soldiers from the 95B MOS (Military Police) and 60 soldiers from the 71L MOS (Administrative Specialist).

In order to conduct preliminary construct validation of the Army Work Environment Questionnaire against performance indices, supervisory and peer ratings of overall Army-wide soldier effectiveness were obtained concurrently for this sample. The performance criteria used in this research were supervisory and peer ratings of overall soldier effectiveness. This behaviorally anchored rating of overall soldier effectiveness was made separately by supervisors and peers who had knowledge of individual soldier performance in 11 categories (e.g., Technical Knowledge/Skill) of Army-wide performance. Specifically, the overall soldier effectiveness criterion for both supervisors and peers was most highly correlated with the following separate dimensions of Army-wide performance: Technical Knowledge (r = .76 supervisor, r = .77 peers), Initiative/Effort (r = .71 supervisor, r = .72 peers), and Leading and Supporting (r = .77 supervisory, r = .71 peers). Consequently, this composite rating of effectiveness is conceptually and empirically defined with respect to those relationships. The overall soldier effectiveness rating was made on a 7-point scale (i.e., ranged from 1 or 2= below standard: soldier performs poorly in important effectiveness areas; does not meet standards nor expectations for adequate soldier performance to 6 or 7 = soldier performs excellently in all or almost all effectiveness areas; exceeds standards and expectations for soldier performance).

RESULTS AND DISCUSSION

Analyses of results from the development and preliminary construct validation of the Army Work Environment Questionnaire focus on: 1) an examination of the discrimination indices (e.g., reliability) of the 14 scales on the AWEQ, 2) a discussion of the relationships between scale scores (dimensions) on the AWEQ and performance criteria, and 3) an assessment of the homogeneity of environmental perceptions across specified units in the 95B MOS (unit-level analysis).

Item-Scale Correlations for the AWEQ

To explore the homogeneity of each of the 14 environmental scales, correlations were computed between scores on individual items and scale scores on the AWEQ. When comparisons are made between the item content of the AWEQ and item-scale correlations, findings show that 87% of the items correlate in the predicted direction with their assigned environmental dimension. Also, items generally correlate highest with their assigned scale. For example, all the items developed for the scales of Resources, Tools, and Equipment (Scale 1); Physical Working Conditions (Scale 4); Perceived Job Importance (Scale 7); Rewards/ Recognition/Positive Feedback (Scale 10); and Discipline (Scale 11) met this condition.

Although item assignments for the previously mentioned five scales were accurate based on obtained correlations, the remaining nine environmental scales contained some items that were more highly associated with other dimensions in the AWEQ. The number of items misclassified on these scales ranged from one to three. Although item-scale correlations are based on a limited n (n=97 to 101 responses per item), this finding suggests that some items may need revision or could be more appropriately assigned to other scales. From future administrations of the AWEQ with comparable Army samples, more internal reliability information will accumulate which will determine how items will be eventually revised.

Intercorrelations of the Scales from the AWEQ

Table 2 presents the intercorrelations between the 14 environmental scales on the AWEQ. Findings suggest that the climate-oriented scales (Scales 10-14) tend to be more highly intercorrelated than those scales which are more representative of job dimensions (Scales 1-9). For example, Scale 10, which involves the organizational reward system, is strongly associated with Recognition and Personal Support (Scale 12, r = .66), Job-Related Support/Guidance (Scale 13, r = .76), and Leader/Peer Role Models for Behavior (Scale 14, r =.61). Conceptually, one would predict a certain degree of correspondence between these scales.

Although job-oriented scales appear to be more distinct, the Job-Relevant Authority factor (Scale 5) shows a fairly strong association with the Job-Relevant Information (Scale 6, r = .59) and Changes in Job Procedures/Equipment (Scale 9, r = .58) scales. The interrelationships among these scales may occur because perceptions of the Job-Relevant Authority dimension are dependent on how supervisors (e.g., chain-of-command) dispense information to workers and display support for work goals. In addition, Scale 5 (Job Relevant Authority) is the job-oriented scale which has the highest correlations with the climate scales (e.g., r = .59 with Scale 10 and r = .57 with Scale 13).

The Workload/Time dimension (Scale 2) appears to be the most homogeneous of the environmental scales, because it has relatively low correlations with other job-oriented scales (except for an r = .56 with Physical Working Conditions), as well as consistently low correlations with the climate scales.

Overall, the item-scale correlations and the scale intercorrelations suggest that revisions in the AWEQ are necessary to increase the homogeneity of the environmental scales. Judicious item revision, possibly a restructuring of the 14 construct system, and future factor analysis work with larger samples should enhance the psychometric properties and construct validity of this measure.

Relationship Between the Scale Scores on the AWEQ and Performance Criteria

These environmental scale scores (predictors) were correlated with both supervisory and peer ratings of overall soldier effectiveness (performance criteria). Intraclass correlation coefficients obtained for the performance criteria indicate that both peer (r = .60) and supervisory (r =.68) ratings had adequate reliability.

Table 3 presents the correlations between the 14 environmental scales and the measures of overall soldier effectiveness. Findings demonstrate that there are significant ($p < .05$) relationships between both supervisory and peer overall soldier effectiveness ratings and six of the 14 environmental scales. Specifically, supervisory ratings are significantly correlated with the more objective job scales of Training (r = .20), Job-Relevant Authority (r = .24), and Work Assignment (r = .23), as well as with the climate-oriented scales of Rewards/Recognition/ Positive Feedback (r = .23), Discipline (r = .20), and Job-Related Support (r = .27).

In constrast, the peer overall effectiveness ratings are significantly correlated ($p < .05$) with scores on the Physical Working Conditions (r = .22), Job-Relevant Information (r = .26), and Changes in Job Procedures (r = .36) scales (job-oriented) from the AWEQ. Also, the peer performance ratings were significantly related ($p < .05$) to the climate dimensions of Rewards/Recognition/ Positive Feedback (r = .20), Job Related Support (r = .22), and Leader/Peer Role Models (r = .24).

Although scores from six environmental scales on the AWEQ significantly correlated with each performance measure, convergence across criteria was only observed for two climate scales (i.e., Rewards/Recognition/Positive Feedback and Job-Related Support). It is interesting that these two climate-oriented scales had significant associations with both supervisory and peer performance measures, because these dimensions tend to be less objective and more

perceptual/idiosyncratic than the job dimensions. In addition, these correlations suggest some divergence with respect to the influence of environmental factors on performance measures. For example, while scale scores on Job-Relevant Information, Changes in Job Procedures/Equipment, and Leader/Peer Role Models were significantly correlated with peer ratings of overall soldier effectiveness, these same dimensions were not related to supervisory effectiveness ratings. Hence, these data suggest that environmental influences may relate differently to different performance criteria.

Unit-level Analysis of Responses to the AWEQ

For years, environmental/climate research in organizations has suffered from both a lack of conceptual clarity with respect to relevant variables, and methodological problems related to the appropriate unit (e.g., individual or organizational unit) for analyses of perceptual questionnaire data. In this research, the issue of the conceptual relevance of environmental variables was addressed through the use of a critical incident method, which had respondents describe what actually happened in their work setting, rather than how they felt about the environment. This descriptive as opposed to an evaluative approach was maintained in the actual construction of the AWEQ through careful attention to the written instructional set and to the item response format (i.e., frequency rating scale).

The data aggregation problem, which requires meaningful integration and analysis across pooled individual descriptions of a common work environment, was in a sense "dodged" in this preliminary work. Analyses were conducted at the individual soldier level. However, it was possible to aggregate environmental scale data for three company-size units of military police (total N = 42) to begin to evaluate within unit agreement in questionnaire responses. Administrative Specialists (71L MOS) worked primarily alone, and therefore this kind of analysis was not conducted for them.

Intraclass correlations, indexing within-unit variability in scale scores compared to across-unit variability in the scores were computed for each of the 14 environmental factors. Several positive intraclasses were found for such dimensions as Training (r = .79), Physical Working Conditions (r = .52), Changes in Job Procedures or Equipment (r = .62), Rewards/Recognition/Positive Feedback (r = .59), and Leader/Peer Role Models (r = .51). These positive intraclass correlations indicate that there is less variability within the unit than across units.

These data indicate that with some of the scales, at least, respondents tended to agree in their descriptions of the environment. This is reassuring, because it suggests (although definitely does not necessarily prove) that soldiers are focusing

on factors that exist in substance rather than those that exist only in terms of idiosyncratic perceptions of environmental phenomena. That most everyone in the unit describes the stability of his/her job conditions pretty much the same way, for example, lends credence to the possibility that a job stability construct is real in organizational practices and can be reliably described by unit members.

CONCLUSIONS

The present research has identified through application of a critical incident methodology, 14 environmental factors, which appear important within the Army work environment. This Army taxonomy, which contains both job and climate-related variables, corresponds reasonably closely with other civilian and military taxonomies (e.g., Eulberg, O'Connor, Peters, & Watson, 1984). A 110-item Army Work Environment Questionnaire (AWEQ) was constructed to measure these environmental factors.

Although previous empirical research (e.g., Peters, O'Connor, & Eulberg, 1984) has found significant relationships between environmental variables and performance only in experimental laboratory settings, this investigation found significant relationships between six scales on the AWEQ (e.g., Rewards/Recognition/Positive Feedback) and performance (i.e., overall ratings of soldier effectiveness).

Despite the correlational nature of these findings, it is encouraging that some significant relationships between environmental predictors and performance measures were obtained in an applied military setting. Prior to this research environmental influences were only associated significantly with performance in laboratory settings. Hence, these results provide tentative support for the theoretical work of Peters and O'Connor (1980), which contend that environmental factors can directly influence performance on the job.

Future research should further explore the contributions to job performance of the "person" versus "situation." Correlations between Army individual difference measures (e.g., ASVAB) and later job performance will describe contributions of person factors to variance in performance. Relationships between environmental factors and performance will indicate the strength of situational factors in job performance.

Importantly, the Project A research program provides an opportunity to refine the AWEQ. This is needed to improve the psychometric properties (e.g., internal scale reliabilities) and generally to enhance the measurement of these environmental factors. The Project A plan also calls for future longitudinal design applications which will be extremely useful for evaluating contributions to performance (and attrition) of the person, the Army environment, and their interaction.

REFERENCES

Bandura, A. (1978). The self system in reciprocal determinism. American Psychologist, 33, 344-358.

Campbell, J. P., & Pritchard, R. D. (1976). Motivation theory in industrial and organizational psychology. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 63-130). Chicago, IL: Rand McNally.

Dunnette, M. D. (1976). Aptitude, abilities, and skills. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 473-520). Chicago, IL.: Rand McNally.

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

Frederiksen, N., Jensen, O., & Beaton, A. (1972). Prediction of organizational behavior. New York: Pergamon Press.

Hackman, J.R., & Oldham, G.R. (1975). Development of the job diagnostic survey. Journal of Applied Psychology, 60, 159-170.

James, L.R., & Jones, A.P. (1974). Organizational climate: A review of theory and research. Psychological Bulletin, 81, 1096-1112.

Kane, W.D. (1981). Task accomplishment in an Air Force maintenance environment. (AD-A101 108/9). Bolling AFB, DC: Air Force Office of Scientific Research (NL) (80-0146).

Magnusson, J. (1981). Toward a psychology of situations: An interactional perspective. Hillsdale, NJ: Lawrence Erlbaum Associates.

Naylor, J.D., Pritchard, R.D., & Ilgen, D.R. (1980). A theory of behavior in organizations. New York: Academic Press.

O'Connor, E.J., Peters, L.H., & Segovis, J. (1980, August). Situational constraints, task-relevant abilities, and experienced frustration. Paper presented at the Annual Meeting of the Academy of Management, Detroit, MI.

O'Connor, E.J., Peters, L.H., Rudolf, C.J., & Pooyan, A. (1982). Situational constraints and employee affective reactions: A partial field replication. Group and Organization Studies, 7, 418-428.

Payne, R.L., & Pugh, S.S. (1976). Organization structure and organization climate. In M.D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago, IL.: Rand McNally.

Peters, L.H., & O'Connor, E.J. (1980). Situational constraints and work outcomes: The influences of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Peters, L.H., O'Connor, E.J., & Rudolf, C.J. (1980). The behavioral and affecting consequences of performance-relevant situational variables. Organizational Behavior and Human Performance, 25, 79-96.

Peters, L.H., Fisher, C.D., & O'Connor, E.J. (1982). The moderating effect of situational control of performance variance on the relationship between individual differences and performance. Personnel Psychology, 35, 609-621.

Peters, L.H., O'Connor, E.J., & Eulberg, J.R. (1984). Situational constraints: Sources, consequences, and future considerations. Research in Personnel and Human Resources Management, 3, forthcoming.

Peters, L.H., Chassie, M.B., Lindholm, H.R., O'Connor, E.J., & Rudolf, C.J. (1981, March). The joint influence of situational constraints and goal setting on performance and affective outcomes. Paper presented at the Southwest Academy of Management Meeting, Houston, TX.

Pooyan, A., O'Connor, E.J., Peters, L.H., Quick, J.C., Jones, N.D., & Kulisch, A. (1982). Supervisory/Subordinate differences in perceptions of performance constraints: Barriers are in the eye of the beholder. Proceedings of the Annual Convention of the Southwest Academy of Management, 170-174.

Porter, L.W., Lawler, E.E., III, & Hackman, J.R. (1975). Behavior in organizations. New York: McGraw-Hill.

Roberts, K.H. & Glick, W. (1981). The job characteristic approach to task design: A critical review. Journal of Applied Psychology, 66, 193-217.

Schneider, B. (1975). Organizational climate: An essay. Personnel Psychology, 28, 447-479.

Schneider, B., & Reichers, A.E. (1983). On the etiology of climates. Personnel Psychology, 36, 19-39.

Sims, H.P., Szilagyi, A.D., & Keller, R. T. (1976). The measurement of job characteristics. Academy of Management Journal, 19, 195-212.

Watson, T.W., O'Connor, E.J., Eulberg, J.R., & Peters, L.H. (1983, October). Measurement and assessment of situational constraints in Air Force Work environments: A brief summary. Proceedings of the 25th Annual Conference of the Military Testing Association.

Wetrogan, L.I., Olson, D.M., & Sperling, H. M. (1983). A systemic model of work performance (Working Paper 83-6). Alexandria, VA: Army Research Institute.

White, M.A., Atwater, L.Y., & Mohr, D.A. (1981, August). A practical methodology for identifying impediments to productivity (NPRDC TR-81-18). San Diego, CA: Navy Personnel Research and Development Center.

Table 1

A Taxonomy of Army Work Environment Factors

1. RESOURCES/TOOLS/EQUIPMENT

* Tools, parts, equipment needed to do the job are not available at all, or not available in sufficient quantity.

* Equipment/tools are of inferior quality, faulty, inadequate for the job, break down frequently, and/or require excessive maintenance time.

versus

* Necessary tools, parts, equipment are always available or easily accessible; an adequate supply of necessary supplies is maintained.

* Tools/equipment are well conditioned and in running order; defective tools or parts are quickly replaced to avoid maintenance down time. Outmoded equipment/tools are replaced with newer models to keep pace with technological changes in the Army.

2. WORKLOAD/TIME

* Workload is too heavy- assigned additional details (e.g., training, inspection preparation) after duty hours; required to work longer shifts due to personnel shortages; good performers given others' tasks to complete in addition to own.

* Too little time given- given unreasonable time limit to complete a specific job, or the assigned workload consistently too great for time limit; no scheduled time for tasks that are low priority but essential (e.g., maintenance); frequent interruptions (e.g., special duties) conflict with task completion.

* Workload too light- too many personnel assigned to a job; unit tasked with too little work, SM must perform "busy work".

versus

* Workload commensurate with available time limits. It is usually possible to finish all assigned tasks within the scheduled time limit. Workload is distributed evenly across unit members.

* Assignments are carefully scheduled so that low priority items can be completed during slow periods. To the extent possible, training activities and special details are scheduled to coincide with slack time in the work schedule.

Table 1 (cont)

A Taxonomy of Army Work Environment Factors

---

3. TRAINING IN MOS SKILLS/OPPORTUNITY TO IMPORVE MOS SKILLS

* Did not receive adequate training in AIT/other schools, etc., or training content conflicts with what is expected on the job; does not receive additional on the job training to correct deficiency.

* Does not receive additional training to keep current in MOS.

* Does not have the opportunity to practice new skills acquired in training due to assignments to non-MOS specific details or assignments out of MOS.

versus

* Received adequate training in AIT/other schools; training content matches well with what's expected on the job.

* Receives on the job training and practice time to improve MOS skills and/or to keep up to speed on MOS skills that are infrequently used (e.g., combat skills).

4. PHYSICAL WORKING CONDITIONS

* Must perform work in unfavorable physical conditions that are not a typical requirement for the MOS. For example, extremely dirty or disorderly workshops and motor pools, office buildings where noise, temperature level, etc. are inadequately controlled.

versus

* In garrison, job sites are well maintained. Offices and workshops are orderly and clean. Efforts are made to keep noise, temperature levels, etc., within an acceptable range.

5. JOB RELEVANT AUTHORITY

* SMs assigned tasks to complete, but due to their rank or failure of supervisors to provide support, they do not have sufficient authority to get the job done, e.g., can not obtain cooperation from other personnel.

versus

* Where SMs' task accomplishment depends on eliciting cooperation from others, they are also delegated relevant authority and supported acccordingly so that they are able to get the job done.

Table 1 (cont)

A Taxonomy of Army Work Environment Factors

---

6. JOB RELEVANT INFORMATION

* SM does not receive information, either from the chain-of-command or immediate work group, that is needed to perform task efficiently, e.g., up-to-date technical documents, notice of regulation or procedural changes, sufficient notification of upcoming events and deadlines, etc.

versus

* SM is kept up-to-date on all information relevant to the job and provided the necessary technical manuals and other documents. SM is promptly notified of changes in procedures or regulations that affect own work.

7. PERCEIVED JOB IMPORTANCE

* SM believes his/her role in the Army, MOS or on a specific task is not important. For example, such SMs do not personally have responsibility for the outcome of their work, or so many personnel are working on the same job that they feel no ownership of outcome; SMs feel their MOS skills are not important because they are never or rarely called upon to use them.

versus

* SM assigned tasks involving some level of responsibility, or his/her job affords an opportunity to perform tasks of obvious significance, e.g., rescue missions.

8. WORK ASSIGNMENT (and underutilization of abilities)

* SMs are not performing at ability level or not using skills acquired in training because they have been assigned to a duty outside their MOS; SMs are assigned within their MOS but given little or no MOS-specific work (e.g., combat MOS or overcrowded MOS). Instead soldier spends most duty hours on post details such as clean-up.

versus

* SMs are assigned to MOS they were trained for and given assignments appropriate to ability and skill level. Where MOS skills are infrequently used (e.g., combat MOS), other opportunities are provided to maintain MOS specific proficiencies. If soldier is assigned outside own MOS, he/she is given the opportunity to keep current in this MOS and to prepare for the appropriate SQT.

Table 1 (cont)

A Taxonomy of Army Work Environment Factors

---

9. CHANGES IN JOB PROCEDURES OR EQUIPMENT

* Nature of MOS tasks change frequently due to changes in procedures, equipment or supervision. Little or no start up time is offered before new procedures go into effect. SM must learn new tasks immediately. Changes may be introduced with little or no explanation of the rationale involved.

versus

* Job tasks tend to be consistent over time. When new equipment or procedures are introduced, sufficient learning time is provided. Rationale behind changes that affect SM's work are explained.

10. REWARDS/RECOGNITION/POSITIVE FEEDBACK

* Good performance ignored, inconsistently or inequitably rewarded either due to Army-wide policies or leadership practices.

versus

* Good Performance consistently and fairly rewarded/recognized by chain of command (e.g. at command level, awards, soldier of month, local recognition; at supervisor level, praise, favorable assignments, promotion recommendation, passes, etc.).

11. PUNISHMENT

* Punishment practices are inconsistent and unfair, some personnel receive no punishment or milder form of discipline for offenses; entire unit is punished for behavior of a few soldiers.

* Discipline inappropriate for offense, overly harsh or severe.

versus

* Punishment is appropriate, targeted to specific individual and nature of offense; individual perceives discipline as a warning and is motivated to reform.

Table 1 (cont)

A Taxonomy of Army Work Environment Factors

---

12. RECOGNITION AND SUPPORT FOR PERSONAL WELFARE

Chain of command, immediate supervisor, work group or other Army personnel soldier comes in contact with:

* Show insensitivity to new soldiers having difficulty coping with Army-life, fail to recognize personal problems contributing to poor performance, fail to take action when problems identified by soldier himself or others (includes administrative errors contributing to severe personal hardship).

* Fail to support soldier in rehabilitative efforts (e.g., drugs/alcohol programs), "write-off" soldier as loser.

versus

* Express an interest in soldier's general welfare, are aware of changes in individual's performance/behavior, sensitive to potential difficulties, encourage communication.

* Recognize serious problems, refer to counseling, support efforts at rehabilitation.

13. JOB-RELATED SUPPORT/GUIDANCE

Chain-of-command, work group, immediate supervisor or other Army personnel soldier works with:

* Fail to recognize individual performance problems (e.g., inadequately trained new soldier, slow learner) and/or do not provide assistance/guidance to soldier with obvious performance weakness; label soldier based on initial performance; do not offer opportunities for good or poor performers to improve job skills.

versus

* Are aware of individual differences in performance, recognize soldier's weaknesses and strengths, offer additional assistance/guidance, provide personal attention and opportunities for improving job skills.

Table 1 (cont)

A Taxonomy of Army Work Environment Factors

14. LEADER OR PEER ROLE MODELS FOR JOB AND SOCIAL BEHAVIOR

* Soldier exposed to leaders or peers who encourage low standards for social behavior and job performance by not adhering to Army regulations, exhibiting a lack of knowledge about their MOS, avoiding participation in Army events, disparging Army life, accepting or promoting negative behavior such as AWOLs, drug/alcohol abuse, etc.

versus

* Soldier observes leaders and peers who adhere to and support Army regulations, are skilled and knowledgeable in their MOS, actively participate in Army events, express an interest in an Army career, avoid negative behaviors, etc.

NOTE. SM means soldierman

Table 2

Scale Intercorrelations for the AWEQ

ENVIRONMENTAL SCALES	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. RESOURCES		34	09	46	41	41	43	14	51	25	26	11	22	55
2. WORKLOAD			18	56	46	36	18	12	35	[illegible]	20	28	36	31
3. TRAINING				38	34	38	42	67	22	36	28	39	46	35
4. CONDITIONS					49	47	37	29	51	33	29	27	34	44
5. AUTHORITY						59	50	27	58	59	41	54	57	49
6. INFORMATION							48	30	54	50	47	54	53	57
7. IMPORTANCE								44	32	42	43	38	41	50
8. ASSIGNMENT									20	24	29	29	28	19
9. PROCEDURES										45	41	33	43	54
10. REWARDS											48	66	76	61
11. DISCIPLINE												46	56	53
12. PERSONAL SUPPORT													77	53
13. JOB SUPPORT														62
14. ROLE MODELS														

Table 3

Correlations Between Scales on the AWEQ and Performance

Environmental Scales[a]	Supervisor Ratings of Performance	Peer Ratings of Performance
1. Resources	.05	-.17
2. Workload	.02	-.02
3. Training	.20*	.18
4. Working Conditions	.18	.22*
5. Job Relevant Authority	.24*	.16
6. Job Relevant Information	.07	.26*
7. Perceived Job Importance	.01	.07
8. Work Assignment	.23*	.15
9. Changes in Job Procedures	.10	.36*
10. Rewards/Recognition	.23*	.20*
11. Discipline	.20*	.19
12. Recognition and Support	.19	.11
13. Job-Related Support	.27*	.22*
14. Leader/Peer Role Models	.13	.24*

* = Correlations which are significant at $p < .05$

a. To ease interpretations of correlations in the table, scoring of the scales has been accomplished so that high scores always mean a relatively favorable environment (e.g., high perceived job importance, few changes in job procedures, etc.)

Reports and papers dealing with methodological issues

Because reports that offer conceptual and methodological advances have immediate relevance for other personnel working in this field, special attention is given to documenting work of this nature.

(1) While hands-on tests of job performance are often the most valid measure of proficiency, they are also expensive to conduct; knowledge tests, on the other hand, cost less but also may not correlate well with performance. Exploring the bases for making such trade-off decisions, Osborn and Hoffman discuss the relationships between the two methods of measurement, and ways of estimating their relative costs. The goal is to select the mix of measures that will, per unit of cost, maximize the content validity of a test.

(2) The personal work constructs that Army officers use in judging work performance are described and analyzed in a paper by Borman. Both the similarities and the differences between officers are found to be substantial.

(3) Following on an iterative series of revisions and refinements as a result of conceptual development and field experiment, a description of the model of soldier effectiveness has been prepared. While the model concepts can be expected to continue to evolve, this report by Borman, Motowidlo, Rose, and Hanser describes 15 dimensions of effectiveness that are now postulated in the areas of organizational commitment, organizational socialization, and morale. (The text of this draft report follows; the appendices, A-D, are reproduced separately in ARI Research Note 85-14, in press.)

The Cost-Effectiveness of Hands-on and Knowledge Measures

William Osborn and R. Gene Hoffman
Human Resources Research Organization

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide information and procedures required to meet military manpower challenges of the future by enabling the Army to enlist, allocate, and retain the most qualified soldiers. This research is funded primarily by Army Project Number 20263731A791 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

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.

Presented at the Annual Convention of the American Psychological Association, Toronto, Ontario, Canada. Prepared for symposium, Performance Measurement: Methodological Issues.

THE COST-EFFECTIVENESS OF HANDS-ON AND KNOWLEDGE MEASURES[1]

William Osborn and R. Gene Hoffman
Human Resources Research Organization

Hands-on tests of task performance are generally conceded to be the most valid or relevant measures of job proficiency. This is not surprising. A well constructed and administered hands-on test calls for the actual job behavior or a facsimile of it under conditions highly similar to those of the job. Yet this kind of test has a major shortcoming: cost. The time, personnel and wear and tear on equipment necessary for administration are often seen as prohibitive. Because of this, paper-and-pencil tests of job knowledge are widely used as substitute or surrogate measures of proficiency.

Knowledge tests are economical because they are group administrable, machine scoreable, and do not require the paraphernalia or conditions of the job setting. Their relevance as a criterion measure of job proficiency is normally evaluated in terms of their correlation with corresponding hands-on measures. Reports of such correlations vary widely, ranging from near zero to as high as .8 (Vineberg & Taylor, 1972; Foley, 1974; Osborn & Ford, 1976). This variation, though seldom explained, most likely results from two factors: the type of task being tested and the quality of the knowledge test.

---

[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.

We would all agree, I think, that a knowledge test of skilled psychomotor performance would not correlate well with a direct or hands-on measure of that performance. On the other hand, a well-constructed knowledge test of a cognitive task would be expected to correlate highly with a measure of actual performance. The quality of a knowledge test further moderates the correlation with a hands-on counterpart. High correlations between knowledge and hands-on measures have been shown (Osborn & Ford, 1976) for procedural tasks where the knowledge items are methodically tied through task analysis to steps in task performance.

Faced with a requirement to assess proficiency within some specified task domain, such as an Army job specialty, it is clear that more tasks can be tested in less time and at less cost with knowledge measures than with hands-on tests. But greater coverage of the task domain must be traded-off against lower validity. If the relationships between the two methods of measurement were established for different types of tasks, and the relative cost of methods were known, cost-effectiveness decisions could be made in selecting the mix of measures that per unit of cost would maximize the content validity of a job proficiency test.

The opportunity to explore this will present itself in the Army's current project (Project A) to validate new soldier selection standards. In this project tests of job proficiency are being developed for a sample of job specialties. Since cost-effective coverage of the task domain is a concern, it is important to use hands-on measures only where knowledge tests will not do. Thus, for many tasks, in each job specialty, both types of measures are being constructed for comparison in a field test. These comparisons, made for various types of tasks, will enable us to evaluate the effectiveness of knowledge measures as surrogates for hands-on tests.

Purpose

The general purpose of this paper is to explore cost-effectiveness of test methods as a way of enhancing content validity. Specifically we will first estimate the relative cost of hands-on and knowledge tests, then examine measures of effectiveness, and finally discuss how these in combination may effect the content validity of a proficiency test.

Cost of Test Methods

The cost of developing a knowledge test of a job task is about the same as that of a hands-on test, since the bulk of the work goes into task analysis which is necessary and similar for both. Developmental trials of a hands-on measure call for more resources than a knowledge measure, but the difference is small relative to subsequent administrative costs. The analysis of cost differences that follows concentrates therefore on resources associated with test administration.

We begin with some assumptions--assumptions, I should point out, that are based on our experience in developing proficiency tests, chiefly but not exclusively, for military technical jobs. These are as follows:

1. The time to administer a hands-on test ranges from five minutes to an hour per task with an average of 20 minutes.

2. One scorer per task plus one monitor per 10 scorers are required to administer hands-on test, and only one person can be tested at a time.

3. A knowledge test averaging about eight items per task can be taken in five minutes.

4. One monitor per 20 examinees is required for knowledge test administration, and all 20 can be tested at once.

5. Equipment and facilities for administering a complete job knowledge test are roughly equivalent to those for one hands-on test station.

From these assumptions one can compile the estimated time and resources necessary to administer each type of test for varying number of tasks to varying numbers of examinees. For example, to administer a hands-on test of five tasks to 20 examinees would take slightly over 73 personnel hours, which includes 33 1/3 examinee hours (5 tasks x 20 min per task x 20 examinees) and 40 staff hours (6 staff per set of five test stations x 1 2/3 hours per examinee x 4 sets of five examinees). Facility hours, which include equipment and space, total 33 1/3 (equivalent to total examinee hours) for the same five-task hands-on test of 20 examinees. A comparable knowledge test on the other hand, would require about 10 1/2 personnel hours (1/2 hour for each of 20 examinees plus one examiner) and a half a facility hour. Resource hours for additional tasks and examinees may be estimated in a similar manner. Sample values are shown in Table 1.

As you might expect the resource hours tend to increase proportionately to increases in number of tasks or examinees. More important to note, however, is the constant ratio of total resource hours for the two types of tests. Hands-on resource requirements are ten times those for a knowledge test regardless of the number of tasks or examinees. This ratio is based on the assumption that a facility hour costs the same as a personnel hour. This may not be a reasonable assumption; a personnel hour may well cost more than a facility hour. If we assume it is ten times more, it may be shown that the total hands-on resource hours are still over seven times those required for administration of knowledge tests. Let us compromise, since we are more interested in method than exact numbers, and set the cost ratio at eight to one.

Table 1

Resource Hours to Administer a Hands-on or Knowledge Test

Number of Tasks	Number of Examinees	Hands-On Test			Knowledge Test		
		Pers Hrs	Fcty Hrs	Total	Pers Hrs	Fcty Hrs	Total
5	20	73	33	106	10.5	.5	11
	80	292	132	424	42	.5	42.5
	320	1168	528	1696	168	1	169
10	20	140	67	207	21	1	22
	320	2240	1072	3312	336	2	338
20	20	280	134	414	42	2	44
	320	4480	2144	6624	672	4	676

That administrative costs of a hands-on test are eight times those of a knowledge test of the same job content is of some interest. But by itself that bit of information is not of much use. Any cost differential must be considered in light of the relative effectiveness of the test methods.

Effectiveness of Test Methods

Generally, a measure of effectiveness would be defined in terms of variance accounted for in the domain of job proficiency. Because of potential deficiency (e.g., motor elements omitted) and contamination (e.g., reading requirements imposed), a knowledge measure of a task presumably would account for less variance than its hands-on counterpart, arguing for use of the hands-on measure. Yet for the price of that hands-on measure it may be possible to tap more total variance with knowledge tests of several tasks. To evaluate that possibility we need to know, ideally, three things in addition to cost: one is the proportion of variance in the domain of job proficiency accounted for by each task; the second is the correlation between a hands-on measure and actual task proficiency; and the third is the correlation between a knowledge and a hands-on measure of a task. How one might go about estimating these variables and using them to select cost-effective measures of job proficiency is the problem.

Let me first suggest an approach for the ideal circumstance in which we have empirical data on the intercorrelation of hands-on and knowledge measures for a set of job tasks. The intercorrelation matrix could be factored using a maximum-likelihood method and forcing a solution with the tasks as factors (Table 2), the loading representing the correlation of each task-by-method measure with the task factor. A loading of one, or near one, may be assigned to each hands-on measure for its corresponding task factor to assure that it is the anchoring measure of task proficiency. If it is

Table 2

Factor Matrix Of Task-By-Method Measures With Tasks As Factors

Task-By-Method Measures	Factor (Task) 1	2	3	4 . . . N
H_1	.9	a_{H12}	a_{H13}	. . .
K_1	a_{K11}	a_{K12}		
H_2	a_{H21}	.9		
K_2	a_{K21}			
H_3	.		.7	
K_3	.			
H_4	.			.9
K_4				
.				
.				
.				
H_N				.9
K_N				

known from task analysis that a hands-on measure excludes a significant portion of task performance--for practical reasons, such as safety--the loading could be scaled down accordingly. Factor or task scores may then be computed and summed over tasks to provide for each person a derived measure of job proficiency. Observed hands-on and knowledge scores may then be correlated with the derived job proficiency score (Table 3). Squaring these correlations provides an estimate of total job variance accounted for by each task measure, what might be termed an effectiveness index. Dividing effectiveness by cost gives us an index of cost-effectiveness or the efficiency of a measure. The cost factor could be approximate, assigning from the earlier cost analysis a one to a knowledge measure and an eight to a hands-on measure, or it could be more precise, calculating the administrative cost in resource hours for the task measures individually.

A correlational approach can now be used to select iteratively the measures that comprise the most cost-efficient mix. Beginning with the measure with the largest efficiency index, a second is added, the one which in combination with the first gives the highest correlation with J, the derived measure of job proficiency. Measures are added one at a time in this manner, adding at each stage the one that in combination with those already selected gives the highest composite correlation with job proficiency. The process, which stops when the multiple correlation ceases to increase, would produce the set of measures with the lowest administrative cost and the highest composite correlation with job proficiency. Or the selection process could be stopped when some fixed level of cost is reached, resulting in the most valid set of measures within the testing budget.

Table 3

Correlation Matrix Of Derived Task
Scores And Observed Task-By-Method Measures

	Hands-On			Knowledge			Tasks			Job		
	H_1	H_2	H_N	K_1	K_2	K_N	T_1	T_2	T_N	$\Sigma T=J$	COST	EFFICIENCY
H_1										r^2_{H1J}	c_{H1}	r^2_{H1J}/c_{H1}
H_2										r^2_{H2J}	c_{H2}	r^2_{H2J}/c_{H2}
H_N										r^2_{HNJ}	c_{HN}	r^2_{HNJ}/c_{HN}
K_1										r^2_{K1J}	c_{K1}	r^2_{K1J}/c_{K1}
K_2										r^2_{K2J}	c_{K2}	r^2_{K2J}/c_{K2}
K_N										r^2_{KNJ}	c_{KN}	r^2_{KNJ}/c_{KN}
T_1												
T_2												
T_N												

This approach suggests how a cost-effectiveness analysis might be conducted if one had performance data for hands-on and knowledge measures of job tasks. Such data will be available in Project A for a few tasks in each of nine job specialties, so we will be able to explore, on a small scale anyway, this kind of analysis.

Rarely, however, are these data available. Test developers normally decide on test methods with little more to go on than judgment of the kinds of skills involved in task performance. To increase test efficiency in this circumstance, developers need some way of estimating the correlation among tasks as well as between methods of testing the tasks.

Wheaton (1977), attempted to approximate correlations among tasks in the domain of tank gunnery by tabulating for every pair of tasks the relative number of identical task elements. These similarity measures were cluster analyzed and the clusters used as a framework for sampling tasks for testing. In Project A we are trying a more direct approach in which several job experts sort tasks into groups on the basis of similarity of procedures required in task performance. Task clusters that emerge from these judgments may also be viewed as an approximation of the real thing--that is, task clusters derived from intercorrelations of actual task performance. The validity of the approximation remains to be seen. But in this case we will have an opportunity to evaluate it by comparing observed task intercorrelations from the field test with those derived from expert judgment.

The correlation between test methods for different types of tasks is also being examined in Project A. As mentioned earlier, we are developing both hands-on and knowledge tests for a range of tasks. This range is

defined by task analysts who rate the tasks in terms of required psychomotor skill, time limitations on performance, and the number of uncued steps in the task procedure. The hypothesis is that the higher the value of these three characteristics, the lower the correlation of a knowledge measure with hands-on performance. This too can be assessed in the field test.

The importance of these procedures is in their attempt to estimate relationships among tasks and test methods. For if developers had even weak estimates of these factors, proficiency test batteries could be made more efficient. Consider a set of data (Table 4) in which job tasks had been clustered from task similarity judgments and a weight, or index of cluster membership, determined for each. Assume as well the following: (1) an index of task variance-accounted-for by the hands-on measure, r^2_H, which would be one except where for practical reasons only part of the task can be tested; (2) an estimate of the correlation between a knowledge and hands-on measure of the task, r_K; and (3) cost estimates for the two methods of testing, C_H and C_K. From these a measure of efficiency may be calculated by task for each test method, with hands-on efficiency defined as the product of cluster weight and hands-on variance accounted for divided by hands-on cost, and knowledge efficiency as the product of hands-on and knowledge variance accounted for times cluster weight divided by cost of the knowledge measure. Then, using a procedure like that described earlier, a set of task-by-method measures may be selected which maximizes cluster variance at the lowest cost. Or test resources could be allocated to clusters proportionally to cluster variance, and task-by-method measures selected within a cluster to maximize variance at the fixed level of cost.

Table 4

Judgment-Based Data Array For Cost-Effective
Selection Of Task-By-Method Measures

CLUSTER	TASK	WEIGHT	HO VAR	KN VAR	HO EFCY	KN EFCY
A	1	w_1	r^2_{H1}	$r^2_H(r^2_{HK})$	$w(r^2_H)/c_H$	$w(r^2_H)(r^2_{HK})/c_K$
	2	w_2	r^2_{H2}	.	.	.
	3	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.			
	.					
	N					
B	N+1					
	.					
	.					
	.					
C	.					
	.					
	.					

The merit of this approach depends entirely on the quality of the underlying task judgments. But if we can extract from task analysts judgmental data that lead to reasonable approximations of actual relationships among tasks and methods of task measurement, it would seem that content validity could be served by this cost-effectiveness approach to test constructions. Surely there is a systematic way for test development to benefit from the fact that knowledge tests, while not always valid measures of task proficiency, are substantially cheaper than hands-on tests.

REFERENCES

Foley, J. P., Jr. (1974). Evaluating maintenance performance: An analysis (AFHRL TR-74-57(I)). Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory.

Osborn, W. C., & Ford, J. P. (1976). Research on methods of synthetic performance testing (HumRRO Final Report FR-CD(L)-76-1). Alexandria, Virginia: Human Resources Research Organization.

Vineberg, R., & Taylor, E. (1972). Performance in four Army jobs by men at different aptitude (AFQT) levels: 4. Relationships between performance criteria (HumRRO Technical Report 72-73). Alexandria, Virginia: Human Resources Research Organization.

Wheaton, G. R., Fingerman, P. W., & Boycan, G. G. (1978). Development of a model tank gunnery test (ARI Technical Report). Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences.

Personal Constructs, Performance Schema, and "Folk Theories" of Subordinate Effectiveness: Explorations in an Army Officer Sample

Walter C. Borman
Personnel Decisions Research Institute

(Selection and Classification Technical Area Working Paper)

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine-year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Abstract

This research employs personal construct theory (Kelly, 1955) to explore the content of categories or schema that might be used in making work performance judgments. Twenty-five experienced U.S. Army officers, focusing on the job of non-commissioned officer (first-line supervisor), generated independently a total of 189 personal work constructs they believe differentiate between effective and ineffective NCOs. The officer subjects defined numerically each of their own 6-10 constructs by rating the similarity between each of these constructs and each of 49 reference performance, ability, and personal characteristics concepts. Correlations were computed between the subject-provided similarity ratings for each construct, and the 189 x 189 matrix was factor analyzed. Six interpretable content factors were identified (e.g., Technical Proficiency, Organization), with 123 of the 189 constructs from 23 of the 25 subjects loading substantially on these factors. Findings here suggest that a core set of concepts is widely employed by these officers as personal work constructs, but that different officers emphasize different combinations of this core set. Thus, substantial between-officer similarities and differences are evident. The personal constructs elicited from officer subjects are likened to performance schema and "folk theories" of job performance. Research is needed to assess the stability of these constructs over time and in different work contexts and to assess the impact of constructs on perceptions and evaluations of job performance.

The study described in this paper explores applications of personal construct theory (Kelly, 1955; Mancuso & Adams-Webber, 1982) to research in performance appraisal. In particular, attention is focused on "folk theories" of work behavior (Borman, 1983), performance constructs used naturally by persons very familiar with a job to make judgments about incumbents' effectiveness on the job. Preliminary data are presented that reveal what these dimensions might look like for experienced Army officer managers. Similarities and differences in construct content are also examined in this officer sample.

In a sense, the personal work constructs elicited here constitute an operational definition of schema (Feldman, 1981; Ilgen & Feldman, 1983; Nathan & Lord, 1983), those dimensions or categories raters use to organize performance information for making effectiveness judgments about persons on jobs. Thus, personal construct theory applied to performance appraisal may facilitate an understanding of the cognitive processes raters employ to, especially, interpret and evaluate ratee behavior related to making performance judgments. Before describing this exploratory work, a brief description of personal construct theory is in order.

Personal Construct Theory

As part of his ambitious psychological theory, Kelly (1955) observed that each person characteristically evolves, for his or her convenience in anticipating events (or other persons' activities), construction

systems reflecting his/her personal way of viewing and interpreting these events. That is, individuals develop personal construct systems which they use to judge events and to make predictions about future events. Most important for the present purpose is that some of these categories are imposed on their person perceptions. These "interpersonal filters" may influence observations and judgments about other people by providing frames-of-reference or sets that make perceivers look for certain kinds of interpersonal information and/or interpret this information according to their constructs (Duck, 1982).

Studies of personal constructs have typically had subjects generate their own categories, usually related to personal characteristics or personality traits. Lines of research have included: (a) the study of thinking sets among clinical groups such as thought-disordered schizophrenics (e.g., Epting, 1984; Widom, 1976); (b) the investigation of cognitive complexity (e.g., Metcalfe, 1974); (c) an examination of constructs and interpersonal relations (e.g., Duck, 1973, 1982); (d) the study of meaningfulness of interpersonal categories and extreme response style (e.g., O'Donovan, 1965); and (e) the investigation of individual differences in personal constructs (e.g., Sechrest, 1968). Important findings resulting from this research are that construct systems of clinical groups have predictably different structures than those of normals, subjects who like each other have constructs more similar to each other's than do strangers (even with appropriate controls for similarities in background),

own personal constructs are rated as more meaningful than experimenter-provided categories, subjects discriminate more between ratees on their own constructs, and substantial individual differences exist in construct structure.

Application of Personal Construct Theory to Performance Rating in Organizations

Personal construct theory has not to my knowledge been applied to the perception of individuals' work performance. Yet it seems reasonable that persons very knowledgeable about a job might develop over time constructs or categories they use to judge incumbents' performance on the job. Of particular interest here are possible similarities and differences in construct content that may have important implications for performance judgments and ratings. First, based on previous investigations of personal constructs in interpersonal perception research, it seems reasonable that there may be important individual differences in work-related constructs that, to a degree, affect what a rater looks for in observing ratee work behavior. Consider, for example, if one rater has an important construct, "getting along smoothly with others on the job," and a second rater does not share that construct or anything like it, the first rater may be more likely than the second to focus on work behavior related directly to that aspect of performance.

Although differences in constructs have been a major emphasis in past research, there may also be substantial similarities in work-related category systems across, especially, experienced supervisors. Such similarities may result from many observations of incumbents on the job that lead supervisors to similar views of what constitutes effective and ineffective performance.

The relationships of personal constructs to perceptions of work behavior may be akin to what might be called "folk theories" of work performance. Interviews with persons about work on jobs or even casual conversations with people about their jobs sometimes reveal what appear to be deeply felt and sometimes idiosyncratic "theories" of job performance. Consider these statements: A sales manager says with conviction, "You know what the key to this (sales) job is? Thinking on your feet with customers." And, a first-line supervisor speaks, "Show me a person who comes to work on time and I'll show you a good employee." Concepts such as these can be viewed as elements of folk theories and may reflect raters' category systems that help shape judgments about the effectiveness of individual employees.

Of course, characteristics of the work situation and employees themselves will in part dictate what raters observe and process when viewing work behavior. When a salesperson makes the largest sale in the history of the region, the regional manager rater is highly likely to attend to that piece of performance information no matter what the

content of his or her personal constructs might be. Also, other features of the situation that increase the salience of a particular construct will make perceivers' use of that construct more likely (Taylor & Fiske, 1978; Tversky, 1977). An example offered by Feldman (1981) is that race is more likely to be a salient construct when a ratee group has only one black than when it contains all blacks.

In spite of potentially relevant situational and ratee factors, the point to be emphasized here is that there may well be important similarities and differences in raters' personal construct systems related to observing and making judgments about work performance. Specifically, raters who have similar construct systems may tend to focus on like aspects of ratee performance and make similar evaluations of its effectiveness; differences in raters' constructs may lead to variations in the work behavior attended to and subsequently recalled in evaluating performance. Thus, personal construct similarities and differences may provide an inherent source of interrater agreement and disagreement. However, before it will be possible to explore this possibility, research is needed to (a) determine if raters actually have and can report meaningful personal constructs related to effectiveness on jobs; (b) examine individual differences in such constructs; (c) evaluate the stability of these constructs in assessing work behavior in different situations and contexts; and (d) assess the impact of these similarities/differences on observations of work behavior and ratings of work performance.

The present work is concerned with (a) and (b) above. Effectiveness constructs were elicited from experienced officers in the U.S. Army, and similarities and differences in these constructs were explored. A trait implication procedure (Borman, 1983) had subjects rate the similarity between each of their constructs and each of 49 reference constructs, yielding subject-provided numerical definitions of the constructs and allowing correlational analyses to describe the degree of similarity in content between different constructs.

METHOD

Subjects

Twenty-five officers in the U.S. Army participated in the research, focusing on the noncommissioned officer (NCO: first-line supervisor) job. All officers had at least two years' experience managing NCOs, and some had as many as twenty years' experience (M = 8.2). The officers were all from different units and had varying specialties (e.g., combat arms, engineering, intelligence).

Procedures

A variant of the Kelly (1955) Reperatory Grid was used to elicit personal work constructs from the officers. Kelly's procedure requires subjects first to identify persons they know who fit certain roles (e.g., mother, best friend, etc.) and then to examine triads of these

role persons (e.g., role person 1 and 3 vs. 7), describing in their own words how the two persons differ from the third. This is done for as many triads as is desired for the particular application.

In this research, officer subjects were asked to think of and record the names of nine NCOs they considered to be effective in their jobs and nine NCOs they considered ineffective in their jobs. Six triad combinations of these 18 role persons were then presented. Three triads consisted of two effective versus one ineffective, and the other three compared two ineffective versus one effective. Each role person appeared in one and only one triad. Subjects were asked (in the two effective vs. one ineffective NCO comparison) to record how the effective NCOs were different from the ineffective NCO; that is, what it was about the effective NCOs that differentiated them from the ineffective NCO. Subjects provided a label and a definition for each of these differentiating constructs.

After they made the six comparisons using the triads and generated six constructs apiece, they were asked to consider the effective and ineffective NCOs as two different groups and to record additional constructs that differentiated the two groups, if others occurred to them. These procedures resulted in a total of 189 personal work constructs for the 25 subjects (mean = 7.56, range = 6-10). Eight example constructs appear in Figure 1.

To obtain a numerical, subject-provided definition of each personal work construct, a trait implication procedure (Borman, 1983) was employed. This method requires a subject to rate the similarity between each of his/her constructs and a number of reference concepts. The vector of similarity judgments for a construct then constitutes a numerical definition of that construct, and correlational analysis can proceed between vectors of similarity ratings across different constructs (within or across subjects).

The critical first step in this procedure is to identify reference concepts. They should be as much as possible exhaustive of the target construct domain because the patterns of similarity ratings for individual constructs of course depends upon the domain represented.

Accordingly, 49 reference dimensions were developed to cover the following domains: (a) personal characteristics and personality traits; (b) cognitive and physical abilities; (c) performance constructs relevant to most or all Army enlisted jobs; and (d) military leadership constructs.

The personal characteristics/personality traits were identified by reviewing the constructs represented in major personality inventories, as well as taxonomic and factor analytic work done in personality research (Hough & Kamp, 1984). Sixteen personality attributes appeared to cover this domain (e.g., energy level, independence, persistence, etc.). The cognitive and physical abilities emerged from reviews of these constructs (Peterson, 1984; Peterson & Bownas, 1982). The nine cognitive and physical abilities included mechanical and verbal ability and physical coordination.

The performance dimensions were identified in a large-scale critical incidents study of enlisted soldier effectiveness (Borman, Motowidlo, & Hanser, 1983). Twelve dimensions (e.g., initiative and effort, adhering to regulations, supporting unit members, etc.) reflected a broad effectiveness domain including elements of technical job performance, organizational commitment, and organizational socialization. Finally, 12 leadership dimensions for NCO first-line supervisors were developed in an analysis of the NCO job (Hubein, Kaplan, Miller, Olmstead, & Sharon, 1983). These included administration of personnel, training soldiers, and organizing and controlling resources.

The 49 reference constructs were named and carefully defined. The intention was to have subjects rate on a 5-point scale the similarity between each of their own personal work constructs and each of the reference constructs (where 4 = my construct is very similar to the reference construct and 0 = my construct and the reference construct are completely different in meaning). However, a pilot test of this trait implication procedure indicated some difficulty, with several subjects indicating that the majority of the reference constructs were very or quite similar to each of their personal constructs. Upon debriefing, pilot subjects stated they could have spread out their similarity ratings to a greater extent for individual personal constructs, but they needed more guidance on how to distribute the similarity ratings. On the other hand, three subjects (of eleven) contributed well-differentiated similarity

judgments. Further, the distributions of these ratings for each construct were quite similar, both within and across subjects. Thus, a modified forced distribution corresponding to these pilot subjects' distributions was developed to serve as a target for subjects. The distribution for individual personal constructs across the 49 reference constructs was: 1-3, 4s; 3-5, 3s; 6-10, 2s; 9-13, 1s; and 20-28, 0s.

Officer subjects then used the 5-point scale, along with guidance on the target distribution, to make judgments about the similarity between each of their personal work constructs and each reference construct. Again, the notion here was to obtain the subject's own definition of his or her personal constructs, but in a numerical form that would allow correlational analyses to index similarities and differences in the content of different constructs.

Data Analyses

The focal analysis involved simply correlating the vectors of similarity ratings within and across subjects. To clarify, the number of variables in this analysis was the total number of constructs generated by the 25 subjects (189), and the N of each correlation was the number of reference constructs (49). The 189 x 189 correlation matrix was factor analyzed to explore the patterns of similarities and differences in content of the personal work constructs, both within and across subjects. In this manner, subject and content factors might be identified. For example, a factor with all constructs from an individual officer loading

substantially on it would suggest the subject has a highly related set of constructs and a comparatively idiosyncratic work construct system, with his/her constructs unrelated to others' constructs (subject factor). A factor highly interpretable and having work constructs from several subjects loading on it might, however, indicate a construct held in common across these officers (content factor).

We should emphasize that the identification of content factors was exploratory at this stage. Thus, factor analysis seemed appropriate for examining the possible existence of constructs shared by different officers. Future efforts to identify similarities in construct content might employ confirmatory factor analysis or other hypothesis testing procedures.

RESULTS

Factor analysis results are summarized in Table 1. The 8-factor solution was selected because of interpretability of factors and a substantial drop in eigenvalues for subsequent factors. Six of the factors are readily interpretable. Factors 3 and 8 appear to be primarily subject factors, and are uninterpretable from the standpoint of content.

To provide a richer description of the eight factors, the example constructs in Figure 1 were selected so that the first construct is one that loaded highly (> .70) on Factor 1, the second loaded highly on Factor 2, etc.

The distinction made previously between subject and content factors is obscured somewhat in the factor analysis results. Table 2 shows that Factors 3 and 8 are most like subject factors in that for each of these factors one or two officers have several constructs loading on it and very few of the other officers have any constructs associated with the factors. The other six factors can be considered more like content factors. Each is very interpretable and is shared by eight or more officers. Of course, some of the officers have two to five of their own factors loading on a single content factor.

Table 2 indicates just how much in common the content factors are across the 25 subjects. Constructs associated with three of the factors are held by the majority of the officers (Initiative/Hard Work, Maturity/Responsibility, and Technical Proficiency), and 11, 8, and 8 officers, respectively, have constructs related to the other three content factors (Supportive Leadership, Assertive Leadership, and Organization).

One way to look at the construct similarities/differences question is to consider the number of constructs loading primarily on the content factors. Table 2 indicates that fully 123 of the 189 personal work

constructs generated (65.1%) have substantial loadings on a content factor, and are thus shared with 7-17 other officer subjects. Of the 66 remaining constructs, 21 (11.1%) loaded on subject factors and 45 (23.8%) had mixed loadings or low communalities.

Focusing idiographically on individual subjects, the construct systems can be characterized one of four ways. The numbers in parentheses indicate the author's assignment of individual officers' construct systems into the four characterizations.

1. Differentiated--Loadings indicate three or more content factors represented, with less than three constructs on any one factor. (8): Subjects 4, 7, 9, 10, 14, 18, 21 and 22.

2. Idiosyncratic--Loadings are primarily on an uninterpretable subject factor or show low communalities. (5): Subjects 1, 5, 8, 20, and 24.

3. Narrow focus--Loadings are on content factors, but only one or two are represented. (3): Subjects 2, 17, and 25.

4. Differentiated but focused--Loadings show three or more content factors represented, but one or two factors are emphasized (with three or more high loadings on a single factor). (9): Subjects 3, 6, 11, 12, 13, 15, 16, 19, and 23.

All but five of the subjects have 50% or more of their constructs loading on content factors. Seventeen officers have three or more content-oriented constructs reflected in their systems, although nine of

these seventeen tend to focus on one or two content areas. Finally, three other officers hold constructs in common with other subjects, but they are heavily focused on just one or two content areas.

DISCUSSION

Results of this exploratory study show that persons very knowledgeable about a job can articulate what appear to be substantive categories of subordinate effectiveness on that job. Thus, personal construct theory (Kelly, 1955), found relevant in the area of interpersonal perception (e.g., Adams-Webber, 1979), apparently has meaningful application to the perception and interpretation of subordinates' work performance. Interestingly, the personal work constructs or "folk theories" of performance reported here demonstrate certain common themes across the 25 officer subjects. Fully 123 of 189 constructs generated by the officers reflect content related to six core construct composites that resulted from the factor analysis. Thus, whereas personal constructs in interpersonal perception research are often interpreted as very different in content across perceivers (Hamilton, 1971; Sechrest, 1968), the overall similarities in job performance constructs for the present subjects are as striking as the differences. Why might this be?

Compared to interpersonal dealings in general, making judgments about people in the performance effectiveness domain may involve fewer possible constructs to consider for successful functioning, and this

1982). Based upon research and theory in the areas of cognitive, personality, and social psychology (e.g., Cantor & Mischel, 1977; Rosch, 1978; Rosenberg & Sedlak, 1972; Srull & Wyer, 1979), performance appraisal researchers have presumed that schema are categories that raters use to help them organize performance information. Schema are like storage bins for perceptions of performance (Feldman, 1981). Feldman (1981) noted that the choice of schema or categories to employ in judging others' performance depends on both situational factors (especially various salience factors discussed previously, such as memorable, outstanding performance on the part of ratees) and person or perceiver factors.

Individual differences in such category systems should affect performance judgments. Within the context of schema, each officer's construct system articulated in this research can be considered as representing a repertoire of categories or schema that can be called up in gathering information about performance, making interpretations regarding ratee behaviors on the job, and evaluating the performance of ratees. Importantly, the study reported here provides a glimpse of the likely content of such schema and gives us an initial idea of similarities and differences in different manager's schema systems.

Future research on personal work construct systems should focus on the stability of these constructs for individual raters over time and in different performance situations and on the impact of constructs on perceptions and evaluations of ratee work performance. Regarding the latter, of special interest is the hypothesis that raters who have very

different construct systems look for and recall different samples of behavioral information and that they form evaluative judgments about performance based on these different samplings, thus providing an inherent reason for interrater disagreement in ratings. More generally, hopefully this study will open another line of research into the cognitive processes underlying performance judgments.

References

Adams-Webber, J. R. (1979). Personal construct psychology: Concepts and applications. New York: John Wiley.

Bannister, D., & Mair, J. M. M. (1968). The evaluation of personal constructs. London: Academic Press.

Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983). A construct approach to a general model of individual effectiveness. Paper presented at the meeting of the American Psychological Association, August, Los Angeles, CA.

Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. Journal of Personality and Social Psychology, 35, 38-48.

Cooper, W. H. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218-244.

Duck, S. W. (1973). Personal relationships and personal constructs: New York: Wiley.

Duck, S. W. (1982). Two individuals in search of agreement: The commonality corollary. In J. C. Mancuso & J. R. Adams-Webber (Eds.), The construing person. New York: Praeger.

Epting, F. R. (1984). Personal construct counseling and psychotherapy. Somerset, NJ: Wiley & Sons.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Hebein, J., Kaplan, A., Miller, R., Olmstead, J., & Sharon, B. (1983). NCO leadership: Tasks, skills, and functions (technical report). Alexandria, VA: Human Resources Research Organization.

Hamilton, D. L. (1970). The structure of personality judgments: Comments on Kuusinen's paper and further evidence. Scandinavian Journal of Psychology, 13, 261-265.

Hough, L. M., & Kamp, J. (1984). Temperament (a working paper). Minneapolis, MN: Personnel Decisions Research Institute.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In L. Cummings & B. Staw (Eds.), Research in organizational behavior (Vol. 5). Greenwich, CT: JAI Press.

Kelly, G. A. (1955). The psychology of personal constructs. New York: Norton.

Lord, R. G., Foti, R. J., & Phillips, J. S. (1982). A theory of leadership categorization. In J. G. Hunt, V. Sekaran, & C. Schriesheim (Eds.), Leadership: Beyond established views. Carbondale, IL: Southern Illinois University Press.

Mancuso, J. C., & Adams-Webber, J. R. (1982). The construing person. New York: Praeger.

Metcalfe, R. J. (1974). Own versus provided constructs in a Rep test measure of cognitive complexity. Psychological Reports, 35, 1305-1306.

Nathan, B. R., & Lord, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

O'Donovan, D. (1965). Rating extremity: Pathology or meaningfulness. Psychological Review, 72, 358-372.

Peterson, N. G. (1984). Indentification of candidate predictor constructs. In H. Wing (Chair), Expert judgments of predictor-criterion validity relationships. Symposium conducted at the meeting of the American Psychological Association Convention, August, Toronto.

Peterson, N. G., & Bownas, D. A. (1982). Skills, task structure and performance acquisition. In E. A. Fleishman & M. D. Dunnette (Eds.), Human performance and productivity: Human capability assessment. Hillsdale, NJ: Erlbaum.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorization. Hillsdale, NJ: Erlbaum.

Rosenberg, S., & Sedlak, A. (1972). Structural representations of implicit theory. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 6). New York: Academic Press.

Sechrest, L. B. (1968). Personal constructs and personal characteristics. Journal of Individual Psychology, 24, 162-166.

Srull, T. K., & Wyer, R. S. (1979). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgment. Journal of Personality and Social Psychology, 37, 1660-1672.

Taylor, S. E., & Fiske, S. T. (1978). Salience, attention, and attributes: Top of the head phenomena. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 11). New York: Academic Press.

Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.

Widom, C. S. (1976). Interpersonal and personal construct systems in psychopaths. Journal of Consulting and Clinical Psychology, 44, 614-623.

Author Note

The research was performed under U.S. Army Research Institute Contract MDA903-82-0531. This research program (Project A) is a long-term, large-scale effort concerned with improving the selection and classification of enlisted soldiers in the U.S. Army. Views expressed here do not necessarily reflect those of the Army or any other agency of the U.S. Government. I thank Elaine Pulakos, Dan Ilgen, Cris Banks, and Jack Feldman for reading a previous version of the manuscript and making several helpful suggestions.

1. Hardworking--Willing to work as long as necessary to accomplish the job; also concerned about the quality of the job.

2. Trustworthy--Once a job has been assigned there is no need to check on him (her).

3. Courage and Candor--Questions dumb rules and speaks own mind.

4. Priorities--Being able to identify those things that must take precedence over others.

5. Technical Proficiency--Knowledge of job and resources to accomplish mission; knows how to do the job better.

6. Firmness--Ability to control personnel and situations without falling apart.

7. Teacher of Soldiers--Always takes the extra time required to ensure soldiers know their task or mission before moving on.

8. Communicates Well--Communicates well with other soldiers, officers, etc., detailed and to the point, tactful, informative, good grammar.

Figure 1. Eight Example Constructs

Table 1

Summary Factor Analysis Results[a] of Correlations Between Personal Work Construct Similarity Judgments

Common Variance Accounted For	Factor	Factor Definition
20.7	1	Initiative/Hard Work--Having initiative to tackle jobs; self-starter; working hard and for long hours; dedication to tasks and the job; high energy and action orientation.
12.6	2	Maturity/Responsibility--Being consistently mature, responsible, and dependable; integrity and honesty; "good citizen."
9.2	3	Subject Factor--(Uninterpretable.)
7.4	4	Organization--Being well-organized; setting priorities; organizing subordinates and resources.
12.3	5	Technical Proficiency--Displaying technical proficiency and competence on job; possessing good job knowledge; knowing where to go for technical information (if needed); learning new concepts quickly and thoroughly.
7.8	6	Assertive Leadership--Working through subordinates to accomplish the mission; being confident and in control of subordinates; inspiring confidence in his/her leadership.

(table continues)

Common Variance Accounted For	Factor	Factor Definition
10.5	7	Supportive Leadership--Displaying concern for subordinates; teaching and providing feedback to help subordinates; supporting and guiding soldiers.
7.9	8	Subject Factor--(Uninterpretable.)
88.4		

[a]A principal factor analysis was conducted with varimax rotation (highest off-diagonal elements placed in diagonals).

Table 2

Summary of Officer Subjects' Personal Construct Systems

Officer Subject	Factor[a] Initiative/ Hard Work	Maturity/ Responsibility	Uninterpretable	Organization	Technical Proficiency	Assertive Leadership	Supportive Leadership	Uninterpretable	Mixed Loadings
1		1	4						1
2		3					3		3
3	3	1			1	1			1
4	2				1	1	1		1
5								4	2
6	1			1		3			1
7	1			1		1			3
8									7
9	1	2			1		2		
10	1	1			1		1		2
11	2	4		1	2				1
12	1	1		3			1	1	1
13	1		1		2		3	1	
14	1	1		1	1		1		
15	4			1	1				1
16	4				1		1		4
17	3			2					3
18		1			1	1		1	1
19	3		1		1		1		1
20		1	4		1				2
21	1	1	1		1	1			
22	2				1	1	1		1
23	4	1		1	2	1			
24					1		1		3
25	1	1	—	—	—	—	—	—	3
Totals	43	32	13	11	21	10	16	8	43

[a]The criteria for loading on a factor were, first, that this was the highest loading for the construct and, second, that it was .30 or above.

Development of a Model of Soldier Effectiveness

Walter C. Borman
Personnel Decisions Research Institute

Stephen J. Motowidlo
The Pennsylvania State University

Sharon R. Rose
Personnel Decisions Research Institute

Lawrence M. Hanser
Army Research Institute

(The appendices for this manuscript are reproduced separately in ARI Research Note 85-14, in press.)

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine-year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as may Army officers and enlisted personnel are participating in this landmark effort.

ABSTRACT

This report introduces a conceptual model of individual effectiveness that extends beyond successful performance on specific job tasks. The model of individual effectiveness suggested here contains elements of organizational commitment, socialization, and morale. The notion is that these broad constructs represent important criterion behaviors that contribute to an individual's worth to his or her organization and to his/her organization's effectiveness. The idea of the model is applied to the "job" of enlisted soldier in the U. S. Army, and 15 dimensions springing from the conceptual model are named and defined.

Empirical research was then conducted to explore these effectiveness constructs. The report presents results of behavioral analysis ~~or~~ BARS (Smith and Kendall, 1963) research to develop dimensions of soldier effectiveness. Seventy-seven Army officers and NCOs in six workshops generated a total of 1315 behavioral examples of soldier effectiveness. Although by no means a formal test of the individual effectiveness model, the content of the examples generated show similarities to elements of the model. Eleven dimensions emerged from behavioral analysis work and these results are discussed. Also discussed are advantages to taking a broader perspective on the performance criterion space in studying individual effectiveness, particularly in a military organization.

INTRODUCTION

This technical report describes work accomplished to define dimensions of U. S. Army first term soldier effectiveness. An initial conceptual and theoretical analysis along with subsequent empirical research was intended to define soldier effectiveness constructs appropriate for all first term enlisted jobs. The purpose of this effort was to develop "Army-wide" criterion constructs to describe first term soldier effectiveness dimensions and to aid in development of rating scales to use in evaluations of soldier effectiveness in any MOS.

Developing a Conceptual Model of Soldier Effectiveness

In developing a model of soldier effectiveness, we sought to expand the set of criterion behaviors considered, to include elements of individual effectiveness not directly related to task performance but related instead to a broader conception of job performance factors. In particular, elements were considered if they appeared to be potentially important contributors to organizational effectiveness in Army units. The notion here was that being a good soldier from the Army's perspective means more than doing the job properly, that is, performing tasks in a technically proficient manner. With this framework, a model of soldier effectiveness may include elements in addition to MOS job performances if they contribute to a soldier's effectiveness in the unit and to his or her "overall worth to the Army". The initial conceptual model development step was seen as useful for guiding thinking during subsequent empirical work to identify and define all elements of the model.

The conceptual model appears in Figure 1. It is a result of preliminary hypotheses about constructs that might be considered under the broad soldier effectiveness domain (Borman, Motowidlo, & Hansen, 1983). These constructs revolve around the areas of organizational commitment, organizational socialization, and morale.

Organizational Commitment -- The concept of organizational commitment (Porter, Steers, Mowday, & Boulian, 1974; Steers, 1977) refers to the strength of a person's identification with and involvement in the organization. It incorporates three kinds of attitudinal and cognitive elements: acceptance and internalization of organizational values and goals; motivation to exert effort toward the accomplishment of organizational objectives; and firm intentions of staying in the organization. The concept transcends job involvement and motivation to perform the specific tasks that comprise the job. It connotes a sense of loyalty to the organization as a whole and a desire to fulfill more general role requirements that come with organizational membership. We argue that the behavioral manifestations of organizational commitment may reflect one aspect of this broad conception of soldier effectiveness.

Organizational Socialization -- Van Maanen and Schein (1979) state, "In its most general sense, organizational socialization is the process by which an individual acquires the social knowledge and skills necessary to assume an organizational role." (p. 211). Some part of this knowledge

and skill is, of course, job-specific. For example, training programs designed to improve the effectiveness with which a person performs job-related tasks are part of the process of organizational socialization. But there are also many other knowledges and skills necessary for effective functioning as an organizational member that are not job-specific. When the socialization process is successful, a person will acquire not only job-related skills but also new patterns of behavior with subordinates, peers, and superiors in the organization, new attitudes, beliefs, and values in line with organizational norms, and new ways of using time not formally dedicated to performing job-related tasks.

Such individual changes are frequently crucial for assuring that the behaviors of different individual organization members will be smoothly coordinated toward accomplishing the organization's mission. As a result, soldier effectiveness might reasonably be regarded as partly a reflection of successful socialization; that is, people whose behavior and attitudes more closely coincide with Army norms might be regarded as more effective soldiers and considered of greater value to the Army.

Morale -- The concept of morale has traditionally been seen as an extremely important element in military organizations. Munson (1921), a former brigadier general writes:

"That their mental state, their will to do, their cooperative effort, their morale--all of which are synonymous--bear a true relation to their output, productivity, and the success of their joint undertaking, is so obvious and has been proven so often as to require no supporting argument." (p. 2)

The concept of military morale is multifaceted. It seems to involve feelings of determination to overcome obstacles, confidence about the likelihood of success, exaltation of ideals, optimism even in the face of severe adversity, courage, discipline, and group cohesiveness. (Motowidlo, Dowell, Hopp, Borman, Johnson, & Dunnette, 1976). Borman, Johnson, Motowidlo, and Dunnette (1975) report results of a study in the Army designed in part to identify behavioral dimensions of morale (see also Motowidlo & Borman, 1977). They found that the following dimensions efficiently describe behavioral expressions of morale among soldiers: community relations; teamwork and cooperation; reactions to adversity; superior-subordinate relations; performance and effort on the job; bearing, appearance, marching, and military courtesy; pride in unit, Army, and country; and use of time during off-duty hours. Because morale seems to figure so prominently as a determinant of unit effectiveness, behavioral dimensions like these may also in part represent important elements of individual soldier effectiveness.

In sum, we expect that the criterion domain of soldier effectiveness and worth to the Army is heavily saturated with elements of organizational

commitment, successful socialization, and morale. Our preliminary hypotheses, then were that soldiers who show high levels of commitment to the Army, acceptance of Army norms, and morale are more effective soldiers in this broader sense and are also of more value to the Army.

These three broad constructs can be viewed in another way that leads to more specific hypotheses about soldier effectiveness. From the combination of morale and commitment emerges a general category that can be labeled "Determination." It is a motivational and affective category that reflects the spirit, strength of character, or "will-do" aspects of good soldiering. Morale and socialization lead to "Teamwork", behaviors that have to do with effective relationships with peers and the unit. Commitment and socialization give rise to "Allegiance". This taps into acceptance of Army norms with respect to authority, faithful adherence to orders, regulations, and the Army life-style, and being adjusted and socialized to the point of wanting to continue in the soldiering role and stay in the Army. Each general category of effectiveness subsumes five more specific dimensions. These dimensions were developed based on the literature referred to above, along with a conceptual analysis of the elements likely to spring from constructs such as determination, teamwork, and allegience.

Figure 1 indicates how all of this fits together. As shown, the most abstract and broad construct, "Soldier Effectiveness", is defined according to somewhat narrower notions of "Morale", "Socialization", and "Commitment", which, with judicious mingling of conceptual elements, produce more concrete

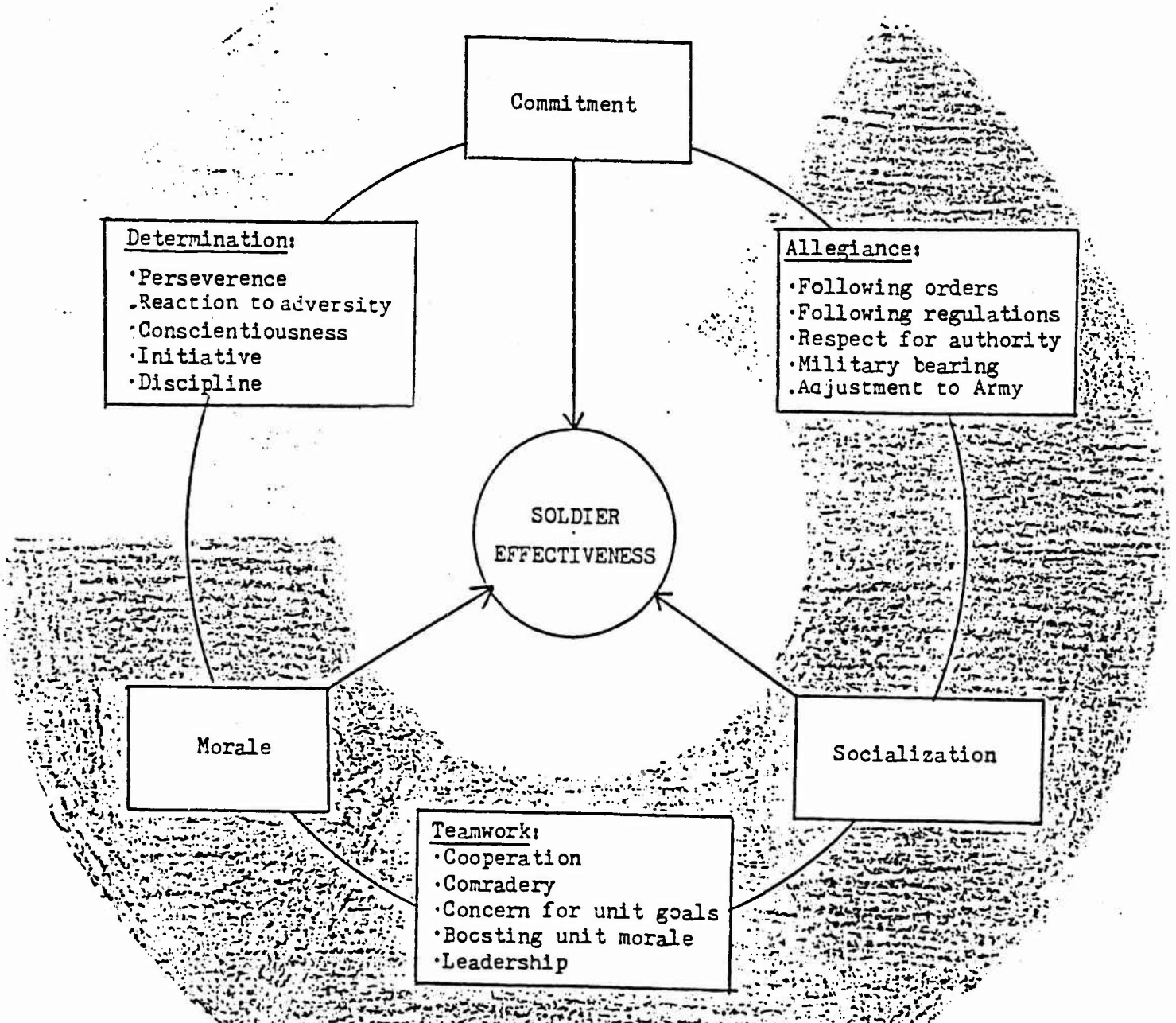
categories of "Determination", "Teamwork", and "Allegiance", which, finally, subsume more specific dimensions of soldier effectiveness. Figure 1 also contains these 15 preliminary dimensions of soldier effectiveness.

A brief caveat is in order here. The model purposely focuses on aspects of soldier effectiveness beyond task performance and other directly job-related performance elements. This is not to say that specific job performance factors are not important for soldier effectiveness. Clearly, they are. Our effort is simply to bring to light some of the factors less obviously related to soldier effectiveness.

As mentioned previously, the conceptual model was considered important to guide thinking in subsequent model development steps. However, we also believed strongly that an empirical strategy should be used to examine the soldier effectiveness domain. Accordingly, a variant of the critical incidents or behavioral analysis (Smith and Kendall, 1963) approach was employed to identify dimensions of soldier effectiveness. Specific procedures and results follow.

Figure 1

A Preliminary Model of Soldier Effectiveness



I. Allegiance

1. Following orders: responds willingly and eagerly to orders; carries out orders promptly and thoroughly; accepts direction from superiors without undue hesitation; versus responds half-heartedly to orders; carries out the letter but not the spirit of orders; refuses to obey orders.

2. Following regulations: complies with rules and regulations; conforms appropriately to standard procedures; tries to correct nonstandard conditions; versus frequently violates rules and regulations; ignores standard procedures when personally inconvenient; follows the letter but not the spirit of rules and regulations.

3. Respect for authority: defers appropriately to superiors' expertise and judgment; shows good military courtesy and respectful demeanor to superiors; speaks respectfully about superiors in conversations with others; versus habitually questions superiors' expertise and judgement; fails to salute properly or show military courtesy and respect in the presence of superiors; speaks disrespectfully about superiors in conversations with others.

4. Military bearing: grooms and dresses to maintain a crisp military appearance; stands, walks, and marches with an erect military posture; shows pride in the uniform and military insignia; versus grooms and dresses sloppily or without regard to military custom; stands, walks, and marches in a slouchy, casual, or careless manner; shows indifference toward the uniform and military insignia.

5. Adjustment to Army: successfully adjusts to military life; shows pride in being a soldier; wants to stay in the Army; versus fails to adjust to military life; shows indifference, dissatisfaction, or embarrassment about being a soldier; wants to leave the Army.

II. Teamwork

6. Cooperation: voluntarily pitches in when necessary to help other unit members with their job and mission assignments; willingly accepts personal inconvenience to aid other unit members with important problems; takes the trouble to listen and support other unit members with personal difficulties; versus pitches in only reluctantly when asked for job- or mission-related assistance; refuses to help other unit members with important problems if personally inconvenient; shows insensitivity and impatience with other unit members who have personal difficulties.

7. Comradery: is popular and well-liked by other unit members; forms close friendships with other unit members; spends off-duty time in group activities with other unit members; versus is unpopular or disliked by other unit members; frequently quarrels or fights with other unit members; remains aloof and spends off-duty time in solitary activities.

8. Concern for unit objectives: puts unit objectives before personal interests; makes personal sacrifices for the unit as a whole; works hard to meet unit objectives even when there is no personal gain; versus refuses to help meet unit objectives when they conflict with personal interests; shows more concern for personal interests than for the welfare of the unit; works for unit objectives only when there is personal gain.

9. Boosting unit morale: helps the unit stick together through hard times; encourages others to keep going when things seem bleak and hopeless; cheers others up when in unpleasant situations; versus shows no concern for unit solidarity; cynically criticizes others who refuse to give in for being foolish and unrealistic; constantly reminds others of the negative or unpleasant aspects of their situation.

10. Emergent leadership: shows good judgment in suggesting ideas for how others in the unit should proceed; persuades others to accept his/her ideas, opinions, and directions; others turn to this soldier for guidance and advice; versus never or rarely has good ideas for how others in the unit should proceed; presents opinions timidly and indecisively or is pushy and strident in rendering opinions, persuading/guiding others, etc.; others ignore this soldier's ideas, opinions, and directions.

III. Determination

11. Perseverence: struggles tenaciously to reach objectives even when the odds of success seem hopeless; sustains maximum effort over long periods of hard duty with unflagging stamina; versus gives up on objectives that cannot be easily reached; tires out quickly and takes frequent rest breaks.

12. Reaction to adversity: shrugs off severely uncomfortable or unpleasant conditions as though they were trivial; adapts and makes the best of hardship conditions without complaint; refuses to let troubles get him/her down; versus exaggerates the severity of minor discomfort and unpleasantness; constantly complains and grumbles about the lack of amenities; loses perspective and becomes demoralized by insignif nt troubles.

13. Conscientiousness: spends extra time and effort to get the job done; consistently completes job and duty assignments promptly on or ahead of schedule; carries out assignments with thoroughness and careful attention to detail; versus refuses to take extra steps to make sure the job gets done; is frequently slow or late in completing assignments; works sloppily and ignores important details.

14. Initiative: volunteers for assignments; anticipates problems and takes action to prevent them; performs extra necessary tasks without explicit orders; versus refuses to volunteer for assignments; waits passively until difficulties occur and reacts only to the immediate problem; does only what explicitly ordered to do.

15. Discipline: devotes full concentration to the job at hand without yielding to the temptation of distractions; controls self-indulgent appetites and does not allow them to interfere with the performance of duty; keeps emotions in check and almost never loses his/her temper; versus easily distracted by opportunities to play, socialize, or pursue other leisure activities; lets too much eating, drinking, sleeping, or other self-indulgent appetites interfere with the performance of duty; fights or destroys property in uncontrolled emotional outbursts with little provocation.

METHOD

Summary of Procedures

The inductive behavioral analysis strategy (Campbell, Dunnette, Arvey, & Hellervik, 1973) requires persons familiar with a job's performance demands to generate examples of effective, mid-range, and ineffective behavior observed on that job. In the present application, "job behavior" was defined broadly as any action related to soldier effectiveness, and NCO and officer participants in workshops were asked to generate behavioral examples from any aspects of what they considered to be the first term soldier effectiveness domain. Behaviors generated were to be appropriate for and applicable to any MOS.

The many behavioral examples emerging from this step were content analyzed to form dimensions or categories of soldier effectiveness and then submitted to a retranslation procedure. In retranslation, officers and NCOs evaluated each example, placing it in a category and rating the level of effectiveness reflected. Those examples that showed good agreement in the retranslation step were used to form behavioral statements anchoring different levels of effectiveness on each of the dimensions.

Behavior Analysis Workshops

Seventy-seven officers and NCOs participated in 6 one-day workshops intended primarily to elicit behavioral examples of soldier effectiveness. Table 1 presents the ranks and sex of all workshop participants.

In each workshop, the leader first provided an introductory briefing, describing the Project A program. He or she explained that Project A is a large scale effort concerned with improving the selection and classification of enlisted soldiers in the U. S. Army. The workshop leader then distributed the orientation materials that appear in Appendix A.

A very important section of these materials is the training program designed to help workshop participants get started writing behavioral examples. As the reader can see in Appendix A, the training has a modeling orientation in which participants are shown improperly written examples and, importantly, these examples corrected to the proper form. Participants were led through this training and then asked to write a first behavioral example. Workshop leaders reviewed the first example and provided corrective guidance as needed. Except for periods taken to discuss behavioral examples or effectiveness dimensions emerging from the content of the examples, the rest of each workshop was devoted to participants' writing and leaders' reviewing the examples. Below are two such examples to provide a flavor for the output from the workshops.

Table 1

Participants in Behavior Analysis Workshops

NCOs (N=30)

Rank	n	Sex	n
SP4	1	Male	28
E-5	5	Female	2
E-6	14		
E-7	12		

Officers (N=47)

Rank	n	Sex	n
First Lt.	3	Male	44
Captain	29	Female	3
Major	15		

Total (N=77)

	Per Cent
NCOs	39
Officers	61
Male	94
Female	6

- This soldier was in a group setting around a tree when a senior officer walked toward them. He called the group to attention and saluted the officer.

- When this soldier was assigned to guard a bivouac area at night on an FTX, he fell asleep at one of the training stations even though he knew he was supposed to be walking the post.

In this manner, 1315 behavioral examples were generated in the six workshops. Details of this data collection appear in Table 2. Duplicate examples and those examples which did not meet the criteria specified in training (e.g., the incident described the behavior of an NCO rather than a first term soldier) were dropped from further consideration. The remaining examples were edited to a common format and content analyzed to form preliminarly dimensions of soldier effectiveness. Three of the authors independently read each example and grouped together those examples which described similar behaviors. The sorted examples were then reviewed and the groupings or dimensions were revised until each author arrived at a set of dimensions that were homogenous with respect to their content. After discussion among project staff and with a small group of officers and NCOs at Fort Benning, a consensus set of 13 dimensions was decided upon. The behavioral examples and dimensions were then readied for retranslation. Specifically, the remaining 1111 non-redundant examples were placed in retranslation booklet form.

Table 2

Soldier Effectiveness Examples Generated

Location	Participants	Number of Examples	Mean Examples Per Participant
Ft. Benning	14 Officers	228	16
Ft. Stewart	13 Officers	266	20
Ft. Stewart	13 NCOs	216	17
Ft. Knox	12 Officers	239	20
Ft. Benning	13 NCOs	149	11
Ft. Carson	8 Officers 4 NCOs	217	18
TOTALS:	77	1,315	OVERALL MEAN: 17

A Retranslation of the Behavioral Examples

Retranslation provides a way of checking on the clarity of individual behavioral examples and of the dimension system. As mentioned, in retranslation, persons familiar with the target domain make two judgments about each example -- the dimension or category it belongs to based on its content and the effectiveness level it reflects. Examples for which there is disagreement related either to category membership or to the rated effectiveness level may be unclear and require revision or elimination from further consideration. Also, confusion between two or more categories in the sorting of several examples may reflect a poorly formed and/or defined category system.

In this project, the retranslation task was divided into five parts, with each subtask requiring a retranslation judge to evaluate 216-225 behavioral examples. The division into subtasks was accomplished to keep reasonable the amount of time each judge would be required to spend on the rating task. Judges were provided with definitions of each dimension to aid in the sorting and a 1-9 effectiveness scale (1 = extremely ineffective; 5 = adequate/average; and 9 = extremely effective) to guide effectiveness ratings. The retranslation materials, including all edited behavioral examples, appear in Appendix B. Sixty-one officer and NCO judges completed retranslation ratings, and the results are presented next.

RESULTS

Table 3 shows the number of behavioral examples reliably retranslated for each of the 13 dimensions. Typically employed acceptance points of greater than 50% for the sorting into a single dimension and less than a 2.0 standard deviation for the effectiveness ratings left 87% of the 1111 examples (78%) included for subsequent scale development work. Appendix B contains effectiveness scale means and standard deviations for each behavioral example, along with the percentage of retranslation raters sorting each example into each dimension.

Results in Table 3 were seen as satisfactory in that sufficient numbers of reliably retranslated examples were available to develop extensive behavior definitions of each dimension. The first two authors considered for each dimension all examples reliably retranslated into that dimension in the above average range (5-9) in writing a behavioral definition of effective performance for that aspect of the model. The same procedure was followed for each dimension in the below average range (1-4.99). The content of the reliably retranslated behavioral examples was summarized in a behavioral definition. The result of this exercise was 13 relatively elaborate definitions of effective and ineffective behavior in each of the model's dimension areas.

The length of the behavioral summary statements was seen as excessive for the rating scales. There was a concern that the amount of reading time required to understand the content of each dimension would lead

Table 3

Number of Behavioral Examples Reliably Retranslated[a]

Into Each Dimension

	Dimension	Number of Examples
A.	Controlling own behavior related to personal finances, drugs/alcohol, and aggressive acts	107
B.	Adhering to regulations, and SOP and displaying respect for authority	158
C.	Displaying honesty and integrity	53
D.	Maintaining proper military appearance	34
E.	Maintaining proper physical fitness	36
F.	Maintaining own equipment	46
G.	Maintaining living and work areas to Army/unit standards	23
H.	Exhibiting technical knowledge and skill	47
I.	Showing initiative and extra effort on job/ mission/assignment	131
J.	Attending to detail on jobs/assignments/ equipment checks	59
K.	Developing own job and soldiering skills	40
L.	Effectively leading and providing motivation to other soldiers	71
M.	Supporting other unit members	65
		870

[a] Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and had standard deviations of their effectiveness ratings of less than 2.0.

raters using the definitions as guides for rating soldiers' effectiveness to lose patience with the rating task or otherwise short-cut the rating procedures. Therefore, developing shorter versions of the behavioral definitions for the rating scales appeared advisable. It was also decided that preparing behavioral definitions for three levels of effectiveness (rather than the two provided by the more elaborate definitions) would help raters to differentiate between ratees. Finally, again in the spirit of shortening the rating task, two pairs of dimensions were combined; leading other soldiers and supporting other unit members were combined to form Leading/Supporting and attending to detail and maintaining own equipment were collapsed to form Maintaining Assigned Equipment. The two collapsings were seen as justifiable because of the conceptual similarity of each of these dimension pairs.

At this point, the first two authors used the reliably retranslated behavioral examples at three levels (1-3.49; 3.5-6.49; 6.5-9) to write behavioral summary statements to capture the content of the specific examples. In the main, this was very straightforward, with the behavioral statements written reflecting the content of many specific examples. For some dimensions, however, because of few examples written to the mid-range of effectiveness, it was necessary to interpolate behavioral content of the high and low effectiveness examples to create the middle level behavioral summary statements. We did not find this difficult to accomplish, but it must be acknowledged that these summary statements are not based quite so solidly on empirical data as are the others.

Development of these behavioral summary statements is the critical step in forming Behavior Summary Scales. This format development procedure has been touted as a highly conceptually sound method for developing rating scales (Borman, 1979). The main advantage of these scales over the traditional behaviorally anchored rating scales is that for a particular dimension and level of effectiveness, the content of all examples reliably retranslated is represented on the scale, not just one of the specific behavioral examples. This makes it more likely that a rater using the scales will be able to match observed performance with performance on the scale. It has been argued (Borman, 1979) that this feature of Behavior Summary Scales is very desirable.

The products from this work are displayed as follows. One of the dimension definitions appears in the text of the report (Figure 2), and all of these definitions[1] are included in Appendix C. Likewise, one of the rating scales appears as Figure 3, and all 11 scales are presented in Appendix D.

---

[1] To conform with the dimension configuration used for the rating scales, dimension definitions previously written for the dimensions that were combined were re-written to be consistent with the 11-dimension system.

Figure 2

An Example Behavioral Definition

F. Maintaining Assigned Equipment

Checking on and maintaining own weapon/vehicle/other equipment

1. Consistently keeping assigned equipment clean, including own weapon and vehicle, as appropriate.
   - Ensuring that weapon and vehicle are constantly up to standard, resulting in high marks on inspections and no deadlining necessary; following proper procedures for cleaning weapon.
   - Painting, polishing or otherwise substantially improving the appearance of assigned vehicle and/or other pieces of equipment when appearance is important.
2. Performing proper checks and preventive maintenance on assigned weapon, vehicle, and other equipment.
   - Properly inspecting all equipment responsible for to make sure it is safe and that no damage will occur as a result of equipment problems (e.g., always checking on water and oil levels on vehicle).
   - Pulling proper services on vehicle according to schedule, and ensuring that all deficiencies are noted; lubricating own weapon and/or other equipment, as necessary.
3. Ensuring that equipment is repaired when necessary.
   - Performing effectively in simple troubleshooting and repair tasks related to maintaining assigned equipment (e.g., weapon, vehicle, etc.).

(continued)

- On more difficult troubleshooting/repair jobs or as regulations/procedures dictate, ensuring that equipment deficiencies are corrected by appropriate support personnel.

Ineffective Performance

1. Maintaining assigned equipment in dirty and/or sloppy condition, including own weapon, vehicle, and/or gear.
   - Often leaving assigned weapon dirty; failing to keep weapon in combat-ready or ready-for-inspection shape, not cleaning weapon before returning it to ammo room, or failing to follow proper procedures in cleaning weapon.
   - Maintaining dirty and/or rusty gear/equipment such as assigned vehicle, sleeping bag, entrenching tools, etc.; refusing to, being reluctant to, or otherwise failing to ready assigned equipment for important inspections or exercises.
2. Failing to perform or improperly performing checks and preventive maintenance on assigned weapon, vehicle, and other equipment.
   - Inspecting equipment haphazardly, skipping steps in servicing sequence, ignoring safety checks on equipment, etc. such that, later, equipment problems may develop.
   - Failing to make daily or other routine checks on assigned pieces of equipment resulting in, at times, no-go inspection marks or even damage to equipment; failing to note deficiencies related to assigned weapon/vehicle/other equipment.

(continued)

3. Having in possession or actually using assigned equipment in need of repair, even when repair job is easy or repair services are available.

   . Being unable to perform simple troubleshooting and repair tasks related to maintaining assigned equipment (e.g., weapon, vehicle, etc.).

   . Even when repair services are available failing to get equipment to them to get it fixed, or unnecessarily delaying getting it into repair.

Figure 3

An Example Behavior Summary Rating Scale

B. Initiative/Effort

Showing initiative and extra effort on the job/mission/assignment

1 2	3 4 5	6 7
Below Standard	Adequate/Mid-Range	Superior
. Refuses to volunteer for assignments or put in extra hours and effort; may even react with hostile attitude when asked to volunteer or work long hours.	. Volunteers for some assignments and puts in extra effort when it's very important to do so.	. Volunteers enthusiastically, takes initiative promptly and effectively when opportunitites arise, and voluntarily works long, extra hours to complete assignments, even without being asked.
. Gives up easily when faced with obstacles, adversity, or discomfort.	. Hangs in there with determination when it's really important to overcome obstacles on the job in the field, etc.	. Refuses to give in to adversity and pushes on with stamina and guts to overcome all obstacles until the assignment is completed.

DISCUSSION

Conceptual and Empirical Model Development

The model described here was designed to portray elements of soldier effectiveness in a context broader than successful performance on job-related tasks. It is an effort to tap elements of criterion behaviors that are important for organizational effectiveness, but that are not necessarily directly task-related. The model presumes that soldier effectiveness involves commitment, socialization, and morale and suggests more specific dimensions that underlie effectiveness in the soldiering role regardless of what the individual's particular job might be.

This approach has the advantage of forcing a broad perspective on the criterion domain. It points out potentially important elements of individual effectiveness that might be overlooked by more "accepted" approaches to job and task analysis. For this reason, we believe the model was useful for guiding efforts to impose structure upon the soldier effectiveness criterion space.

In particular, for empirical model development work, the conceptual model suggested the potential usefulness of asking officers and NCOs to refer to a broad conception of soldier effectiveness when contributing behavioral examples. It was reasoned that if conceptual model development can yield

such a rich sampling of effectiveness criteria, systematic empirical identification of such criteria may also result in an expanded array of important criterion elements.

Thus, behavioral analysis efforts proceeded, with officer and NCO workshop participants contributing behavioral examples of soldier effectiveness pertaining to several facets of this criterion domain. Although by no means a formal empirical test of the conceptual model, the behavioral analysis work did yield dimensions similar to those hypothesized by the earlier model. Eleven dimensions emerged and were thoroughly defined based on the content of many behavioral examples of soldier effectiveness. Also, behavioral rating scales were developed with shorter behavioral summary statements defining and anchoring three different effectiveness levels of each scale.

Model Dimensions as Criteria in Selection Research

The point was made that criteria of individual effectiveness such as organizational commitment/socialization and morale may be important as contributors to organizational effectiveness, even though they are not directly task-related. Discussion concerning these links between individual characteristics and organizational effectiveness suggest this may be the case (e.g., Mowday, Porter, & Steers, 1982). Also, recent work on the closely related construct of "organizational citizenship" (Bateman & Organ, 1983; Smith, Organ, & Near, 1983) assumes this kind of

linkage between organization members' standing on the dimensions of Altruism (helping other organization members) and Generalized Compliance (a more impersonal form of conscientious citizenship) and positive effects on organizational unit functioning. Confirmation of substantive links between these individual characteristics and organizational effectiveness is hard to come by because of difficulties measuring the effectiveness of organizations (Campbell, 1977). However, on balance, we believe that constructs such as commitment, socialization, and morale are likely important in this regard. Organizations with members who are commited, well adjusted to unit norms, etc. should tend to be more effective, at least along certain dimensions.

It follows, then, that in the interests of enhancing organizational effectiveness, an important question is what are the antecedents and "causes" of a unit members' commitment, socialization, morale, citizenship, and specific factors identified in the empirical model of soldier effectiveness? Considerable literature presumes that organization-related factors such as job characteristics (Hackman & Oldham, 1975), control a good deal of the variance in the kinds of variables considered in the model. However, it is also possible that to some extent individuals enter organizations with proclivities toward high or low levels of commitment, adjustment or morale. This phenomenon could take the form of an interaction between person and organization, where individuals have personal characteristics that make it likely they will be commited or not commited, well adjusted or poorly adjusted, etc.

in organizations with certain features. Or, less likely, the phenomenon could take the form of a main effect, where individuals have personal characteristics that impact upon their commitment, adjustment, etc. related to any organization.

This is not new. Although conventional wisdom states that organizational factors control most of the variance in these kinds of dependent variables, Locke, for example, (1969, 1976) has argued for the existence of a person-situation interaction in determining levels of satisfaction (closely related to morale). Individual differences are posited to interact with organizational factors to determine satisfaction. This suggests that although features of the organization are important in this context, characteristics the person brings with him or her may also contribute to satisfaction and perhaps impact on the other criteria in the model discussed here.

Related views have been expressed by Blood (1969), Schneider (1976), Schmitt and Schneider (1983), and Pulakos and Schmitt (1983). Blood (1969) found that individual differences in worker values were related to subsequent job satisfaction. Schneider (1976) and Schmitt and Schneider (1983) suggested that individuals' personal characteristics might be important contributors to their satisfaction on jobs, and Pulakos and Schmitt (1983) demonstrated that for graduating high school students certain needs related to jobs correlated positively with satisfaction nine and twenty weeks into their first job experience.

Related to the model of soldier effectiveness, we submit that other criteria potentially important for organizational effectiveness (in addition to satisfaction) may fit into this framework. That is, individuals' organizational commitment/socialization and other elements of the model, as well as morale/satisfaction probably make important contributions to an organization's effectiveness, and further, it may be possible to identify personal characteristics in job candidates that portend high commitment, socialization, morale/satisfaction, etc. in the hiring organization.

The main point then is that the soldier effectiveness model's criterion elements that extend beyond directly task-related performance criteria may also fit into a personnel selection framework. Provided these elements are important for organizational effectiveness and that these criteria can be predicted by the skills, abilities, and personal characteristics individuals bring with them to the organization, the model's dimensions should definitely be considered in addition to job performance criteria in selection research and practice.

Potential Problems Related to Consideration of These Criteria in Selection Context

Although there are potential advantages to broadening the scope of the performance effectiveness domain to include elements of effectiveness that are not job-specific, this approach also carries inherent risks. As we

move from the relative concreteness and immediacy of effectiveness in specific job-related tasks, the importance of less job-related elements such as following orders and military bearing for organizational effectiveness becomes less obvious and direct.

Even though the workshop participants cited examples of behavior that indicate these kinds of elements are important for soldier effectiveness, it is not obvious that soldiers who are exceptionally good or poor in those areas necessarily contribute to or detract from the success of the Army's overall mission. It is much more obvious that soldiers who perform their jobs well or poorly contribute to or detract from organizational effectiveness. On the other hand, although such dimensions might seem somewhat removed from effective contribution to the Army's mission, we believe they may help shed light on patterns of behavior that have important implications for Army organizational effectiveness.

In sum, the model of soldier effectiveness, as depicted in the behavioral definitions and the rating scales, offers a behavior-based description of the criterion elements important for first term soldier effectiveness. These criterion elements, some of them directly relevant to task performance, others related to a broader view of soldier effectiveness, are appropriate for evaluating first term soldiers in any MOS. The behavioral definitions springing from the model provide an in-depth description of the performance requirements for first term soldiers. And, the rating scales provide a format for generating supervisor, peer, and self assessments of effectiveness in all important aspects of the domain.

References

Bateman, T. S., & Organ, D. W. (1983). Job satisfaction and the good soldier: The relationship between affect and employee "citizenship". Academy of Management Journal, 26, 587-595.

Blood, M. R. (1969). Work values and job satisfaction. Journal of Applied Psychology, 53, 456-459.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

Borman, W. C., Johnson, P. D., Motowidlo, S. J., & Dunnette, M. D. (1975). Measuring motivation, morale and job satisfaction in Army careers. Minneapolis: Personnel Decisions, Inc.

Campbell, J. P. (1977). Organizational effectiveness as a construct. In P. Goodman & L. H. Pennings, (Eds.), New perspectives in organizational effectiveness. San Francisco: Jossy-Bass.

Campbell, J. P., Dunnette, M. D., Arvey, R., & Hellervik, L. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

Eaton, N. K., & Shields, J. L. (1982, August). U. S. Army Soldier Selection, Classification, and Utilization Research Program. Paper presented to 90th Annual Convention of the American Psychological Association.

Hackman, R. J., & Oldham, G. (1975). Development of the job diagnostic survey. Journal of Applied Psychology, 60, 159-170.

Locke, E. A. (1969). What is job satisfaction? Organizational and Human Performance, 4, 309-336.

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally.

Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 62, 177-183.

Motowidlo, S. J., Dowell, B. E., Hoppe, M. A., Borman, W. C., Johnson, P. D., & Dunnette, M. D. (1976). Motivation, satisfaction, and morale in Army careers: A review of theory and measurement (ARI Technical Report TR-76-A7). (NTIS No. AD-A036390). Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Munson, E. L. (1921). The management of men. New York: Holt.

Porter, L. W., & Lawler, E. E. (1965). Properties of organization structure in relation to job attitudes and job behavior. Psychological Bulletin, 64, 23-51.

Porter, L. W., Steers, R. M., Mowday, R. T., & Boulian, P. V. (1974). Organizational commitment, job satisfaction, and turnover among psychiatric technicians. Journal of Applied Psychology, 59, 603-609.

Pulakos, E. D., & Schmitt, N. (1983). A longitudinal study of a valence model approach for the prediction of job satisfaction of new employees. Journal of Applied Psychology, 68, 307-312.

Schmitt, N., & Schneider, B. (1983). Current issues in personnel selection. In K. M. Rowland & J. Ferris (Eds.), Research in personnel and human resources management (Vol. 1). Greenwich, CT: JAI Press.

Schneider, B. (1976). Staffing organizations. Santa Monica, CA: Goodyear Publishing.

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. Journal of Applied Psychology, 68, 653-663.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for ratings scales. Journal of Applied Psychology, 47, 149-155.

Steers, R. M. (1977). Antecedents and outcomes of organizational commitment. Administrative Science Quarterly, 22, 46-56.

Van Maanen, J., & Schein, E. H. (1979). Toward a theory of organizational socialization. In B. M. Staw (Ed.), Research in organizational behaviors (Vol. 1). Greenwich, CT: JAI Press.

Appendices A - D of

DEVELOPMENT OF A MODEL OF SOLDIER EFFECTIVENESS

are reproduced in ARI Research Note 85-14 (in press)

III. PREDICTOR MEASUREMENT

Norman G. Peterson

The major activities completed during the second year of Project A with respect to predictor measure development were:

(1) The definition and identification of the most promising predictor constructs.

(2) The administration and initial analysis of the Preliminary Battery.

(3) The development, tryout, and pilot testing of the first version of the Trial Battery, called the Pilot Trial Battery.

(4) The development and tryout of psychomotor/perceptual measures, using a microprocessor-driven testing device.

All of these activities were aimed primarily at developing the Trial Battery, which will be completed and administered to a large sample of soldiers in the third year of Project A in accordance with the concurrent validation research design. Figure 16 is a flow chart of the major activities devoted to predictor measurement on Project A and shows the relationships between these activities. The numbers on the figure correspond to the activities listed above. Each of these activities is described briefly.

Predictor Development

Construct Definition. The first activity, defining and identifying the most promising predictor constructs, was accomplished in large part by using experts to provide structured, quantified estimates of the empirical relationships of a large number of predictors to a set of Army job performance dimensions (the dimensions were defined by other Project A researchers). By pooling the judgments of 35 experienced personnel psychologists, we were able to more reliably identify the "best" measures to carry forward in Project A.

These estimates were combined with other information (from the literature review and Preliminary Battery analyses) and evaluated by consortium and ARI scientists and members of the Scientific Advisory Group (SAG). A final, prioritized list of constructs was identified.

This effort also produced a heuristic model, based on factor analyses of the experts' judgments, that organizes the predictor constructs and job performance dimensions into broader, more generalized classes and shows the estimated relationships between the two sets of classes. This effort is fully described in Wing, Peterson, and Hoffman (1984).

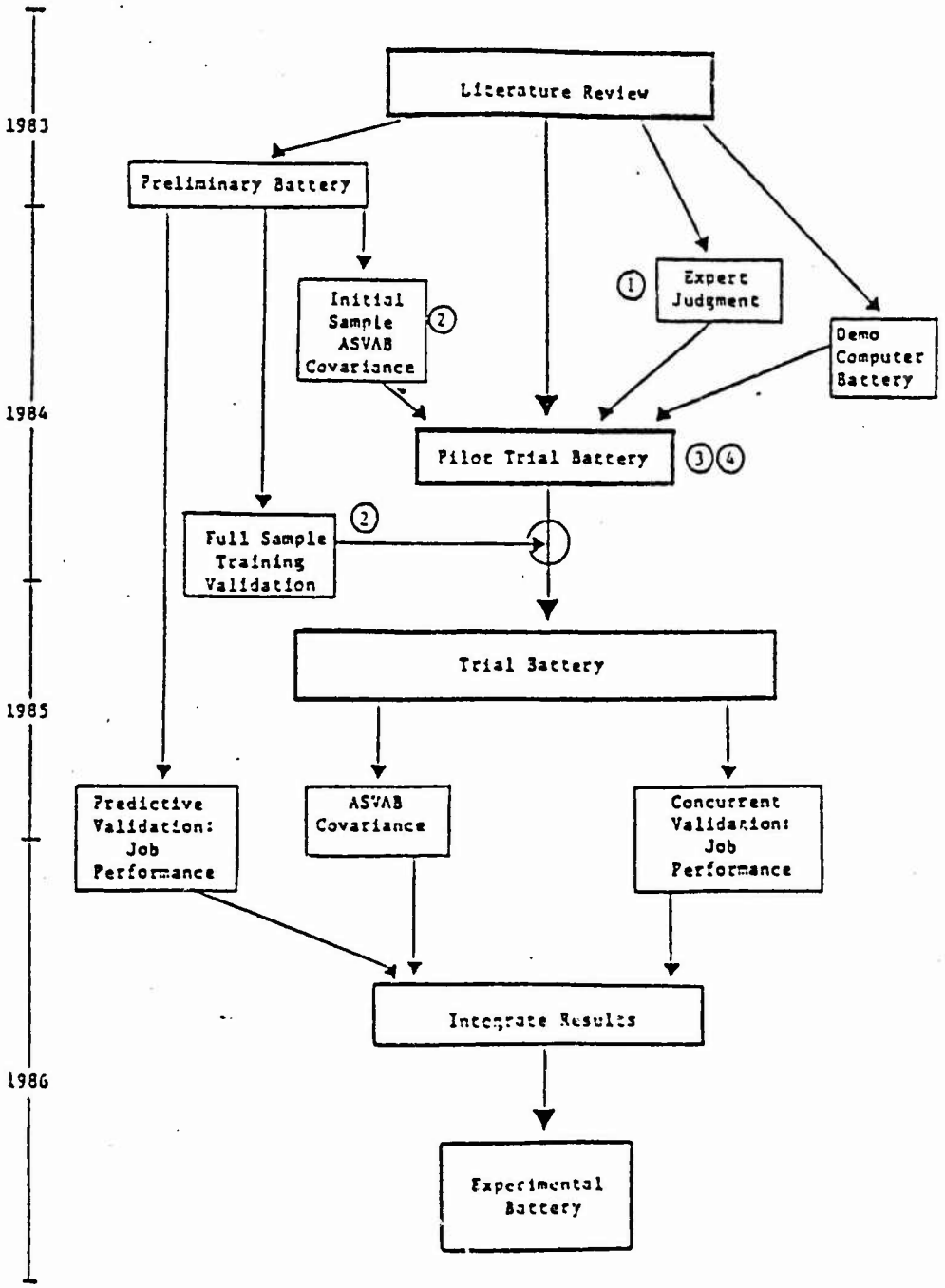


Figure 16. Flow Chart of Predictor Measure Development Activities of Project A

Preliminary Battery. Similarly, the initial analyses of Preliminary Battery data provided empirical results to guide our Pilot Trial Battery test development efforts. Data were collected with the Preliminary Battery on four MOS during the second year of the project. These four MOS were 05C (Fort Gordon), 19E/K (Fort Knox), 63B (Fort Dix and Fort Leonard Wood), and 71L (Fort Jackson).

The first 1800 cases from this sample were used in the initial analyses. These analyses enabled us to tailor the Pilot Trial Battery tests more closely to the enlisted soldier population.

They also demonstrated the relative independence of cognitive ability tests and non-cognitive inventories of temperament, interest, and biographical data. This effort is fully reported in Hough et al. (1984).

A total of just over 11,000 Preliminary Battery cases were collected during Project A's second year. These data will be further analyzed to verify and extend the findings of the initial analyses. Most important, as Figure 16 indicates, the PB measures will be correlated with training performance measures to provide data for use in revising the Pilot Trial Battery during the third year of the project.

Pilot Trial Battery. The information from the first two activities fed into the third activity: the development, tryout, revision, and pilot testing of new predictor measures, collectively labeled the Pilot Trial Battery. New measures were developed to tap the ability constructs that had been identified and prioritized. These measures were tried out on three separate samples, with improvements being made between tryouts. The tryouts were conducted at Forts Carson, Campbell, and Lewis with approximately 225 soldiers participating.

At the end of the second year, the final version of the Pilot Trial Battery underwent a pilot test on a larger scale. Data were collected to allow investigation of various properties of the battery, including distribution characteristics, covariation with ASVAB tests, internal consistency and test-retest reliability, and susceptibility to faking and practice effects. About 650 soldiers participated in the pilot test.

Computerized Measures. The development, tryout, revision, and pilot testing of computerized measures is actually a subset of the Pilot Trial Battery development effort, but is worthy of separate mention. During the first year of the project, the literature review, site visits to military laboratories currently investigating computerized measures, and the programming of a demonstration battery laid the groundwork for FY84 activity.

Several objectives were reached during 1984. An appropriate microprocessor was identified and six copies were obtained for developmental use. The ability constructs to be measured were identified and prioritized. Software was written to utilize the microprocessor for measuring the abilities and to administer the new tests with an absolute minimum of human administrators' assistance. A customized response pedestal was designed and fabricated so that responses would be reliably and straightforwardly obtained from the people being tested. The software and hardware were put through an iterative tryout and revision process.

Pilot Trial Battery

Shown next is a general overview of the content of the Pilot Trial Battery, including the general ability area, method of measurement, number of tests or inventories, time to complete the tests, and total number of items.

Perceptual/Psychomotor Measures - Computer

Ten Tests
100 Minutes
343 Items

Cognitive Measures - Paper-and-Pencil

Ten Tests
100 Minutes
343 Items

Non-cognitive Measures - Paper-and-Pencil

Two Inventories
90 Minutes
Assessment of Background and Life Experiences (ABLE):
- Four Validity Scales
- Eleven Substantive Scales
- 270 Items

Army Vocational Interest Career Examination (AVOICE):
- Twenty-four Basic Interest Scales
- Six Organizational Climate/Environment Scales
- 309 Items

Figures 17 and 18 provide more detail about the substance of the Pilot Trial Battery. The cognitive/perceptual/psychomotor measures are shown in Figure 17. The predictor categories (left column) are the predictors that were identified as most promising, as described earlier. The Pilot Trial Battery test names are given in the right column. Note that ASVAB also appears in this column. This denotes that there is an ASVAB subtest that at least partially measures that predictor. Tests marked with an asterisk are administered via the computer-driven testing device.

Figure 18 shows the content of the two non-cognitive inventories, the Assessment of Background and Life Experiences (ABLE) and the Army Vocational Interest Career Examination (AVOICE). The AVOICE is a modified version of an inventory developed by the U.S. Air Force. Note that the Climate Environment Scales were not identified as essential predictors, but have been included at this point to measure individuals' perceptions of their organizations' environment.

Predictor Category	Pilot Trial Battery
Verbal	ASVAB
Memory	*Short Term Memory *Number Memory
Number Facility	ASVAB *Number Memory
Perceptual Speed and Accuracy	ASVAB *Perceptual Speed and Accuracy *Target Identification
Reasoning/Induction	Reasoning Test 1 Reasoning Test 2
Information Processing	*Simple Reaction Time *Choice Reaction Time
Spatial: Orientation	Orientation 1 Orientation 2 Orientation 3
Closure/Field Independence	Shapes
Spatial: Visualization	Object Rotations Assembling Objects Path Mazes
Mechanical Information	ASVAB
Multilimb Coordination	*Target Shoot *Target Tracking 2
Precision	*Target Shoot *Target Tracking 1
Movement Judgment	*Cannon Shoot

*Computerized

Figure 17. Cognitive/Perceptual/Psychomotor Measures In the Pilot Trial Battery

Predictor Category	Pilot Trial Battery	
	AVOICE Scales	
Realistic vs. Artistic	Mechanics	Drafting
	Heavy Construction	Audiographics
	Marksman	Electronic Communication
	Electronics	Infantry
	Outdoors	Armor/Cannon
	Agriculture	Vehicle Operator
	Law Enforcement	Adventure
		Aesthetics
Investigative	Medical Service	
	Mathematics	
	Science/Chemical	
	Automated Data Procesing	
Enterprising Interests	Leadership	
Social Interaction	Teaching/Counseling	
Conventionality	Office Administration	
	Food Service	
	Supply Administration	
(N/A)	Climate Environment Scales	
	Achievement	Status
	Safety	Altruism
	Comfort	Autonomy
	ABLE Scales	
Stress Tolerance/Adjustment	Emotional Stability	
	Self-esteem	
Dependability/ Conscientiousness	Non-delinquency	
	Traditional Values	
	Conscientiousness	
Achievement/Work Orientation	Work Orientation	
Physical Condition/Athletic Abilities/Energy	Physical Condition	
	Energy Level	
Potency/Leadership	Dominance	
Locus of Control/ Work Orientation	Internal Control	
Agreeableness/Likability/ Sociability	Cooperativeness	

Figure 18. Non-cognitive Measures in the Pilot Trial Battery: The Army Vocational Interest and Career Examination (AVOICE) and the Assessment of Background and Life Experiences (ABLE)

Summation

At the end of the second year, the Pilot Trial Battery had been developed to measure a carefully identified and prioritized set of predictor constructs. It had been subjected to an iterative process of writing, trying out, and revising that resulted in a 6.5-hour battery of tests. Pilot test data were collected that will provide information for further refinement of the Pilot Trial Battery, especially a reduction in length. Ultimately this process will result in the Trial Battery that will be administered to over 12,000 soldiers in Year 3 of the project. In addition, more than 11,000 soldiers had completed the Preliminary Battery. Analyses of these data had informed the development of the Pilot Trial Battery, and further analyses will affect the refinement and reduction of the Pilot Trial Battery.

Associated Reports and Papers

Several reports have been prepared to record the details of early analyses in the various prediction studies.

(1) The validity, in predicting success in training, of the cognitive tests that make up the present ASVAB was tested by Martin, Rossmeissl, and Wing for both the Aptitude Area composites and the Armed Forces Qualification Test (AFQT). For 11 MOS with sufficient data to permit assessing prediction, the overall corrected validity coefficient was .66 for the composites and .64 for the AFQT. Initial data were obtained on prediction by racial and gender subgroups.

(2) Results from the technical review of possible predictor and criterion measures were presented in a paper by Wing, Peterson, and Hoffman. Expert judgments of the validity of each of 53 predictors against each of 72 criteria were obtained and analyzed. (The text of this report follows this section; Appendices A-L are being reproduced in ARI Research Note 85-14, in press.)

(3) Data from the first two months of the administration of the Preliminary Battery were analyzed in a paper by Hough, Dunnette, Wing, Houston, and Peterson. Covariance analyses of the results from the cognitive and the noncognitive measures provided guidance for continuing modification of materials for the Trial Battery.

(4) The potential power and the hazards of using meta-analysis techniques were pointed out in an introductory paper by Wing. The technique provides a way to combine the results of research from different studies.

(5) A review of the theory and methods involved in processing verbal information, prepared by Mitchell during FY83, is being published as ARI Technical Report 648. It included development of a general model of text processing, which was used as a conceptual framework in assessing the cognitive processing contributions to verbal subtest performance.

BLANK PAGE

Validity of Cognitive Tests in Predicting Army Training Success

Clessen J. Martin
Paul G. Rossmeissl
Hilda Wing

Army Research Institute

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

This paper was presented at the Psychonomics Society, San Diego, November 1983.

Validity of Cognitive Tests in Predicting Army Training Success

Clessen J. Martin
Paul G. Rossmeissl
Hilda Wing
U.S. Army Research Institute for the
Behavioral and Social Sciences

Introduction

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple Cognitive abilities test battery used by all the military services for selection and classification of enlisted personnel. ASVAB Forms 8/9/10 were introduced in October, 1980. ASVAB 8/9/10 is comprised of ten subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto/Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). For purposes of Army Selection and Classification these subtests are combined into aptitude area (AA) composites as illustrated in Table 1.

Table 1

The Composition of the ASVAB Composites

Operational Army Composites

AFQT	= VE + AR + .5NO
Electronics (EL)	= AR + EI + MK + GS
Operators/Foods (OF)	= NO + VE + MC + AS
Surveillance/Communications (SC)	= NO + CS + VE + AS
Motor Maintenance (MM)	= NO + EI + MC + AS
Clerical (CL)	= NO + CS + VE
Skilled Technical (ST)	= VE + MK + MC + GS
Combat (CO)	= AR + CS + MC + AS
Field Artillery (FA)	= AR + CS + MC + MK
General Technical (GT)	= VE + AR
General Maintenance (GM)	= MK + EI + GS + AS

Within the Army Composites, the AFQT is used for the initial selection of personnel and the other composites are used for the assignment of soldiers to the various MOS (Military Occupational Specialities) or jobs within the Army.

The purpose of this research is to determine the validity of ASVAB Forms 8/9/10 both in relation to AFQT and to the ten Army Composites in predicting success in training. While the Army uses both training and job performance (Skill Qualification Tests) criteria for test validation, the requirements associated with the conduct of this research necessitated only the use of training criteria.

Method

Criteria and Sample

The Army does not routinely record end-of-training grades on a soldier's personnel file. For this reason, during calendar year 1981 the Army Research Institute (ARI) collected training data for all MOS with 100 or more entrants per year. Included in these data were end-of-course grades for each soldier. Collection was terminated at 1000 for the high density MOS, and at the end of the year for the remaining MOS. It is these end-of-course grades which formed the criterion measures for this research. It was not possible to find useful criteria for all MOS. Many did not show sufficient variance in the end-of-course grade to be useful in the assessment of predictor validities. For example, in the MOS 16E, 92% of the grades reported were at the maximum value of 100. The analyses of this research were, therefore, limited to a sample of 11 MOS shown in Table 2. These MOS were selected because they all had a fairly large N (defined as 90 or greater) and a training score standard deviation greater than five. Summary statistics for the criterion measures from these MOS are given in Table 3.

Table 2

MOS included in the Research

MOS	Name	Army Composite
05G	Signal/Security Specialist	SC
16P	Short Range Missile Crewman	OF
16S	MANPADS Crewman	OF
32D	Tech Controller	EL
33S	Electronic Warfare Systems Repairer	ST
61B	Watercraft Operator	MM
61C	Watercraft Engineer	OF
67Y	Attack Helicopter Repairer	MM
68J	Attack Fire Control Repairer	EL
71D	Legal Clerk	CL
76P	Material Control & Accounting Specialist	CL

Table 3

Summary Statistics for Training Criteria

MOS	N	Training Score Mean	Training Score S D
05G	91	84	7.3
16P	101	83	14.2
16S	514	79	8.3
32D	120	81	14.2
33S	103	82	9.0
61B	92	80	7.7
61C	150	83	6.9
67Y	137	83	6.3
68J	128	86	6.1
71D	96	73	22.9
76P	613	87	5.1

Analyses

The data for the MOS listed in Table 3 required further editing before any validation analyses were performed. First, scores for all soldiers who had attrited from training for non-academic reasons were dropped. Standard scores were then computed for those remaining. Academic attrites were assigned a score of one standard deviation below the minimum passing score and academic recycles were assigned a score that was one-half of a standard deviation below the minimum passing score. This differential score assignment to attrites and academic recyles has been a conventional procedure in ARI validation research involving pass/fail training criteria and does reflect different underlying considerations between these two failure groups.

Two sets of predictors were validated against the criteria measures from each MOS: AFQT and the appropriate Army Aptitude composite. Uncorrected validities for these predictors were obtained using standard regression analyses. In addition, a stepwise regression (Draper & Smith, 1966) based upon the ten ASVAB subtests was conducted for each MOS. The results of this analysis can be interpreted as the "best" fit of the ASVAB subtests to the criterion data and, therefore, could be used as an index for the fit of the other predictors. Validities for the composite predictors corrected for restriction in range were obtained using Lawley's (1943) general case method. This method can be shown to be mathematically identical to that proposed by Gulliksen (1950).

Finally, whenever the N within an MOS was sufficiently large to perform meaningful subgroup analyses, the above procedures were repeated for subgroups within the MOS. The variables of race and

gender were used in the definition of these subgroups.

Results and Discussion

The results and discussion of this research will be divided into two topics: AA composites and subgroup analyses.

Operational Army Composites

Table 4 presents the validity coefficients obtained from each of the 11 MOS for both AFQT and the appropriate AA composite. For each validity coefficient, the uncorrected and corrected value for restriction in range has been computed. Also reported for each MOS is the uncorrected stepwise best fit estimate based upon all 10 subtests of the ASVAB. Inspection of the uncorrected validity coefficients for the stepwise best fit analysis reveals that in all cases, these values are higher than for either the AFQT or for the corresponding AA composite. The average increment among the 11 MOS for the uncorrected stepwise values in comparison to the AA composite value was .10.

Table 4

Corrected and Uncorrected Validities for Operational Army Composites

MOS	Uncorrected Stepwise Best Fit	AFQT Uncorrected/ Corrected	Army Composite Uncorrected/ Corrected
05G	.61	.55/.81	(SC) .48/.79
16P	.28	.15/.30	(OF) .21/.36
16S	.28	.17/.40	(OF) .23/.44
32D	.46	.44/.67	(EL) 43/.67
33S	.66	.46/.84	(ST) 56/.87
61B	.51	.49/.69	(MM) .45/.65
61C	.58	.45/.73	(OF) .45/.75
67Y	.45	.29/.66	(MM) .39/.75
68J	.53	.28/.62	(EL) .44/.73
71D	.41	.38/.65	(CL) .27/.64
76P	.48	.40/.68	(CL) .26/.60

Somewhat surprising is that, for four of the 11 MOS, AFQT yielded a higher corrected validity coefficient than did the corresponding Army composite. In no instance was the increase greater than .08 as in 76P. Correspondingly, the increased predictive validity for the Army composites in relation to AFQT was greatest in 68J and 67Y, where the increase was .11 and .09, respectively.

Inspection of the validity coefficients for the Army composites corrected for restriction in range, revealed validities ranging from .36 for 16P to .87 for 33S with an average validity coefficient of .66. The average corrected validity coefficient for AFQT was .64. The largest validities were obtained for the Skilled Technical(ST) composite (.87) and for Surveillance/ Communications(S/C) composite (.79) and the lowest average validity was for the Operators/Food (O/F) Composite (.52).

Subgroup Analyses

There were sufficient sample sizes in 16S to examine the validity coefficients separately for Blacks and Whites. Since 16S is currently not available to women because it is a combat specialty, gender comparisons were not possible. However, in 76P it was possible to examine both race and gender differences. Table 5 presents the uncorrected validities separately for Blacks and Whites in 16S for AFQT and the Operators/ Food composite. For the present Army O/F composite there was relatively little difference between the corrected validity coefficients for Blacks and Whites (.53 and .51, respectively). A somewhat larger difference between Blacks and Whites was observed for the AFQT (.47 and .68, respectively).

Table 5

Validities for 16S
by Race
Uncorrected/Corrected

	Blacks	Whites
n	159	333
AFQT	.03/ 47	.21/.68
O/F Composite	.16/.53	.28/.51

Table 6 presents the uncorrected and corrected validities in 76P, separately for Blacks and Whites as well as males and females, on AFQT and the Clerical composite. As in 16S, the corrected validities are somewhat higher for Whites than for Blacks and especially so for White females. However, due to the relatively small sample size for White females, this difference should be interpreted with caution.

Table 6

Validities for 76P
by Race and Gender
Uncorrected/Corrected

	Blacks	Whites
Males		
n	273	143
AFQT	.28/.69	.60/.73
CL component	.12/.57	.47/.65
Females		
n	116	38
AFQT	.26/.62	.51/.77
CL composite	-.02/.46	.41/.69

Conclusions

Given the relatively small sample sizes and small number of the Army MOS analyzed, the results of this validation research were generally favorable with respect to the validity of ASVAB 8/9/10. The overall corrected validity coefficient for the Army composites was .66. In the two MOS where it was possible to analyze validities separately for Blacks and Whites, the average corrected validity coefficient for Blacks was .52 and .62 for Whites. In 76P where it was possible to compare the corrected validity coefficients for males and females, the resulting values were .61 and .58, respectively. While the overall validity of the six Army composites analyzed was .66, the average validity for AFQT across all 11 MOS was .64. This result suggests that the Army composites examined in this research contribute relatively little in differential prediction of training criteria. This is not surprising given the limited focus of this research to training performance in a relatively few MOS. However, ongoing research at ARI with a larger number of more heterogeneous MOS, using both training and job performance criteria, is expected to provide more definitive results.

References

Draper, N.R. & Smith, H. Applied Regression Analysis. New York, Wiley, 1966.

Gulliksen, H. Theory of Mental Tests. New York, Wiley, 1950.

Lawley, D.M. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 273-287.

Expert Judgments of Predictor-Criterion Validity Relationships

Hilda Wing
Army Research Institute

Norman G. Peterson
Personnel Decisions Research Institute

R. Gene Hoffman
Human Resources Research Organization

August 1984

(The appendices for this manuscript are reproduced separately in ARI Research Note 85-14, in press)

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

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.

Presented at the Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Expert Judgments of Predictor-Criterion Validity Relationships

Since World War II decisions regarding selection and classification of Army enlisted personnel have been based on instruments which have been empirically validated. The primary criterion has been training performance and the predictors have been cognitive abilities. In recent years an increased interest and emphasis have been placed on job performance subsequent to training. This has led to consideration of a predictor-criterion space expanded to include the noncognitive variables of perceptual and psychomotor abilities and of vocational interests and temperament. Empirical linkages among such a variety of variables are incompletely established. The U.S. Army Research Institute's Project A, Improving the Selection, Classification, and Utilization of Army Enlisted Personnel (U.S. Army Research Institute, 1983),is both developing new predictor and criterion measures as well as improving existing ones, to investigate this expanded predictor-criterion space.

The identification, for development or improvement, of an efficient and effective set of initial or pre-induction predictors of soldier performance in the expanded predictor and criterion domains requires that the expected validities of predictor-criterion combinations be hypothesized. Such validities reflect the degree of congruence between test performance scores on selection measures and performance scores on job criteria collected sometime later within the job contexts where persons have been placed. The criterion-related validity coefficients gathered by the Army over the past four decades, designed to identify consistent patterns of predictor-criterion relationships, provide one basis for developing hypotheses. The major limitation of this set of statistics is that its coverage of the expanded predictor and criterion domains is incomplete.

Other strategies are possible for developing hypotheses which link predictors with job criteria. For example, Wernimont and Campbell (1968) suggested that with greater behavioral congruence between predictors and criteria, higher predictive validities may be expected. This is a content-oriented strategy which relies on the logical analyses of job activity components to identify directly measurable tasks and knowledge. Thus, predictor measures selected should actually sample elements of the job performance domain. Based on this advice, the Army has examined the predictive validity of job samples using simulations of the job context (Campbell & Black, 1982; Johnson, Jones & Kennedy, 1984). However, criteria have often been training measures rather than on-the-job measures of performance, and results have been disappointing, perhaps because of the difficulty of simulating the job. Another strategy for developing hypotheses about predictor and criterion relationships is more construct-oriented and relies on the theoretical analysis of job activity requirements to determine behavior patterns at a more abstract level than the content-oriented approach. An increasing number of scholars (e.g., Cronbach, 1980; Dunnette & Borman, 1979; Guion, 1980; Messick, 1980; Peterson & Bownas, 1982) are emphasizing the interrelatedness of all three (content, construct- and criterion-related) approaches. Each of these approaches suggests important issues for consideration.

Hypothesizing the predictive validity of potential selection measures

must begin with a content-oriented strategy focusing on the criterion domains of interest. In the present context, the areas of performance to be considered are effectiveness in specific Army tasks associated with Army jobs, general adjustment to a career in the Army, and success in Army school training. Each of these areas must be analyzed to collapse specific indices into content categories which summarize the criterion domain. Once criterion categories are determined, abstractions concerning the psychological attributes or constructs which are required by the criterion categories must be generated according to existing behavioral theory. The identification of these psychological constructs, each linked to specific criterion categories, then provide the hypotheses about potentially valid measures for predicting soldier effectiveness. Of course, the criterion-related strategy is then required for confirmation of the proposed linkages with empirically demonstrated statistical relationships.

In the research literature, a large number of empirical validity coefficients are available for some of the predictor-criterion combinations cusrently of interest to the Army. Hunter, Schmidt, and Jackson (1981) have presented methods for use in helping to decide whether or not such validity coefficients and results may be generalized across different situations and populations. Fluctuations in observed sample validities arise from fluctuations in observed sample range restrictions, criterion and predictor reliabilities, and, most importantly, from sampling error resulting from the small sample sizes used in most validity research. Several recent investigations (Dunnette, Rosse, Houston, Hough, Toquam, Lammlein, King, Bosshardt, & Keyes, 1981; Pearlman, Schmidt, & Hunter, 1980 ; Schmidt, Hunter, & Caplan, 1981; Callender & Osburn, 1981; Schmidt, Hunter, & Pearlman, 1981) of large data sets have shown that validities of cognitive tests are very likely to be generalized across employment settings, and for both training and job performance criteria. Such generalizability means that tests with non-zero validities for one job in one setting tend to have non-zero validities for other jobs in other settings. On the other hand, the magnitude of those non-zero validity coefficients may fluctuate such that important moderator effects occur when tests are combined with differential weighting (Schmidt, Hunter, & Pearlman, 1981) The variance in validity coefficients which does remain, apart from situational artifacts, indicates that different jobs do carry unique requirements, (Dunnette et al., 1981, Linn, Harnisch, & Dunbar, 1981.

Such "unique" influences may include noncognitive variables which are not addressed in the existing set of predictor or criterion measures. Another possibility is that certain specialized variables (e.g., spatial ability, psychomotor skills) may be highly valid for some occupations but minimally important for others, suggesting the importance of such variables for classification purposes after selection decisions have been made. Thus, it is important that the development of hypotheses about relationships between potential predictor constructs and the full range of criterion categories be exhaustive. Yet that undertaking must be manageable.

The approach used here is to (1) identify criterion categories, (2) identify an exhaustive range of psychological constructs which may be potentially valid predictors of those criterion catregories, and (3) obtain expert judgments about the relationships between the two. Schmidt, Hunter, Croll, and

McKenzie (1983) showed that pooled expert judgments, obtained from experienced personnel psychologists, were as accurate in estimating the validity of tests as actual, empirical criterion-related validity research using samples in the hundreds of subjects. That is, experienced personnel psychologists are effective "validity generalizers" for cognitive tests. They do tend to underestimate slightly the true validity as obtained from empirical research. Would such judges be as effective validity generalizers for noncognitive tests as well? Schmidt et al. do not know and suggest additional research.

Hence, one way to identify the "best bet" set of predictor variables and measures is to use a formal judgment process employing experts such as that followed by Schmidt et al. (1983). Peterson and Bownas (1982) provide a complete description of the methodology which has been used successfully by Bownas and Hechman (1976), Peterson, Houston, Bosshardt, and Dunnette (1977), Peterson and Houston (1980), and Peterson, Houston, and Rosse (1984) to identify predictors for the jobs of firefighter, correctional officer, and entry-level occupations (clerical and technical), respectively. Descriptive information about a set of predictors and the job performance crierion variables are given to "experts" in personnel selection and classification, typically personnel psychologists. These experts make estimates of the relationships between predictor and criterion variables by rating or directly estimating the value of the correlation coefficients.

The result is a matrix with predictor and criterion variables as the columns and rows, respectively. Cell entries are experts' estimates of the degree of relationship between the particular predictors and various criteria. The interrater reliability of the experts' estimates is checked first. If the estimate be sufficiently reliable (previous research shows values in the .80 to .90 range for about 10 to 12 experts), the matrix of predictor-criterion relationships can be analyzed and used in a variety of ways. By correlating the columns of the matrix, the covariances of the predictors can be estimated on the basis of the profiles of their estimated relationships with the criteria. These covariances can then be factor analyzed in order to identify predictors which function similarly with regard to predicting performance criteria. Similarly, the criterion covariances can be examined to identify clusters of criteria predicted by a common set of predictors.

Such procedures facilitate the identification of redundancies and overlap in the predictor set. The common sets or clusters of predictors and of criteria are an important product for a number of reasons. First, they provide an efficient and organized means of summarizing the data generated by the experts. Second, the summary is in a form which makes for easier comparison with the results of meta-analyses of criterion-related validity coefficients. Conflicting or absent evidence is a sure guide to important research questions. Certain clusters may require reconfiguration based on new data. Third, more indirect but potentially more important, these clusters provide a model or theory of predictor-criterion performance space. "Nothing is as practical as a good theory," although it is difficult to predict just what practical contributions will be made. The contributions for personnel selection and classification are most obvious. Other possibilities include the enhancement of systems design, to include these empirically derived dimensions of individual differences early in the process so that new equipment will be as effective as

it should be. Training for such new systems could also be based on this model.

Method

Identification of Predictor Variables

The list of predictor variables should include the relevant areas of the predictor space (cognitive, perceptual, psychomotor, biographical, vocational interest, temperament) while the variables chosen should provide maximum accuracy in selection and classification. Variables pertaining to physical strength and stamina, while important, were excluded from coverage in this research because it was judged that the necessary expertise (in physiology) would not be available among the proposed judges. Further, other Army research units are responsible for developing and implementing standards in this area.

The first step was the usual "exhaustive" literature search. The search was conducted by three research teams, each responsible for a broadly defined area of human abilities or characteristics. The three areas were cognitive abilities; noncognitive characteristics such as vocational interests, biographical data, and measures of temperament; and, psychomotor/perceptual abilities. These areas or domains proved to be convenient for purposes of organizing and conducting literature search activities, but they were not used as (nor were they intended to be) a final taxonomy of possible predictor measures.

Several methods were used to insure as comprehensive a search for sources as possible. We conducted computerized searches of seven data bases that we judged most relevant for our purposes. Their names and descriptions are shown in Appendix A. Over 10,000 sources were identified via the computer search. In addition to the computerized searches, we obtained reference lists from recognized experts in each of the three areas, emphasizing the most recent research in the field. We also obtained several annotated bibliographies from military research laboratories. Finally, we scanned the last several years' editions of research journals that are frequently used in each ability area as well as more general sources such as textbooks, handbooks, and appropriate chapters in the Annual Review of Psychology.

As is usually the case with such exhaustive search techniques, the majority of the sources identified were not directly relevant for our purposes, the identification and development of promising measures for personnel selection in the U.S. Army, and they were screened out in step 2. Relevance was broadly defined, measures that could appropriately be applied to a population of "normally" functioning, adult individuals were retained for further scrutiny. (Examples of non-relevant measures: those intended to detect neurological problems; end-of-course or achievement tests targeted at specific content areas, primary-school achievement tests.) For the most part, we were able to make the relevance judgment based on abstracts of the sources obtained in the search.

In step 3, the relevant sources were reviewed and two record forms were completed for each source: an article review form and a predictor or variable review form (several of the latter could be completed for each source). These forms were designed to capture the essential information in a standard format. (Copies of these forms are shown in Appendix B.) This was necessary because of the incredible diversity of reporting formats that occurs across journal articles, technical reports, and books. Part of the review process for each variable included categorizing it into an initial taxonomy of predictors, reported in Peterson and Bownas (1982).

Each variable was then evaluated using the twelve factors shown in Figure 1. At least two researchers independently rated each variable on each factor. Discrepancies were resolved by discussion. We note here that the information available for the variables did not always allow us to make an evaluation of each variable on all twelve factors. The evaluations were used to select the final set of predictor variables for use in the expert judgment process. Variables were included if they received generally high marks and if they added to the comprehensiveness of coverage for a particular domain of predictor variables. At this point, we began to depart somewhat from our initial taxonomy and to create a new one that we felt best represented the entire predictor domain relevant for our goal: selection and classification of enlisted recruits for the U.S. Army. There were 53 members in the final set of 53 predictor variables. The names and definitions of these variables are shown in Table 1.

The fifth and final step was the preparation of materials describing each fo the 53 variables. The expert judges were experienced psychologists and were generally familiar with psychometric information and, in varying degrees, knowledgable about the 53 variables in our final list. Therefore, the descriptive material was designed to transmit a large amount of information, but as efficiently as possible. Each packet contained a sheet that named and defined the variable, described how it was typically measured, and provided a summary of the reliability and validity of measures of the variable. Following this sheet were descriptions of one or more specific measures of the variables. These descriptions included the name of the test, its publisher, the variable it was designed to measure, a description of the items and the number of items on the test (in most cases, sample items were included), a brief description of the administration and scoring of the test, and brief summaries of studies of the reliability and validity of the measure. Appendix C includes an example of one of these packets.

Identification of Criterion Variables

Specific Job Task Criterion Categories. The purpose of this portion of the work was to reduce the job task domain of job performance to a set of descriptors that could be used as criterion variables against which the potential effectiveness of predictor measures could be judged. Short of enumerating all job tasks in the nearly 280 entry level job specialties, the nature of the performance domain had to be characterized in a way that was at once comprehensive, understandable and usable by judges. Since many jobs share similar tasks, the abstraction of generic task categories was possible. Two approaches were tried and the results of one chosen for use.

One approach began with a preliminary analysis of task statements for fourteen of the nineteen jobs previously selected for intensive research in Project A. These jobs were the ones for which survey data were available from the Army's Comprehensive Occupational Data Analysis Program (CODAP). These statements were first clustered by their verbs. Only 812 unique verbs appeared in 8,721 task statements from the fourteen jobs. Based on the similarity of their meanings, these 812 unique verbs were reduced to 138 categories. The range of application of these verbs was characterized by the object words from the original task statements. For example, "adjust" and "align" were verbs judged to have similar meanings and therefore placed in one category. Objects related to that verb category included cargo doors, telescopes, and brakes.

This procedure was then repeated for the entire set of jobs for which CODAP data were available. Thus some 11,000 verbs were identified from approximately 69,000 task statements. These verbs were reviewed for common meanings which resulted in 727 verb categories being identified. At this point a decision was made to reduce further the number of criterion categories by again collapsing categories of similar verbs. This process was repeated twice. The result was a three-tier hierarchy of verbs covering 69,000 task statements with the top and most general level of verbs totalling 30 categories. The level below that consisted of 136 verbs which were slightly more specific in focus. The next level down consisted of 727 verbs, again slightly more specific in meaning. These represented the essential meanings of the approximately 11,000 unique verbs in the data set of 69,000 tasks.

Since the focus was on the verbs of the task statements, the categories tended to be characteristics of human behaviors with the tasks in each category requiring similar behaviors. That is, in a category at any level of the hierarchy, tasks within the category were judged more similar in their behavior requirements than tasks between categories. Job descriptors which focus on behavior requirements are typically termed worker-oriented. However, within each category, the tasks varied widely in terms of the objects of the behavior. Consequently, each category tended to include tasks from a variety of different jobs. The thirty most general worker-oriented criterion categories appear in Appendix D.

The second approach was based on more general job descriptions of a representative sample of 111 jobs that had been previously clustered by personnel experts familiar with Army jobs. Twenty-three clusters had been identified. Criterion categories were developed by reviewing the descriptions of the jobs in these clusters to determine common job activities. Emphasis was placed on determining what a soldier in each job might be observed doing and what he or she might be trying to accomplish. The categories were constructed to connote a set of actions that typically occur together (e.g., transcribe, annotate, sort, index, file, retrieve) leading to some common objective (e.g., record and file information). Criterion categories often included reference to the use of equipment or other objects. Once criterion categories were identified for the common actions in the 23 clusters, additional categories were identified to cover unique aspects of jobs in the sample of 111. In all, 53 catego-

ries were generated. Most of these categories applied to several jobs, and most of the jobs were characterized by activities from several categories.

The emphasis of this second approach on job accomplishments and objects created criterion categories much different from those of the first approach. The first, which produced worker-oriented categories, focused on generalized human behaviors. From this perspective, tasks within each criterion were associated by virtue of the judged similarity of the human behavior requirements. The second approach produced job-oriented categories, with the tasks in each category associated by their tendencies to occur together in order to produce some identifiable job product or outcome.

For two major reasons, the decision was made to use these job-oriented criterion categories rather than the worker-oriented categories in the subsequent research of estimating predictor effectiveness. First, the job-oriented categories were more similar to the criterion specifications used in typical validation research. That is, they represent job requirements rather than behavior requirements. The second reason, which is related to the first, concerns reliability. The purpose of this research is to determine human attributes (psychological constructs) which are predictive of job performance. In other words, the links between psychological predictor constructs and job requirements are to be hypothesized by judgments about the strengths of the relationships between each of the psychological constructs and each of the job criterion categories.

These judgments would have been simplified if the criterion categories were worker-oriented with relatively homogenous behavior requirements. This simplification would occur because the process of abstracting behavior requirements would have been done prior to the criterion categories being given to the judges. That is, the abstracting of behavior requirements would have been transferred from one set of judges (the subjects of this research) to another (the designers of the research). Because these two sets of judges both represented the same population (personnel psychologists), and because greater numbers were to be involved in the task of estimating predictor-criterion relationships, it seemed more reasonable to structure the abstracting of behavior requirements so that it occurred concurrently with, and was aided by, the process of judging the relationships between predictor constructs and criterion categories. Thus, the 53 job-oriented criterion categories were chosen to represent the job task domain of job performance. Their names and definitions are shown in Table 2

Initial Training Categories. Two sources of information were used to identify appropriate training performance variables: First, archival records of soldiers' performance in training were examined, and second, interviews with trainers were conducted. This information was obtained for eight military occupational specialties (MOS). These were Radio/Teletype Operator, MANPADS Crewman, Light Vehicle/Power Generation Mechanic, Motor Transport Operator, Food Service Specialist, M60 and M1 Armor Crew, Administrative Specialist, and Unit Supply Specialist. These specialties represented a heterogeneous group with respect to type of work and were, for the most part, high density MOS.

Five or six trainers were interviewed for each MOS. The format of the

interview was a modified critical incidents approach. Trainers were asked "what things do trainees do that tell you they are good (or bad) trainees?" Generally, trainers responded with fairly broad, trait-like answers and appropriate follow-up questions were used to obtain more specific, behaviorally-oriented information.

The review of archival records was intended to identify the type of measures used to evaluate training performance, since the content was, obviously, specific and unique to each MOS.

After conducting the interviews and examining the archives, we pooled and categorized the information from both sources. We found much overlap across MOS in the way training performance was evaluated. Furthermore, we could not include content-specific variables since this would require several hundred training performance variables (one for each MOS, at least) nor did we wish to do so, since the task or MOS specific performance variance was covered elsewhere (see section entitled Specific Job Task Criterion Categories).

In the end, we decided that four variables adequately represented training performance. Their names and definitions are shown in Table 3.

Generalized Army Effectiveness Categories. The identification of variables representing generalized Army effectiveness was carried out in three steps. First, a preliminary conceptual model was developed based on relevant theory and empirical findings. Second, empirical research using the inductive behavioral analysis methods was carried out to verify and modify the preliminary model. Finally, several criterion variables that are common across all MOS but not behavioral in nature were added to the final list. We briefly summarize those steps here. A more complete description of this research can be found in a paper presented by Borman, Motowidlo, and Hanser (1983) at last year's APA convention.

The preliminary model revolved around three concepts: organizational commitment, organizational socialization, and morale. Each of these was thought to contribute to generalized Army effectiveness.

The concept of organizational commitment (Porter, Steers, Mowday, & Boulian, 1974; Steers, 1977) refers to the strength of a person's identification with and involvement in the organization and incorporates three kinds of attitudinal and cognitive elements: acceptance and internalization of organizational values and goals, motivation to exert effort toward the accomplishment of organizational objectives, and firm intentions of staying in the organization.

Organization socialization is defined as the process through which an individual acquires the knowledge and skills necessary to assume an organizational role (Van Maanen & Schein, 1979). Some of this process is job specific, but much is not. Thus, generalized Army effectiveness might reasonably be regarded as due, in part, to the degree of successful socialization to the Army in general.

Morale has traditionally been regarded as an extremely important element

in military organizations. This concept is multifaceted and seems to involve feelings of determination to overcome obstacles, confidence about the likelihood of success, exaltation of ideals, optimism even in the face of severe adversity, courage, discipline, and group cohesiveness (Motowidlo, Dowell, Hopp, Borman, Johnson, & Dunnette, 1976). Other reports (Borman, Johnson, Motowidlo, & Dunnette, 1975; Motowidlo and Borman, 1977) found that the following dimensions efficiently describe behavioral expressions of morale among soldiers: community relations; teamwork and cooperation; reactions to adversity; superior-subordinate relations; performance and effort on the job; bearing, appearance, marching, and military courtesy; pride in unit, Army, and country; and use of time during off-duty hours. Because morale seems to figure so prominently as a determinant of unit effectiveness, behavioral dimensions like these may also, in part, represent important elements of generalized Army effectiveness. The preliminary model is shown in Figure 2, reproduced from Borman et al. (1983). Definitions of the fifteen dimensions shown under Determination, Allegiance, and Teamwork are also reproduced from that report and shown in Appendix E.

Behavioral analysis workshops were carried out in order to verify and extend this model. Very briefly described, these workshops involve asking persons knowledgable about a job to generate behavioral examples of effective and ineffective performance in all aspects of the job. These examples are then content-analyzed and performance categories are formed. Army NCOs and officers participated in such workshops and generated several hundred examples of general soldier effectiveness. These examples were content-analyzed by Project A staff and the resulting categories were compared to the dimensions in the preliminary model. There was considerable overlap, but some modifications were made to the preliminary model dimensions. After making these modifications, nine behavioral dimensions were named and defined. These are shown in Table 4.

In the final step, six more criterion variables were added. They are named and defined in Table 5. The first two, "Survive in the field," and "Maintain physical fitness," were added because they represent tasks that all soldiers are expected to be able to perform but did not emerge elsewhere. The last four listed are all important "outcome" criterion variables. That is, they represent outcomes of individual behavior that have negative or positive value to the Army, but the outcomes could occur because of a variety of individual behaviors.

Subjects

The experts were 35 industrial, measurement, or differential psychologists with experience and knowledge in personnel selection research and/or applications. Each expert was an employee of or consultant to one of the four organizations involved in Project A: U.S. Army Research Institute, Personnel Decisions Research Institute, Human Resources Research Institute, American Institutes for Research. Not all of the employees were directly involved with Project A although all of the consultants were.

Instructions and Procedures

A copy of the instructions may be found at Appendix F. A brief summary of these instructions is given here. First, each judge was provided with information about the concept of "true validity," criterion-related validity corrected for such artifacts as range restriction and unreliability, and unaffected by variation in sample sizes. Judges were asked to make estimates of the level of true validity rather than of estimated validity, on a nine point scale. A rating of "1" mean meant a true validity in the range of .00 to .10; "2", .11 - .20; and so forth, to "9", .81 - .90.

Second, descriptions of the 53 predictor variables were placed into three groups, Groups A, B, and C, two groups of 18 and one of 17. The 72 criterion descriptions were in one group. Each rater was encouraged to skim the materials for a few predictors and for all the criteria before beginning.

Third, each judge estimated the validity of each predictor for each criterion. The order of the predictor groups (A, B, C) were counterbalanced across judges, such that about one-third of the 35 judges began with Group A (Predictors 1 - 18), another one-third with Group B (Predictors 19 - 36), and the rest with Group C (Predictors 37 -53). Judgments of predictors were to proceed in numerical order (1 - 53; 19 - 53, 1 - 18; 37 - 53, 1 - 36).

Ratings were made on separate Judgment Record Sheets. Prior to making any judgments about a predictor, the expert was to read the descriptive information and review the examples given to measure it. Judgments were to be made about the predictor as a construct, not about the variable as measured by any specific measurement instrument. Judges were then to read the description of the first criterion and to estimate the validity of that predictor for that criterion. Judgments could be either positive or negative; positive signs were not to be entered. The judges were then to read the description of the second criterion and rate the validity of the same predictor for that criterion. The validities of the first predictor variable for all 72 criteria were to be estimated before moving to the next predictor.

When complete, all materials were packaged and returned to the second author. The average amount of time taken to perform the task was twelve hours; all judges completed the task during the first week of October, 1983.

Analyses

Reliability. A two-way analysis of variance (53 Predictors by 72 Criteria, repeated across 35 Judges) was used to estimate reliability. Reliability estimates are calculated with the formula:

Marginal reliability (across all levels of a single dimension) =

$$\frac{MS_{effect} - MS_{error}}{MS_{effect}}$$

Individual reliability (average for each level of the dimension) =

$$\frac{\text{MSeffect} - \text{MSerror}}{\text{MSeffect} + \text{df(MSerror)}}$$

Reliability estimates were made for the marginal and individual reliabilities across predictors, criteria, and predictors by criteria.

Descriptive Statistics. Means and variances for each cell in the 53 x 72 predictor-criterion matrix were calculated. Initial results indicated some aberration with one variable, Field Dependence/Independence (Predictor # 11). Closer inspection of the data indicated that not all of the judges had rated this variable in the same direction. The majority appeared to assume that high scores meant field dependence while 12 judges appeared to assume that high scores meant field independence. The ratings for this predictor were reversed for these 12 judges before analyses continued. (This correction would not affect the reliability calculations.)

For assistance in interpeting later findings, ordered statistics for each predictor and criterion were prepared. That is, for each predictor, all criteria with average estimated validaties of .20 or higher were listed in order of decreasing size. The same procedure was followed for each criterion: All predictors with average estimated validities of .20 or higher were listed, in order of decreasing size.

Factor and cluster analyses. The matrix of mean ratings was factor analyzed (principal components, varimax rotation) both by columns (predictors) and by rows (criteria). The most reasonable solutions were selected. In combination with the descriptive statistics, the factors for each of the two analyses were further subdivided into clusters. The primary data used in such clustering were the patterns or profiles of factor weights evidenced by a given variable.

Estimated validities for predictor-criterion combinations. For both the factor and the cluster analyses, matrices were developed to display the mean estimated validity for each predictor-criterion combination, along with the standard deviation of this mean across variables. Available for comparisons were summary tables of empircial criterion-related validity coefficients from prior research.

Results

Reliability

The results of the reliability analysis of variance may be found at Appendix G. For predictors, the overall reliability was .974 while for individual raters the average reliability was .518. For criteria, the overall figure was .988 while for individual raters it was .709. The reliability of cell means across all raters was .961 while the average value for individual raters was .411. These are satisfactory statisics. Subsequent analyses were performed on the cell means having the reliability of .96.

Descriptive Statistics

The mean estimates of validity across 53 predictors and 72 criteria are displayed in Table 6. The variance (across judges) of these estimates may be found at Appendix H. The ordered criterion means, for each predictor, may be found at Appendix I, while the ordered predictor means, for each criterion, may be found at Appendix J.

Factor and Cluster Analyses

Predictors. Solutions with two through 24 factors were calculated; eigenvalues diminished below 1.0 after nine or ten factors. No more than eight factors were interpretable. The communalities and factor loadings for the solutions with two through 13 factors may be found at Appendix K. The nine factor solution was selected as most reasonable and is displayed as Table 7. The eight interpretable factors were as follows: I: Cognitive Abilities; II: Psychomotor; III: Motivation/Stability; IV: Visualization/Spatial; V: Social Skills; VI: Vigor; VII: Information Processing; VIII: Mechanical.

These eight factors appeared to be composed of 21 clusters, based on the loadings of each predictor variable on all nine factors. These are displayed in Table 8. Inspection of the profiles clarifies the meanings both of the factors and the clusters, as follows.

The eight predictor factors divide the predictor domain into reasonable-appearing parts. The first five refer to abilities and skills in the cognitive, perceptual, and psychomotor areas while the last three refer to traits or predispositions, in the noncognitive area. Most of the representative measures of the constructs defining the first five factors are of maximal performance while most of the representative measures of the last three factors are of typical performance, with the exception of the interest variables. Three of the interest constructs were more related to the abilities area while the remaining three were more related to the traits area. These first three (Realistic, Investigative, Artistic) also appear to refer more to things while the second three (Social, Enterprising, Conventional) refer more to people.

The first four factors, which include 11 clusters of 29 predictor constructs or variables, are cognitive-perceptual in nature. The first factor, labeled "Cognitive Abilities", includes seven clusters, five of which appear to consist of more traditional mental test variables: Verbal Ability/General Intelligence, Reasoning, Number Ability, Memory, Closure. The Perceptual Speed and Accuracy cluster is linked to measures having a long history of inclusion in traditional mental tests. The seventh cluster, Investigative Interests, refers to no cognitive test at all but does tap interest in things intellectual, the abilities for which are evaluated in this factor.

The second factor, Visualization/Spatial, consists of only one cluster but includes six constructs which have some history of assessment of spatial ability. Two of the clusters from the Cognitive Abilities factor, Reasoning and Closure, have some affinity to this second factor, as may be seen in the factor analysis data. This may be due to the setting tasks used to illustrate the assessment of the constructs, which are to solve problems of a visual and

nonverbal nature. The third factor, Information Processing, also consists of only one cluster, with the three constructs referring more directly to cognitive-perceptual functioning rather than accumulated knowledge and/or structure.

The fourth factor, Mechanical, includes two clusters, one of which consists only of the construct of Mechanical Comprehension while the other is, again, an interest cluster consisting of a positive loading for Realistic Interests and negative loading for Artistic Interests. Some (Humphreys, personal communication, 1984) suggest that tests of technical knowledge are legitimate indices of interest in technical areas. Our experts appear to agree.

The fifth factor, Psychomotor, consists of three clusters which include the nine psychomotor constructs. The first cluster, Steadiness/Precision, refers to aiming and tracking tasks, where the target may move steadily or erratically. The second cluster, Coordination, indexes the large-scale complexity of the response required in a psychomotor task while the third factor, Dexterity, appears to index the small-scale complexity of responses.

The remaining three factors, noncognitive in character, refer more to interpersonal activities. The Social Skills factor consists of two clusters. The first, Sociability, refers to a general interest in people while the second, Enterprising Interests, refers to a more specific interest in working successfully with people. The seventh factor is called "Vigor" as it includes two clusters which both refer to general activity level. The first, Athletic Abilities/Energy, includes two constructs which point towards a physical perspective while the second cluster, Dominance/Self-Esteem, points towards a psychological perspective. The eighth and last factor, Motivation/Stability, includes three clusters or facets. The first, Traditional Values, includes both temperament measures and interest scales, and refers to being rule-abiding and a good citizen. The second cluster, Work Orientation, refers to temperament measures which index attitudes towards the individual vis a vis his/her efforts in the world. The third cluster, Cooperation/Stability, appears to refer to skill in getting along with people, including getting along with oneself in a healthy manner.

Criteria. Solutions with two through 24 factors were calculated; eigenvalues diminished below 1.0 after eight or nine factors, for all solutions. The communalities and factor loadings for solutions with two to eleven factors may be found at Appendix L. Only five factors were interpretable. The six factor solution was selected as most reasonable and is displayed as Table 9. The five factors were as follows: I: Technical Skills; II: Commitment/Initiative; III: Personnel Interaction; IV: Combat; V: Clerical/Data.

These five factors appeared to be composed of 16 clusters, based on the loading of each criterion variable on all factors. These are displayed in Table 10. Inspection of these profiles clarifies the meanings both of the clusters and factors, as follows.

The criterion space also appears to be reasonably divided by the factor analyses, with a five-factor solution providing the best solution here. Four of the five factors refer to job specific performance while the remaining

factor references most of the generalized or Army-wide criteria. The first two of the specific performance factors emphasize cognitive and perceptual abilities while the third and fourth also emphasize, respectively, physical/psychomotor and personal interaction. There was no factor specific to training: The training constructs were divided among one of the specific performance factors and the Army-wide factor.

The first and largest factor, Technical Skills, includes nine clusters and 35 criterion constructs. One of these clusters, Training Performance, includes three of the training criterion variables, those variables referring to products rather than attitudes. The remaining clusters refer to different types of job performance in Army enlisted occupations which include different kinds of working with different kinds of equipment, but primarily with cognitive and perceptual abilities. There are no special physical or psychomotor demands although few of the constructs refer to much of the traditional desk job attributes. The first two clusters refer directly to dealing with equipment, the first being Inspect/Troubleshoot with the second being Repair/Install. The fifth cluster refers to operating equipment but apparently from a (comparatively) stable console rather than in a moving unit. Certainly the operated equipment does not include the function of moving itself, a function which is important in the third factor. The third cluster of the first factor, Construction/Repair, incorporates building and maintaining structures with, it can be assumed, appropriate equipment. The fourth cluster, Parachute Preparation/Field Placement of Equipment, and the sixth cluster, Battlefield Perception/Planning, incorporate the preparatory decision making and planning activities involved in combat actions. The fourth cluster includes constructs referring to more detailed procedural guidance than does the sixth, which also includes more of the actual battlefield activity. (Some would argue that the latter rarely goes according to procedures.) The seventh cluster, Food Prepartion/Procedures, is somewhat of a catchall but includes constructs which appear to reference more detailed procedural guidance than do the constructs in the fourth cluster, as well as with lower or less direct penalty for inadequate or untimely performance. The ninth and final cluster, Air Traffic Control, is a singleton but appears to be placed within the correct factor: Cognitive-perceptual skill requirements for dealing with equipment.

The second factor, Clerical/Data, refers to criterion variables requiring information transmission rather than dealing with equipment, as in the first factor. The first cluster here, Clerical, includes the more typical and less demanding information handling activities while the second cluster, Translate/Decode Data, implies greater cognitive demands.

The third factor, Combat, incorporates physical and psychomotor abilities as well as cognitive. The first cluster, Physical Combat Tasks, emphasizes the physical and interpersonal aspects of combat survival while the second cluster, Operate Heavy Artillery and Vehicles, references the cognitive and psyhchomotor abilities required to operate the heavy machinery used in the modern battlefield.

The fourth factor, Personal Interaction, consists of only one cluster but of those seven constructs which require effective interpersonal skills for successful performance. It is not simply interest in people, it is competence

in different aspects of social interaction and communication.

The fifth and last factor, Commitment/Initiative, includes two clusters of criterion constructs which are applicable to every Army soldier. The first, Commitment, is fairly broad and references many aspects of being a good soldier, obeying the rules and being satisfied with one's life as a soldier. The second cluster, Initiative, refers to how much effort the soldier makes to be this good soldier. The remaining training construct is included in this cluster as it references effort directly.

Estimated validities. The mean values for various predictor-criterion group combinations summarize the estimated ratings. Table 11 contains a 5 x 8 matrix, showing the estimated validities for each of the five criterion factors of each of the eight predictor factors. Table 12 presents the same statistics for the 16 criterion clusters by the 21 predictor clusters.

The data described here are estimated validity coefficients. How do they compare with empirically derived values? Vineberg and Joyner (1982) summarized the research on the prediction of military job performance. Table 13 summarizes these statistics, a variety of predictors for a variety of criteria.

Comparing estimated validities (Tables 11 and 12) with actual validities (Table 13) is awkward because the criterion spaces are not divided the same way. For the observational data Vineberg and Joyner subdivided the criterion space by method of criterion (Job knowledge, Task Performance) while the estimations were based on the type of the criterion. While the actual validities can be cross-referenced to the predictors the use of different divisions of predictor and criterion space as well as minimal or absent data in some cells makes comparisons tentative at best.

What conclusions may be drawn about the estimated validity coefficents? First, none of the values is very large. Although the judges were instructed to make their estimates corrected for several possible attenuating factors, these instructions were not, apparently, completely successful. It appears that the judges still underestimated to some extent. Second, some of the estimated values are larger than others, indicating which predictor-criterion combinations might be fruitful to explore in more detail. Such exploration depends of course on the costs and benefits of establishing a formal selection and classification system, which in turn depends on the numbers of individuals who will go through the system as well as the values of different levels of performance in different occupational areas. In this report we will limit inferences to those based on the factor combinations of estimated validities (Table 11) rather than the cluster combination (Table 12).

The best single predictor factor is that of Cognitive Abilities, even for he Combat and Commitment criterion factors. These are, of course, the criteion factors least well predicted. The second predictor factor, Visualization/Spatial, is estimated to be as predictive of Technical Skills criteria as are cognitive predictors, but less predictive of the other four criterion areas. Information Processing, the third predictor factor, is estimated to be more useful in the Clerical/Data criterion area than elsewhere, but moderately

important to the other three specific criterion factors. Mechanical comprehension, the fourth predictor factor, is estimated to be important for those two criterion facotrs which emphasize working with complex and/or heavy equipment. The Psychomotor predictor factor is estimated to be most important for the Combat criterion factor although the overall estimated validity for the constructs of this factor is not high.

The three trait or predisposition predictor factors are estimated to provide different and/or additional validity. The predictor factor of Social Skills was judged to be of minimal importance except for the criterion factor of Personal Interaction. Vigor, the seventh predictor factor, was estimated to be more important for those criterion factors which included more than cognitive-perceptual abilities: Combat, Personal Interation, Commitment/Initiative. The physical component of Vigor is more important for the Combat criterion factor while the psychological component would be more important for the remaining two criterion factors, both of which require dealing with people. Finally, the eighth predictor factor, Motivation/Stability, is estimated to have moderate relationships with all criterion factors but a much stronger relationship to the fifth criterion factor, Commitment/Initiative. Indeed, this predictor factor is by far the strongest estimated antecedent of this criterion group.

Discussion

Specific concerns of this research will be mentioned first, followed by a more general discussion of the broader issues. For the predictor space, the estimates indicate a fairly definitive split into two areas: Cognitive and noncognitive, abilities and traits, data/things and people. Interest measures, according to our expert judges, may also be split into these two subdomains. Criterion-related evidence relating to this demarcation of the predictor space would be most interesting.

The initial list of predictors was incomplete. The absence of physical performance predictor constructs has been mentioned. However, there were criterion constructs which did incorporate physical performance, those constructs which helped to define the Combat criterion factor. While physical performance measures remain outside of the scope of Project A it would be important to determine how judges such as ours would have dealt with physical performance predictors, and where such predictors would have been found in an expanded predictor space. Other predictor omissions were less intentional but, in due time, no less obvious. We included only one example of a predictor construct of technical knowledge, Mechanical Comprehension. Currently operational selection and classification procedures for the Army include other measures of technical information, specifically, information about electronics, automobiles, shop, and general science. We suspect that such measures would end up in the same predictor factor, if not cluster, which we called "Mechanical" here. This would lead to a name change, perhaps, to "Technical Information."

There were few predictor measures which referenced specifically auditory skills although at least two clusters (Verbal, Information Preocessing) reference them indirectly. The history of standardized testing includes an impor-

tant yet small amount of space to measures involving audition. The equipment problems of such assesment remain formidable even in the current age of microprocessors. The assumption that many of the skills required of auditory processing are also required of visual and verbal processing may be reasonable. More information about the limitations of this assumption would be helpful.

Finally, our identification of the separate clusters of Reasoning, Closure, and Visualization/Spatial may strike some readers as arbitrary. It so struck us. It is important to remember that the data base on which our identification of independent predictor (and criterion) factors is composed, consists of judgmental data of affinity rather than observational data. If one were to develop a battery which included most if not all of the 53 predictor constructs rated here, and administered this battery to a representative sample sufficiently large for statistical generalization, it is not likely that the same predictor factors would emerge in the same fashion. To the extent that our experts based their judgments on empirical data there would, of course, be a great deal of correspondence. The cognitive-perceptual factors might be more closely interrelated; the interest constructs might well form their own factors. However, the 21 predictor clusters identified here are being used now as preliminary dimensions along which to develop a comprehensive battery. This battery will be administered to large samples of soldiers as Project A proceeds. The results of this battery administration will be compared to the results displayed here.

The criterion space of this research was developed to cover the range of performance expected from most Army enlisted personnel. Here, too, our experts distinguished between people and things but in a different way. Two of the criterion factors included variables related to social activities but one referred to social skills while the other referred to personological attributes. Perhaps one should not be surprised that the clusters of the five factors of the criterion space were so sensible, as psychologists can be quite sophisticated with post hoc explanations. Nevertheless, it was interesting that four of the five sets of criterion groups or clusters could be easily described by skill-relevant aspects such as type of equipment, use of equipment, purpose of interaction with equipment, etc. The remaining factor referred less to the "doing" of the activities referencing the first four factors and more to the quality of "being" a good soldier. Finally it is interesting that the training criterion constructs did not form a separate factor but allied themselves with other criterion variables on the basis of content. This result suggests that our experts were unwilling to perceive initial training as a unique entity. Perhaps they were assuming that training in the Army is a constant, and continues past initial formal schooling. Or, perhaps they were making their estimates based on inferences that Army training is representative of later field activity. Again, data expected later in Project A will shed some light on this. We do know that criterion-related validity coefficients for training tend to be larger than coefficients for on-the-job performance. Apparently this difference in level of relationship did not impact here on estimates of types of relationship.

The estimated relations between the predictor and criterion factors, between the predictor and criterion clusters, provide both a summary of the data as well as food for thought. That cognitive variables are most predic-

tive of all types of criteria should come as no surprise; what is more intriguing is that noncognitive predictors, particularly those indexing motivation, may also be predictive of a broad variety of behaviors as well. Thus, selection might be enhanced by broad-band temperament or disposition measures. The other predictor factors or clusters may be less effective for selection but may be very useful for differential classification. This might be especially true for the psychomotor constructs.

The potential utility of the estimated mean validities for personnel selection and classification needs several qualifications. The most important is that these data should not be overinterpreted. They are judges' estimates, not actual values of validity coefficients. While these estimates appear comparable to actual validities where such values exist, they are different in a systematic way: They are lower. Data are not currently available to develop corrections for our estimates; their relative values should be given more credence than their absolute sizes. And some of the values may be low because there is less actual data for specific predictor-criterion combinations. However, in a recent review of validity generalization, Hunter and Hunter (1984) noted that validity coefficients for military performance were lower than those for civilian performance. They cited the review of Vineberg and Joyner which is summarized here in Table 13.

Some of the incompleteness in the initial sets of predictors and criteria was planned. Physical abilities were not included among the predictors and the criterion space is restricted to that of Army entry-level enlisted occupations, equivalent to mostly blue-collar and white-collar skilled and semi-skilled trades. Other omissions were not planned but became obvious as the results were evaluated, as indicated above. While the authors judge that the 21 predictor clusters represent a delineation of the relevant predictor space for Army jobs, this delineation is not final. As Project A and related research proceeds, the delineation will be modified. Other researchers will wish to make other refinements and elaborations of both the predictor and the criterion domains. The five factor, 16 cluster partition of the criterion space does appear to be fairly complete, however, in its coverage of enlisted Army occupations.

On the other hand, this research does show that personnel experts can estimate the validity of a wide variety of predictor-criterion relationships. These estimates were made with a high degree of reliability and a reasonable amount of accuracy. More definitive information about such accuracy will be provided by the criterion-related validity research now underway in Project A. Not all of the specific predictor-criterion construct combinations will be evaluated but a large number of them will be. The amount and kind of correspondence between the estimates presented here and these validity coefficients will indicate more clearly the potential and limitations of the methods used here.

References

Borman, W. C., Johnson, P. D., Motowidlo, S. J., & Dunnette, M. D. (1975). Measuring motivation, morale and job satisfaction in Army careers. Minneapolis, MN: Personnel Decisions, Inc.

Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983). Developing a model of soldier effectiveness: A strategy and preliminary results. Alexandria, VA: U.S. Army Research Institute. Presented at the 91st Annual Convention of the Psychological Association, Anaheim, California.

Bowmas, D. A., & Heckman, R. W. (1976). Job analysis of the entry-level firefighter position. Minneapolis, MN: Personnel Decisions, Inc.

Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. Journal of Applied Psychology, 66, 274-281.

Campbell, C. H., & Black, B.A. (1982). Predicting trainability of MI crewmen. (TR 592). Alexandria, VA: U.S. Army Research Institute.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.) New directions for testing and measurement: No. 5 Measuring achievement: Progress over a decade. San Francisco: Jossey-Bass.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. In M. R. Rosenzweig & L. W. Porter (Eds.) Annual Review of Psychology; Palo Alto, CA: Annual Reviews, Inc., 30, 477-525.

Dunnette, M. D., Rosse, R. L., Houston, J. S., Hough, L. M., Toquam, J., Lammlein, S., King, K. W., Bosshardt, M. J., & Keyes, M. A. (1981). Development and validation of an industry-wide electric power plant operator selection system. Minneapolis, MN: Personnel Decisions Research Institute.

Guion, R. M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-390.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Journal of Applied Psychology, 96; 72-98.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Advanced meta-analysis: Quantitative methods for cumulating research findings across studies. Beverly Hills, CA: Sage.

Johnson, J. H., Jones, M. B., & Kennedy, R. S. (1984). Cognitive predictors of tank commander performance (TR in press) Alexandria, VA: U.S. Army Research Institute.

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first year grades in law school. Applied Psychological Measurements, 5, 281-289.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Motowidlo, S. J. & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 62, 177-183.

Motowidlo, S. J., Dowell, B. E., Hopp, M. A., Borman, W. C., Johnson, P. D., & Dunnette, M. D. (1976). Motivation, satisfaction, and morale in Army careers: A review of theory and measurement. (TR-76-A7). Alexandria, VA: U. S. Army Research Institute (NTIS No. AD-A036390).

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Peterson, N. G., & Bownas, D. A. (1982). Task structure and performance acquistion. In M. D. Dunnette & E. A. Fleishman (Eds.) Human capability assessment. New York: Lawrence Erlbaum & Associates, 49-105.

Peterson, N. G., & Houston, J. S. (1980). The prediction of correctional officer job performance: Construct validation in an employment setting. Minneapolis, MN: Personnel Decisions Research Institute.

Peterson, N. G., Houston, J. S., Bosshardt, M. D., & Dunnette, M. D. (1977). A study of the correctional officer job at Marion Correctional Institution, Ohio: Development of selection procedures, training recommendations and an exit information program. Minneapolis, MN: Personnel Decisions Research Institute.

Peterson, N. G., Houston, J. S., & Rosse, R. L. (1984). The LOMA job effectiveness prediction system, technical report #4: Validity analyses. Atlanta, GA: Life Office Management Association.

Porter, L. W., Steers, R. M., Mowday, R. T., & Boulian, P. V. (1974). Organizational commitment, job satisfaction, and turnover among psychiatric technicians. Journal of Applied Psychology, 59, 603-609.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalization results for two job groups in the petroleum industry. Journal of Applied Psychology, 66, 261-273.

Schmidt, F. L., Hunter, J. E., Croll, P. R., & McKenzie, R. C. (1983). Estimation of employment test validities by expert judgment. Journal of Applied Psychology, 68, 590-601.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 66, 166-185.

Steers, R. M. (1977). Antecedents and outcomes of organizational commitment. Administrative Science Quarterly, 22, 46-56.

U.S. Army Research Institute (1983). Improving the selection, classification, and utilization of Army enlisted personnel. Project A: Research Plan. (Research Report 1332). Alexandria, VA: Author.

Van Maanen, J., & Schein, E. H. (1979). Toward a theory of organizational socialization. In B. M. Staw (Ed.) Research in organizational behaviors (Volume I). Greenwich, CT: JAI Press.

Vineberg, R., & Joyner, J. N. (1982). Prediction of job performance: Review of military studies. (TR 82-37). San Diego, CA: Navy Personnel Research and Development Center (DTIC No.AD A113208).

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples and criteria. Journal of Applied Psychology, 52, 372-276.

Acknowledgements

We would like to thank the many individuals who assisted in developing the predictor and criterion constructs. These include Carol Manning from the American Institute for Research, Charlotte Campbell of the Human Resources Research Organization and Walter Borman and Marvin Dunnette of the Personnel Decisions Research Institute. We would also like to thank our expert raters, from the above organizations and the U.S. Army Research Institute.

1. Discriminability--extent to which the measure has sufficient score range and variance, i.e., does not suffer from ceiling and floor effects with respect to the applicant population.

2. Reliability--degree of reliability as measured by traditional psychometric methods such as test-retest, internal consistency, or parallel forms reliability.

3. Group Score Differences (Differential Impact)--extent to which there are mean and variance differences in scores across groups defined by age, sex, race, or ethnic groups; a high score indicates little or no mean differences across these groups.

4. Consistency, Robustness of Administration and Scoring--extent to which administration and scoring is standardized, ease of administration and scoring, consistency of administration and scoring across administrators and locations.

5. Generality--extent to which predictor measures a fairly general or broad ability or construct.

6. Criterion-Related Validity--the level of correlation of the predictor with measures of job performance, training performance and turnover/attrition.

7. Construct Validity--the amount of evidence existing to support the predictor as a measure of a distinct construct (correlational research, experimental research, etc.).

8. Face Validity/Applicant Acceptance--extent to which the appearance and administration methods of the predictor enhance or detract from its plausibility or acceptability to laymen as an appropriate test for the Army.

9. Differential Validity--existence of significantly different criterion-related validity coefficients between groups of legal or societal concern (race, sex, age); a high score indicates little or no differences in validity for these groups.

10. Test Fairness--degree to which slopes, intercepts, and standard errors of estimate differ across groups of legal or societal concern (race, sex, age) when predictor scores are regressed on important criteria (job performance, turnover, training); a high score indicates fairness (little or no differences in slopes, intercepts, and standard errors of estimate).

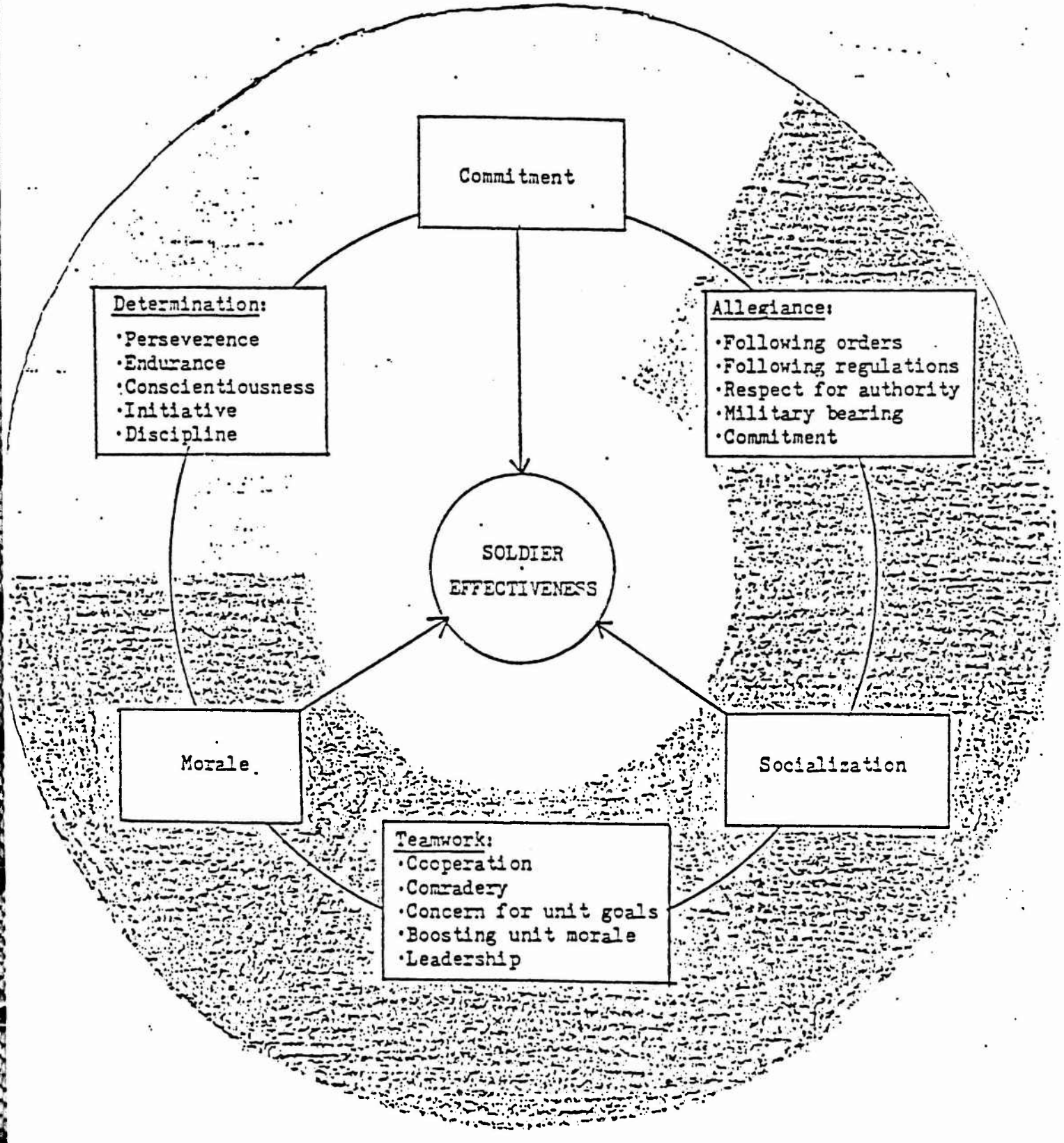
11. Usefulness of Classification--extent to which the measure or predictor will be useful in classifying persons into different specialties.

12. Overall Usefulness for Predicting Army Criteria--extent to which predictor is likely to contribute to the overall or individual prediction of criteria important to the Army (e.g., AWOL, drug use, attrition, unsuitability, job performance, and training).

FIGURE 1. Factors Used to Evaluate Predictor Measures

Figure 2

A Preliminary Model of Soldier Effectiveness



List of 53 Predictor Variables Identified For Inclusion in the Expert Judgment Task

PREDICTOR VARIABLES

Construct Name	Definition
Verbal Comprehension	Measures knowledge of the meaning of words and their relationships to each other.
Numerical Computation	Measures speed and accuracy in performing simple arithmetic operations, i.e., addition, subtraction, multiplication and division.
Use of Formulations and Number Problems	Measures the ability to correctly use algebraic formulae to solve number problems.
Word Problems	Measures the ability to select and organize relevant information to correctly solve mathematical word problems.
Reading Comprehension	Measures the ability to read and understand written material.
Two-Dimensional Mental Rotation	Measures the ability to identify a two-dimensional figure when seen at different angular orientations within the picture plane.
Three-Dimensional Mental Rotation	Measures the ability to identify a three-dimensional object, projected on a two-dimensional plane, when seen at different angular orientations either within the picture plane or about the axis in depth.
Inductive Reasoning: Concept Formation	Measures the ability to discover a rule or principle and apply it in solving a problem.
Spatial Visualization	Measures the ability to mentally manipulate the components of a two- or three-dimensional figure into other arrangements.
Deductive Logic	Ability to use logic and judgment in drawing conclusions from available information. Given a test of facts and a set of conclusions, deductive logic refers to the ability to determine whether the conclusions flow logically from the facts.
Field Dependence	Ability to find a simple form when it is hidden in a complex pattern. Given a visual percept or configuration, field dependence (or independence, more accurately) refers to the ability to hold it in mind so as to disembed it from other well-defined perceptual material.
Perceptual Speed and Accuracy	Ability to perceive visual information quickly and accurately and to perform simple processing tasks with it (e.g., comparisons). This requires the ability to make rapid scanning movements without being distracted by irrelevant visual stimuli, and also measures memory, working speed, and sometimes eye-hand coordination.

Construct Name	Definition
Mechanical Comprehension	Ability to learn, comprehend, and reason with mechanical terms. More specifically, this is the ability to perceive and understand the relationship of physical forces and mechanical elements in practical situations.
Rote Memory	Measures the ability to recall previously learned but unrelated item pairs.
Place Memory (Visual Memory)	Ability to remember the configuration, location, and orientation of figural material.
Ideational Fluency	Ability to rapidly generate ideas about a given topic or exemplars of a class of objects.
Follow Directions	Measures ability to follow simple and complex directions.
Analogical Reasoning	Measures the ability to identify the underlying principles governing relationships between pairs of objects.
Figural Reasoning	Measures ability to generate and apply hypotheses about principles governing the relationship among several figures.
Spatial Scanning	Measures the ability to visually survey a complex field to find a particular configuration representing a pathway through the field.
Omnibus Measures of Intelligence/Aptitude	Measures general mental ability or general aptitude.
Word Fluency	Ability to rapidly think of words.
Verbal and Figural Closure	Measures ability to identify objects or words given sketchy or partial information.
Processing Efficiency	Speed of reactions to simple stimuli.
Selective Attention	This is the ability to attend to a target stimulus when presented with two or more stimuli simultaneously.
Time-Sharing	Time-sharing is the ability to perform two or more tasks simultaneously.
Multilimb Coordination	Multilimb coordination is the ability to coordinate the simultaneous movement of two or more limbs. This ability is general to tasks requiring coordination of any two limbs (e.g., two hands, two feet, one foot and one hand). It is most common to tasks where the body is at rest (e.g., seated or standing) while two or more limbs are in motion.

TABLE 1 Continued

Construct Name	Definition
Control Precision	Control precision is the ability to make fine, highly controlled (but not over-controlled) muscular movements necessary to adjust or position a machine or equipment control mechanism. This ability is general to tasks requiring motor adjustments in response to a stimulus whose speed and/or direction of movement are perfectly predictable. This ability is critical in situations where the motor adjustments must be both rapid and precise. The ability extends to arm-hand movements as well as to leg movements.
Rate Control	Rate control is the ability to make continuous anticipatory muscular movements necessary to adjust or position a machine or equipment control mechanism. This ability is general to tasks requiring motor adjustments or movements in response to a moving stimulus which is changing speed and/or direction in a random or unpredictable manner. The ability applies to compensatory tracking of the stimulus as well as following pursuit of the stimulus.
Manual Dexterity	Manual dexterity is the ability to make skillful, coordinated movements of the hand or the arm and hand. This ability most typically applies to tasks involving manipulation of moderately large objects (e.g., blocks, pencils, etc.) under speeded conditions.
Finger Dexterity	Finger dexterity is the ability to make skillful, coordinated, highly controlled movements of the fingers. This ability applies primarily to tasks involving manipulation of objects with the fingers.
Track Tracing Test	Designed to measure arm-hand steadiness.
Wrist-Finger Speed	The ability to carry out very rapid, discrete movements of the fingers, hands, and wrists. This ability applies primarily to tasks in which the accuracy of the movement is <u>not</u> a major concern. This ability is determined entirely by the speed with which the movement is carried out.
Aiming	The ability to make very precise, accurate hand movements under highly-speeded conditions. This ability is dependent upon very precise eye-hand coordination.
Speed of Arm Movement	This ability involves the speed with which discrete arm movements can be made. The ability deals with the speed with which the movement can be carried out <u>after</u> it has been initiated.

TABLE 1 Continued

Construct Name	Definition
Involvement in Athletics and Physical Conditioning	Frequency and degree of participation in sports, exercise and physical activity. Individuals high on this dimension actively participate in individual and team sports and/or exercise vigorously several times per week.
Energy Level	Characteristic amount of energy and enthusiasm. The person high in energy level is enthusiastic, active, vital, optimistic, cheerful, zesty, and has the energy to get things done.
Cooperativeness	Characteristic degree of pleasantness versus unpleasantness exhibited in interpersonal relations. The highly cooperative person is pleasant, tolerant, tactful, helpful, not defensive, and generally easy to get along with. His/her participation in a group adds cohesiveness.
Sociability	Outgoingness. The person high in sociability is talkative, relates easily to others, is responsive and expressive in social environments, readily becomes involved in group activities, and has many relationships.
Traditional Values	Personal views in areas such as authority, discipline, social change, and religious commitment. The person with traditional values accepts authority and the value of discipline, is likely to be religious, values propriety, and is conventional, conservative, and resistant to social change.
Dominance	Tendency to seek and enjoy positions of leadership and influence over others. The highly dominant person is forceful and persuasive at those times when adopting such characteristics is appropriate.
Self-esteem	Degree of confidence in one's abilities. A person with high self-esteem feels largely successful in past undertakings and expects to succeed in future undertakings.
Conscientiousness	Characteristic amount of behavioral self-control. The highly conscientious person is dependable, planful, well organized, and disciplined. This person prefers order and thinks before acting.
Locus of Control	Characteristic belief in the amount of control people have over rewards and punishments. The person with an internal locus of control expects that there are consequences associated with behavior and that people control what happens to them by what they do. The person with an external locus of control believes that what happens to people is beyond their personal control.

TABLE 1 Continued

Construct Name	Definition
Emotional Stability	Characteristic degree of stability vs. reactivity of emotions. The emotionally stable person is generally calm, displays an even mood, and is not overly distraught by stressful situations. He/she thinks clearly and maintains composure and rationality in situations of actual or perceived stress.
Nondelinquency	Amount of respect for laws and regulations as manifested in attitudes and behavior. The nondelinquent person is honest, trustworthy, wholesome, and law-abiding. Such persons will have histories devoid of trouble with schools and legal agencies.
Work Orientation	Tendency to strive for competence in one's work. The work-oriented person works hard, sets high standards, tries to do a good job, endorses the work ethic, and concentrates on and persists in completion of the task at hand.
Realistic Interests	Preference for concrete and tangible activities, characteristics and tasks. Persons with realistic interests enjoy, and are skilled in, the manipulation of tools, machines and animals but find social and educational activities and situations aversive.
Investigative Interests	Preference for scholarly, intellectual, and scientifice activities and tasks. Persons with investigative interests enjoy analytical, ambiguous, and independent tasks but dislike leadership and persuasive activities.
Enterprising Interests	Preference for persuasive, assertive and leadership activities and tasks. Persons with enterprising interests may be characterized as ambitious, dominant, sociable and self-confident.
Artistic Interests	Preferences for unstructured, expressive and ambiguous activities and tasks. Persons with artistic interests may be characterized as intuitive, impulsive, creative and non-conforming.
Social Interests	Preferences for social, helping and teaching activities and tasks. Persons with social interests may be characterized as responsible, idealistic, and humanistic.
Conventional Interests	Preferences for well-ordered, systematic and practical activities and tasks. Persons with conventional interests may be characterized as conforming, unimaginative, efficient, and calm.

CRITERION CONSTRUCTS

1. Inspect mechanical systems--test, measure, and/or use diagnostic equipment as well as visual, aural and tactile senses, in conjunction with technical information, to compare the operating status of mechanical equipment (e.g., engines, transmissions, machineguns) and mechanical components (e.g., bearings in an electrical generator) to standards of operating efficiency, and to identify malfunctions.

   Actions may include: analyze, read, operate

2. Troubleshoot mechanical systems--use test, measuring, and diagnostic equipment, in conjunction with technical information, to determine the cause of malfunctions in mechanical equipment (e.g., engines, transmissions, machineguns) and mechanical components (e.g., bearings in an electrical generator).

   Actions may include: analyze, read, calculate

3. Repair mechanical systems--perform corrective actions on previously diagnosed malfunctions of mechanical equipment or mechanical components using appropriate tools (e.g., wrenches, screwdrivers, gauges, hammers) in conjunction with technical information.

   Actions may include: adjust, assemble/disassemble, install, fix, read, work metal

4. Inspect fluid systems--use test, measuring, and diagnostic equipment, as well as visual, aural and tactile senses, in conjunction with technical information, to determine the operating status of fluid systems (e.g., hydraulic, refrigeration, engine cooling, compressed air) in comparison to standards of operating efficiency, and to identify malfunctions.

   Actions may include: analyze, read, operate

5. Troubleshoot fluid systems--use test, measuring and diagnostic equipment, in conjunction with technical information, to determine the cause of malfunctions in fluid systems (e.g., hydraulic, refrigeration, engine cooling, compressed air).

   Actions may include: analyze, read, calculate

6. Repair fluid systems--perform corrective actions on previously diagnosed malfunctions of fluid systems using appropriate tools (e.g., wrenches, pressure gauges, soldering equipment) in conjunction with technical information.

   Actions may include: adjust, assemble/disassemble, install, fix, read

7. Inspect electrical systems--use test, measuring, and diagnostic equipment, as well as visual, aural and tactile senses, in conjunction with technical information, to determine the operating status of electrical systems (e.g., generators, wiring harnesses, switches, relays, circuit breakers, motors, lights) in comparison to standards of operating efficiency and to identify malfunctions.

   Actions may include: Analyze, read, operate

8. Troubleshoot electrical systems--use test, measuring and diagnostic equipment, in conjunction with technical information, to determine the cause of malfunctions in electrical systems (e.g., generators, wiring harnesses, switches, relays, circuit breakers, motors, lights).

   Actions may include: analyze, read, calculate

9. Repair electrical systems--perform corrective actions on previously diagnosed malfunctions of electrical systems and electrical components using appropriate tools (e.g., pliers, wire strippers, soldering irons) in conjunction with technical information.

   Actions may include: adjust, assemble/disassemble, install, fix, read

10. Inspect electronic systems--use test, measuring and diagnostic equipment, and to a limited extent, visual, aural, and tactile senses, in conjunction with technical information, to compare the operating status of electronic systems (e.g., communications equipment, radar, missile and tank ballistics controls) to standards of operating efficiency and to identify malfunctions.

    Actions may include: analyze, read, operate

11. Troubleshoot electronic systems--use test, measuring, and diagnostic equipment, in conjunction with technical information, to determine the cause or location of malfunctions in electronics systems (e.g., communication equipment, radar, missile and tank ballistics controls).

    Actions may include: analyze, read, calculate

12. Repair electronic systems--perform corrective actions on previously diagnosed malfunction of electronic systems and electronic components using appropriate tools (e.g., test sets, screwdrivers, pliers, soldering guns) in conjunction with technical information.

    Actions may include: adjust, assemble/disassemble, install, fix, read

13. Repair metal--perform corrective actions (e.g., bend, cut, drill, saw, weld, rivet, hammer, grind, solder, paint) to refabricate metal structures.

    Actions may include: calculate, assemble/disassemble, fix, construct, read, work metal

14. Repair plastic and fiberglass structures--perform corrective actions (e.g., measure, cut, saw, drill, sand, fill, paint, glue) to refabricate plastic and fiberglass structures.

    Actions may include: calculate, assemble/disassemble, fix, construct, read

15. Construct wooden buildings and other structures--perform carpentry activities (e.g., measure, saw, nail, plane) to frame, sheath and roof buildings, or to erect trestles, bridges, piers, etc.

    Actions may include: calculate, assemble/disassemble, install, construct, read

16. Construct masonry buildings and structures--perform masonry activities (e.g., measure, lay brick, pour concrete) to construct walls, columns, field fortifications, etc.

    Actions may include: construct, calculate, assemble/disassemble, read

17. Prepare parachutes--inspect cargo and personnel parachutes, repair or replace faulty parachute components, and prepare (i.e., pack) parachute for future air drop.

    Actions may include: adjust, assemble/disassemble, pack/unpack, fix, sew, read

18. Prepare equipment and supplies for air drop--fabricate and assemble platforms, cushions, and rigging to parachute supplies, equipment and vehicles; load, position and secure supplies and equipment in aircraft.

    Actions may include: adjust, assemble/disassemble, pack/unpack, construct, transport

19. Install electronic components--place and interconnect electronic and communication components and equipment (e.g., radios, antennas, telephones, teletypewriters, radar, power supplies) and check system for operation.

    Actions may include: adjust, assemble/disassemble, install, read

20. Operate electronic equipment--set and adjust the controls of electronic components to operate electronic systems (e.g., radio, radar, computer hardware, missile ballistics controls).

    Actions may include: adjust, operate

21. Send and receive radio messages--use standardized radio codes and procedures to transmit and receive information.

    Actions may include: signal, communicate, read

22. Operate keyboard device--type information using a typewriter, teletype or keypunch, or computer terminal.

    Actions may include: process, operate

23. Use maps in the field--read and interpret map symbols and identify geography features in order to locate geography features and field positions on the map, and to locate map features in the field.

    Actions may include: analyze, identify, read, calculate

24. Plan placement or use of tactical position and features--using maps and on-site inspection, identify geographic positions or areas to be used for cover and concealment or to place fortifications, mines, detectors, chemicals, etc.

    Actions may include: analyze, calculate, read

25. Place tactical equipment and materials in the field--without using heavy equipment (e.g., lifts, dozers), place mines, detectors, chemicals, camouflage or other tactical items into position on the battlefield.

    Actions may include: use weapons, maneuver, transport, install

26. Detect and identify targets--using primarily sight, with or without optical systems, locate potential targets, and identify type (e.g., tanks, troops, artillery) and threat (friend or foe); report information.

    Actions may include: communicate, analyze

27. Prepare heavy weapons for tactical use--transport, position and assemble heavy tactical weapons such as missiles, field artillery, anti-aircraft systems.

    Actions may include: adjust, assemble/disassemble, install, pack/unpack

28. Load field artillery or tank guns--manipulate breech controls and handle ammunition (stow and load) to prepare guns for firing.

    Actions may include: use weapons, pack/unpack

29. Fire heavy direct fire weapons (e.g., tank main guns, TOW missile, infantry fighting vehicle cannon)--using optical sighting systems, manipulate weapon system controls to aim, track and fire on designated targets.

    Actions may include: use weapons, operate, adjust

30. Operate fire controls of indirect fire weapons (e.g., field artillery)--using map coordinates and ballistics information determine elevation and azimuth needed for firing at designated targets; adjust weapon using fire controls.

    Actions may include: analyze, calculate, read, adjust

31. Fire individual weapons--aim, track and fire hand operated weapons such as rifles, pistols, and machineguns at designated targets.

    Actions may include: use weapons

32. Engage in bayonet and hand-to-hand combat--use offensive and defensive body maneuvers to subdue hostile individuals.

    Actions may include: maneuver, apprehend

33. Operate wheeled vehicles--use various vehicle controls to drive wheeled vehicles from point to point, generally over paved and unpaved roads, observe traffic regulations; secure cargo.

    Actions may include: maneuver, transport, operate

34. Operate track vehicles--use various vehicle controls to drive track vehicles (e.g., tanks, APCs, scout vehicles, bulldozers); steer in response to terrain features.

    Actions may include: maneuver, transport, operate

35. Operate lifting, loading and grading equipment--operate heavy equipment (e.g., fork lifts, cranes, loader, back-hoes, graders) to load, unload, or move heavy equipment, supplies, construction materials (e.g., culvert pipes, building or bridge trusses), or terrain features (e.g., earth, rock, trees).

    Actions may include: construct, operate

36. Operate power excavating equipment--use pneumatic hammers and drills, paving breakers, grinders, and backfill tampers, in the fabrication and modification of concrete, stone and earthen structures.

Actions may include: construct, operate

37. Reproduce printed materials--operate duplicating machines and offset presses to reproduce printed materials; collate and bind materials using various types of bindery equipment.

Actions may include: adjust, operate, photograph, calculate

38. Make movies and videotapes--use motion picture cameras or videotape equipment to record visual and auditory aspects of assigned subject matter to be used for intelligence analyses, training or documentation.

Actions may include: adjust, photograph

39. Draw maps and overlays--use drafting, graphics, and related techniques to prepare and revise maps, with symbols and legends, from aerial photographs.

Actions may include: analyze, process, draw

40. Write and deliver presentations--prepare scripts for formal presentation including radio and television broadcast; make oral presentations.

Actions may include: analyze, write

41. Record and file information--collect, transcribe, annotate, sort, index, file, and retrieve information (e.g., training rosters, personnel statistics, supply inventories).

Actions may include: process, dispose

42. Receive, store and issue supplies, equipment and other materials--inspect material and review paperwork upon receipt; sort, transport, and store material; issue or ship material to authorized personnel or units.

Actions may include: analyze, calculate, process, send, pack/unpack, transport

43. Prepare technical forms and documents--follow standardized procedures to prepare or complete forms and documents (e.g., personnel records and dispositions, efficiency reports, legal briefs).

Actions may include: process, write, analyze

44. Translate or decode data--use standardized coding systems and decoding rules to convert coded information to some more usable form (e.g., interpret radar information, decode Morse code, translate foreign languages).

    Actions may include: analyze

45. Analyze intelligence data--determine importance and reliability of information; integrate information to provide identification, disposition and movement of enemy forces and estimate enemy capabilities.

    Actions may include: communicate, analyze, read

46. Prepare food--prepare food and beverages according to recipes and meal plans (measure, mix, bake, etc.); inspect fresh food and staples for freshness; maintain sanitary work area.

    Actions may include: cook, read, sanitize, dispose, calculate

47. Receive clients, patients, guests--schedule, greet and give routine information to persons seeking medical, dental, legal or counseling services.

    Actions may include: administer, communicate, process

48. Interview--verbally gather information from clients, patients, witnesses, prisoners, or other persons.

    Actions may include: communicate

49. Provide medical and dental treatment--give medical attention to soldiers in the field, or medical or dental clinic, or to animals (e.g., CPR, splinting fractures, administering injections, dressing wounds).

    Actions may include: treat, sanitize, photograph

50. Select, lay-out and clean medical or dental equipment and supplies--prepare treatment areas for use by following prescribed procedures for laying-out instruments and equipment; clean equipment and area for subsequent use.

    Actions may include: sanitize, assemble/disassemble, pack/unpack, dispose

51. Perform medical laboratory procedures--conduct various types of blood tests, urinalysis, cultures, etc.

    Actions may include: sanitize, analyze, calculate, adjust

TABLE 2 Continued

52. Control individuals and crowds--apprehend suspected criminals, capture enemy soldiers, guard prisoners, participate in riot control operations, etc.

Actions may include: apprehend, communicate, administer

53. Control air traffic--coordinate departing, en route, arriving and holding aircraft by monitoring radar equipment and communicating with aircraft and other air traffic control facilities.

Actions may include: communicate, analyze, send, operate, signal

TABLE 3

Initial Training Performance Variables

1. Training progress/success—successfully completing formal training course in normal amount of time versus washing out, being reassigned, being "set back" or "recycled."

2. Effort/motivation in training—the degree of effort, motivation, and interest that a soldier puts into his/her training, as evidenced by such things as curiosity about course content, not being afraid to be "wrong" or to ask questions, taking notes, being attentive in class, studying on own time, seeking out the instructor to clarify course content.

3. Performance of theoretical, or "classroom" parts of training—learning the theoretical part of a course; performing well on quizzes, tests, and examinations given in a classroom setting that tests the acquisition of concepts, principles, facts, or other information, e.g., learning the basic food groups, understanding the principles of internal combustion, learning the nomenclature of a weapon.

4. Performance of practical, "hands-on" part of training—applying the theory or principles of a course to practical problems and situations, either during simulations, field exercises, or other "hands-on" parts of training, e.g., cooking a meal, repairing an engine, firing a weapon, etc.

TABLE 4

Nine Behavioral Dimensions of
Generalized Army Effectiveness

1. Following regulations—consistently complying with Army rules and regulations; conforming appropriately to standard procedures; following the spirit as well as the letter of military and civilian laws, regulations, written orders, etc.

2. Commitment to Army norms—adjusting successfully to Army life; displaying appropriate military appearance and bearing; showing pride in being a soldier.

3. Cooperation with supervisors—responding willingly to orders, suggestions, and other guidance from NCOs and officers; deferring appropriately to superiors' expertise and judgment and being supportive of superior officers/NCOs.

4. Cooperation with other unit members—pitching in when necessary to help other unit members with their job and mission assignments or during training; encouraging and supporting other unit members, as appropriate; showing concern for unit objectives over and above personal interests.

5. Hard work and perseverance—working hard on the job and during training; sustaining maximum effort over long periods of hard duty and on daily assignments; coping well with hardship or otherwise unpleasant conditions to continue to work toward mission completion.

6. Attention to detail—carrying out assignments carefully and thoroughly; consistently completing job and duty assignments on time or ahead of schedule; being conscientious in maintaining own and unit's equipment, and taking care to ensure that own quarters are clean and neat.

7. Initiative—willingly volunteering for assignments; performing extra necessary tasks without explicit orders; anticipating problems and taking action to prevent them.

8. Discipline—consistently concentrating on the job or duty assignment rather than being distracted by opportunities to socialize or otherwise stop working; controlling own emotions and not allowing them to interfere with performance of duty; keeping under control alcohol and other drug intake so that performance is not affected.

9. Emergent leadership—displaying good judgment in making suggestions to others in the unit regarding the job, duty assignments, etc.; appropriately taking charge when placed in a leadership position; where appropriate, persuading others in the unit to accept his/her ideas, opinions, and directions.

TABLE 5

Six General Army Effectiveness Variables

10. Survive in the field--react to direct or indirect fire; construct individual fighting position; camouflage self and equipment; use challenge and password; protect against NBC attack.

11. Maintain physical fitness--keep self at physical fitness level appropriate for state of battle readiness.

12. Disciplinary problems--having a record of disciplinary problems as reflected by AWOLS, Article 15s, civil arrests, etc.

13. Attrition--separating from the Army for "negative" reasons such as discipline or drug-related problems.

14. Reenlistment--signing on for a second tour of duty.

15. Job satisfaction/morale--being satisfied with own MOS and Army life.

Table 6

Means of Validity Estimates of 35 Experts
for 53 Predictors and 72 Criteria

	CRITERION FACTORS (MEANS)																			
PREDICTORS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.26	.31	.25	.26	.31	.26	.27	.31	.26	.27	.32	.26	.17	.18	.18	.17	.18	.18	.27	.25
2	.32	.36	.29	.33	.36	.28	.34	.37	.29	.36	.38	.30	.24	.24	.27	.26	.18	.20	.31	.31
3	.26	.31	.22	.26	.31	.22	.27	.32	.24	.29	.35	.26	.18	.18	.22	.20	.14	.15	.25	.24
4	.30	.37	.26	.30	.38	.27	.31	.38	.28	.32	.40	.29	.20	.19	.24	.21	.16	.17	.28	.27
5	.29	.33	.27	.29	.33	.27	.29	.33	.27	.30	.34	.28	.17	.17	.20	.18	.17	.17	.27	.26
6	.27	.28	.26	.26	.26	.23	.25	.26	.23	.23	.24	.22	.27	.27	.31	.28	.23	.22	.25	.19
7	.29	.30	.27	.28	.29	.26	.27	.27	.24	.26	.26	.24	.24	.24	.30	.28	.22	.22	.27	.17
8	.29	.42	.26	.29	.41	.25	.30	.43	.28	.30	.43	.28	.19	.19	.21	.20	.18	.18	.27	.22
9	.34	.35	.32	.32	.32	.31	.30	.30	.29	.30	.29	.28	.32	.33	.35	.32	.24	.26	.29	.20
10	.31	.41	.25	.30	.41	.25	.31	.42	.26	.32	.43	.27	.18	.18	.20	.18	.16	.17	.27	.22
→ 11	-.22	-.22	-.16	-.21	-.22	-.16	-.22	-.22	-.17	-.21	-.23	-.17	-.15	-.14	-.17	-.16	-.13	-.12	-.18	-.10
12	.28	.26	.22	.28	.26	.22	.29	.25	.23	.29	.26	.23	.20	.20	.19	.19	.25	.20	.23	.23
13	.48	.51	.47	.43	.46	.43	.38	.40	.38	.35	.38	.36	.35	.34	.36	.35	.22	.26	.32	.24
14	.17	.18	.16	.17	.18	.16	.17	.18	.16	.17	.19	.16	.14	.14	.14	.14	.15	.14	.18	.19
15	.24	.22	.19	.24	.22	.18	.23	.22	.17	.23	.21	.18	.17	.17	.19	.18	.16	.17	.20	.16
16	.11	.16	.09	.12	.16	.10	.12	.16	.10	.12	.16	.10	.09	.09	.10	.10	.08	.09	.10	.11
17	.28	.30	.30	.28	.30	.30	.28	.30	.30	.28	.30	.30	.24	.24	.26	.26	.28	.27	.30	.27
18	.22	.30	.20	.21	.30	.20	.22	.30	.20	.22	.30	.20	.15	.15	.17	.17	.14	.14	.21	.17
19	.36	.42	.33	.32	.39	.30	.34	.41	.32	.35	.41	.33	.27	.27	.29	.28	.24	.23	.28	.23
20	.30	.34	.24	.30	.34	.25	.37	.41	.31	.37	.40	.30	.19	.18	.22	.21	.20	.19	.30	.19
21	.30	.39	.28	.30	.38	.28	.32	.40	.29	.32	.41	.30	.22	.22	.24	.23	.23	.23	.32	.27
22	.13	.17	.12	.13	.17	.12	.14	.18	.13	.14	.18	.13	.11	.11	.12	.11	.10	.11	.15	.15
23	.23	.23	.16	.21	.22	.15	.22	.22	.16	.22	.23	.16	.15	.16	.18	.16	.15	.14	.18	.14
24	.21	.22	.16	.21	.22	.16	.21	.22	.16	.22	.22	.16	.13	.13	.13	.13	.12	.12	.18	.21
25	.19	.20	.14	.19	.20	.14	.19	.21	.14	.19	.21	.14	.11	.11	.11	.11	.12	.10	.15	.20
26	.14	.17	.12	.14	.17	.12	.15	.18	.12	.15	.18	.13	.10	.10	.11	.11	.09	.09	.14	.20
27	.09	.10	.16	.09	.10	.16	.09	.10	.16	.09	.10	.16	.18	.18	.20	.20	.15	.14	.14	.15
28	.11	.14	.21	.11	.14	.21	.12	.15	.22	.12	.16	.22	.15	.15	.14	.13	.10	.10	.19	.22
29	.07	.09	.12	.07	.09	.12	.07	.09	.12	.07	.10	.12	.11	.11	.11	.10	.08	.08	.12	.16
30	.14	.17	.30	.14	.17	.30	.14	.17	.29	.14	.17	.28	.31	.32	.31	.30	.26	.26	.25	.21
31	.14	.15	.24	.14	.15	.24	.14	.15	.28	.15	.17	.28	.19	.19	.17	.16	.20	.16	.28	.20
32	.09	.10	.19	.09	.10	.20	.10	.12	.21	.10	.12	.21	.19	.19	.15	.16	.11	.11	.19	.13
33	.08	.08	.10	.08	.08	.10	.08	.08	.10	.08	.08	.10	.10	.10	.10	.10	.12	.10	.11	.11
34	.07	.07	.08	.07	.07	.08	.07	.08	.09	.07	.07	.09	.08	.08	.08	.07	.07	.07	.10	.08
35	.06	.06	.07	.06	.06	.07	.06	.06	.07	.06	.06	.07	.08	.08	.09	.10	.08	.08	.07	.07
36	.06	.06	.07	.06	.06	.07	.05	.05	.06	.05	.05	.06	.10	.10	.18	.17	.08	.12	.06	.05
37	.18	.19	.19	.18	.18	.19	.19	.19	.18	.18	.18	.18	.20	.20	.22	.22	.18	.19	.17	.15
38	.08	.08	.08	.08	.08	.08	.08	.08	.08	.08	.08	.08	.08	.08	.12	.12	.09	.10	.08	.08
39	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.05	.04	.04	.04	.04	.05
40	.07	.06	.07	.07	.06	.07	.07	.06	.07	.07	.06	.07	.08	.08	.08	.08	.09	.08	.07	.08
41	.07	.07	.06	.07	.07	.06	.07	.07	.06	.07	.07	.06	.06	.06	.08	.08	.07	.07	.06	.06
42	.16	.18	.16	.16	.18	.16	.16	.17	.16	.16	.18	.16	.15	.16	.16	.16	.17	.17	.17	.16
43	.24	.24	.21	.24	.24	.21	.24	.25	.22	.24	.25	.21	.19	.19	.19	.19	.29	.26	.23	.19
44	.16	.17	.14	.16	.17	.14	.16	.17	.14	.16	.17	.14	.14	.14	.15	.15	.17	.17	.16	.14
45	.14	.16	.15	.14	.16	.15	.14	.16	.15	.15	.17	.16	.15	.15	.14	.14	.18	.16	.15	.16
46	.12	.12	.12	.12	.12	.11	.12	.12	.11	.12	.12	.11	.11	.11	.11	.11	.14	.13	.12	.11
47	.25	.26	.26	.25	.26	.26	.25	.26	.26	.25	.26	.26	.25	.25	.26	.26	.27	.26	.25	.24
48	.30	.29	.32	.30	.29	.32	.30	.29	.31	.30	.29	.31	.32	.32	.33	.33	.25	.27	.31	.25
49	.12	.22	.04	.12	.22	.04	.13	.22	.06	.13	.22	.06	.02	.02	-.00	-.00	.01	-.01	.07	.06
50	-.05	-.04	-.04	-.05	-.04	-.04	-.05	-.04	-.05	-.05	-.04	-.05	-.05	-.05	-.04	-.04	-.04	-.03	-.05	-.05
51	-.15	-.09	-.14	-.15	-.09	-.14	-.14	-.09	-.13	-.14	-.09	-.13	-.11	-.09	-.06	-.08	-.14	-.14	-.12	-.14
52	-.13	-.13	-.14	-.13	-.13	-.14	-.13	-.14	-.14	-.13	-.14	-.14	-.14	-.14	-.14	-.14	-.09	-.10	-.13	-.11
53	.11	.09	.11	.11	.08	.11	.11	.08	.10	.11	.08	.10	.10	.10	.10	.10	.12	.11	.10	.12

TABLE 6 continued

CRITERION FACTORS (MEANS)

	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	.34	.27	.29	.28	.16	.23	.17	.14	.17	.26	.13	.09	.15	.15	.16	.13	.18	.31	.28	.56
2	.26	.25	.35	.31	.18	.21	.18	.14	.24	.37	.15	.09	.14	.15	.16	.13	.20	.25	.37	.23
3	.19	.16	.28	.28	.15	.19	.15	.11	.17	.32	.13	.08	.11	.11	.12	.11	.16	.18	.31	.17
4	.26	.20	.34	.31	.18	.22	.17	.11	.18	.31	.12	.08	.12	.13	.14	.11	.18	.24	.34	.29
5	.29	.24	.30	.30	.16	.19	.16	.11	.16	.25	.11	.08	.12	.12	.13	.10	.20	.31	.29	.52
6	.13	.12	.39	.37	.23	.33	.16	.14	.21	.25	.18	.10	.16	.17	.20	.15	.14	.23	.39	.11
7	.11	.11	.36	.36	.22	.33	.17	.14	.20	.24	.17	.10	.15	.17	.21	.17	.12	.23	.35	.11
8	.20	.15	.29	.32	.18	.26	.15	.11	.18	.23	.13	.11	.11	.12	.14	.11	.12	.21	.25	.26
9	.12	.12	.36	.35	.23	.30	.20	.16	.23	.26	.21	.12	.17	.19	.24	.20	.12	.26	.39	.12
10	.20	.14	.28	.32	.16	.24	.15	.11	.18	.25	.12	.09	.13	.13	.14	.12	.13	.20	.26	.30
(11)	-.10	-.07	-.30	-.26	-.14	-.37	-.10	-.07	-.18	-.16	-.18	-.07	-.10	-.12	-.11	-.10	-.06	-.19	-.27	-.07
12	.24	.33	.26	.20	.16	.32	.12	.13	.24	.24	.25	.18	.18	.18	.18	.16	.20	.19	.24	.12
13	.15	.14	.15	.16	.21	.13	.27	.20	.20	.21	.16	.10	.19	.22	.27	.23	.15	.15	.16	.09
14	.31	.21	.21	.16	.12	.22	.12	.12	.15	.18	.11	.08	.11	.12	.12	.11	.13	.12	.15	.18
15	.14	.17	.39	.35	.20	.25	.14	.11	.16	.19	.14	.08	.17	.18	.16	.15	.11	.18	.31	.11
16	.12	.10	.11	.14	.10	.12	.08	.07	.08	.10	.07	.07	.08	.08	.09	.07	.09	.22	.13	.40
17	.26	.24	.25	.22	.24	.17	.24	.22	.22	.27	.17	.11	.20	.20	.21	.19	.21	.20	.23	.18
18	.19	.13	.24	.28	.14	.18	.13	.09	.12	.18	.10	.09	.11	.11	.12	.11	.11	.20	.19	.32
19	.21	.18	.39	.37	.22	.31	.19	.13	.22	.28	.17	.11	.16	.18	.19	.14	.16	.26	.33	.24
20	.15	.15	.44	.40	.21	.25	.16	.11	.22	.26	.18	.10	.18	.21	.19	.14	.12	.19	.33	.14
21	.29	.23	.36	.39	.22	.30	.20	.14	.20	.30	.14	.11	.15	.17	.18	.15	.17	.30	.34	.48
22	.24	.18	.14	.15	.10	.13	.10	.08	.11	.13	.07	.07	.08	.09	.10	.07	.12	.21	.18	.47
23	.18	.14	.27	.24	.14	.38	.11	.08	.18	.18	.14	.08	.12	.13	.13	.10	.11	.20	.29	.17
24	.32	.29	.17	.16	.13	.29	.11	.14	.22	.21	.26	.26	.19	.20	.19	.16	.12	.15	.12	.14
25	.33	.22	.16	.15	.10	.31	.09	.11	.22	.20	.24	.20	.19	.20	.20	.15	.11	.18	.14	.13
26	.26	.17	.14	.13	.09	.22	.09	.11	.20	.19	.18	.16	.18	.19	.21	.15	.09	.16	.11	.13
27	.12	.19	.07	.07	.15	.08	.16	.22	.22	.18	.20	.31	.30	.34	.37	.26	.11	.11	.09	.06
28	.14	.16	.07	.07	.10	.09	.09	.10	.27	.25	.29	.13	.18	.20	.21	.16	.10	.16	.18	.07
29	.12	.12	.06	.06	.08	.12	.08	.08	.29	.23	.33	.16	.19	.21	.20	.14	.08	.15	.09	.06
30	.15	.23	.07	.06	.19	.07	.18	.26	.19	.20	.24	.25	.18	.19	.21	.19	.17	.12	.19	.06
31	.17	.36	.06	.06	.10	.06	.08	.10	.11	.13	.13	.07	.08	.08	.08	.07	.12	.12	.24	.06
32	.12	.14	.06	.06	.11	.07	.09	.12	.18	.15	.34	.13	.12	.12	.14	.13	.09	.17	.22	.06
33	.15	.28	.06	.06	.08	.06	.08	.10	.11	.11	.13	.14	.09	.09	.10	.10	.10	.07	.11	.06
34	.09	.13	.06	.06	.07	.07	.07	.09	.22	.18	.30	.14	.09	.10	.11	.09	.07	.11	.13	.06
35	.07	.12	.06	.06	.08	.06	.10	.18	.13	.13	.17	.28	.10	.11	.12	.10	.09	.07	.07	.06
36	.05	.05	.06	.05	.13	.06	.19	.19	.12	.10	.16	.42	.10	.12	.15	.16	.06	.05	.04	.04
37	.14	.15	.13	.14	.19	.15	.20	.20	.16	.15	.18	.29	.15	.16	.18	.21	.13	.17	.14	.23
38	.12	.08	.06	.08	.11	.08	.12	.12	.09	.10	.07	.03	.08	.08	.09	.08	.08	.15	.08	.18
39	.06	.04	.05	.04	.04	.04	.05	.04	.04	.04	.03	-.01	.04	.05	.05	.05	.04	.10	.04	.23
40	.08	.07	.06	.06	.07	.07	.07	.07	.07	.06	.09	.09	.07	.07	.07	.07	.07	.06	.07	.07
41	.08	.06	.08	.11	.09	.08	.07	.06	.08	.08	.10	.19	.07	.07	.08	.08	.06	.10	.06	.22
42	.16	.15	.19	.20	.17	.18	.15	.14	.15	.16	.18	.23	.14	.15	.16	.15	.13	.18	.16	.31
43	.20	.21	.21	.24	.20	.20	.20	.19	.19	.22	.19	.13	.18	.18	.17	.17	.19	.19	.26	.21
44	.13	.13	.16	.17	.14	.14	.14	.13	.15	.16	.16	.21	.13	.14	.14	.13	.12	.18	.16	.22
45	.18	.14	.15	.17	.17	.19	.15	.17	.21	.21	.26	.28	.15	.16	.14	.14	.13	.16	.14	.24
46	.12	.11	.11	.11	.11	.10	.11	.10	.11	.12	.10	.07	.12	.10	.10	.11	.10	.11	.11	.12
47	.23	.24	.23	.24	.25	.23	.25	.23	.22	.23	.20	.17	.23	.23	.24	.24	.23	.25	.26	.25
48	.18	.14	.17	.20	.26	.15	.27	.25	.26	.25	.25	.22	.27	.27	.29	.28	.10	.02	.09	-.07
49	.03	-.00	.17	.19	.02	.16	-.00	-.01	.02	.05	.03	-.01	.01	.01	.01	.00	-.02	.08	.13	.02
50	.01	-.04	-.01	.00	-.02	.02	-.00	-.02	.02	.02	.02	.05	-.00	.01	.00	-.00	-.04	.08	-.03	.30
51	-.10	-.16	-.02	.01	-.13	-.03	-.14	-.14	-.13	-.12	-.14	-.17	-.12	-.13	-.14	-.18	-.06	.27	.17	.22
52	-.01	-.06	-.04	-.04	-.08	-.06	-.10	-.10	-.10	-.10	-.10	-.13	-.07	-.08	-.08	-.10	-.08	.07	-.03	.19
53	.15	.22	.06	.05	.07	.06	.11	.10	.10	.10	.08	.04	.12	.12	.12	.11	.22	-.05	.09	-.04

TABLE 6 continued

CRITERION FACTORS (MEANS)

	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	.34	.29	.40	.46	.46	.20	.28	.45	.34	.20	.30	.23	.34	.20	.12	.14	.13	.12	.17	.15
2	.30	.29	.33	.42	.39	.24	.15	.15	.28	.18	.39	.15	.38	.13	.10	.11	.10	.11	.21	.11
3	.22	.21	.24	.34	.34	.20	.11	.12	.24	.13	.34	.12	.33	.12	.09	.10	.10	.10	.17	.10
4	.26	.25	.29	.41	.45	.22	.16	.20	.29	.16	.34	.14	.37	.15	.11	.13	.12	.11	.17	.12
5	.34	.30	.40	.46	.50	.22	.24	.34	.31	.18	.32	.15	.34	.21	.12	.13	.12	.12	.17	.13
6	.11	.12	.11	.19	.18	.10	.08	.09	.19	.17	.18	.10	.39	.08	.06	.07	.07	.08	.14	.08
7	.10	.13	.11	.18	.20	.10	.08	.10	.22	.15	.17	.10	.38	.08	.07	.08	.08	.08	.13	.08
8	.19	.17	.20	.38	.48	.13	.13	.26	.30	.14	.24	.18	.36	.14	.10	.11	.10	.10	.14	.12
9	.11	.13	.12	.16	.20	.12	.09	.10	.20	.16	.19	.12	.38	.09	.07	.08	.08	.08	.13	.08
10	.21	.18	.22	.38	.47	.14	.15	.31	.30	.16	.27	.20	.34	.15	.10	.12	.12	.10	.14	.13
11	-.08	-.07	-.08	-.15	-.20	-.06	-.04	-.05	-.14	-.09	-.12	-.11	-.30	-.04	-.02	-.03	-.03	-.02	-.09	-.05
12	.33	.29	.28	.36	.25	.13	.11	.12	.21	.18	.22	.16	.40	.08	.07	.07	.07	.09	.22	.09
13	.10	.10	.10	.12	.14	.13	.08	.09	.16	.13	.18	.12	.19	.09	.08	.08	.08	.09	.10	.08
14	.26	.24	.21	.33	.24	.14	.12	.16	.20	.18	.22	.11	.27	.14	.08	.08	.08	.07	.14	.07
15	.17	.18	.11	.16	.20	.10	.08	.11	.15	.16	.15	.12	.33	.08	.07	.06	.06	.07	.12	.07
16	.13	.12	.16	.20	.30	.09	.14	.24	.15	.08	.12	.12	.14	.06	.06	.06	.07	.08	.06	.10
17	.29	.27	.30	.28	.21	.31	.14	.14	.23	.27	.31	.17	.25	.28	.17	.22	.16	.14	.24	.06
18	.18	.16	.18	.30	.38	.11	.12	.23	.24	.12	.19	.15	.24	.11	.09	.10	.09	.09	.11	.11
19	.20	.20	.20	.34	.40	.17	.16	.22	.26	.16	.26	.17	.39	.12	.10	.10	.10	.10	.18	.13
20	.16	.16	.14	.20	.22	.12	.10	.11	.16	.15	.18	.14	.39	.10	.08	.08	.08	.09	.16	.09
21	.30	.28	.34	.46	.53	.19	.23	.38	.44	.23	.36	.24	.44	.18	.13	.15	.17	.16	.21	.20
22	.26	.19	.26	.33	.34	.13	.23	.36	.21	.13	.16	.15	.23	.11	.08	.10	.10	.10	.12	.10
23	.16	.14	.14	.31	.32	.10	.11	.15	.14	.11	.15	.12	.32	.08	.08	.08	.08	.09	.15	.09
24	.16	.14	.14	.24	.20	.11	.11	.14	.18	.11	.15	.17	.36	.08	.07	.07	.07	.08	.12	.07
25	.16	.14	.13	.23	.21	.10	.13	.17	.18	.10	.14	.20	.43	.08	.08	.08	.08	.08	.18	.07
26	.12	.12	.11	.21	.18	.14	.13	.16	.19	.09	.12	.20	.43	.07	.06	.06	.06	.08	.11	.07
27	.07	.08	.07	.07	.06	.12	.06	.06	.16	.10	.12	.18	.13	.06	.07	.06	.06	.07	.08	.07
28	.07	.07	.07	.08	.06	.09	.06	.06	.17	.10	.17	.08	.18	.06	.06	.06	.06	.07	.09	.07
29	.06	.06	.06	.07	.06	.08	.06	.06	.10	.08	.11	.09	.17	.06	.06	.05	.06	.07	.08	.06
30	.09	.11	.08	.07	.06	.16	.05	.06	.24	.19	.19	.12	.12	.06	.06	.06	.06	.07	.08	.06
31	.11	.07	.08	.07	.06	.14	.05	.05	.26	.17	.22	.06	.11	.06	.06	.06	.06	.06	.08	.06
32	.07	.06	.06	.05	.05	.11	.05	.06	.23	.13	.19	.07	.09	.05	.06	.05	.06	.06	.07	.06
33	.09	.07	.07	.07	.06	.10	.06	.06	.14	.11	.12	.08	.10	.05	.06	.06	.06	.06	.07	.06
34	.08	.06	.06	.06	.06	.08	.06	.06	.14	.08	.13	.08	.09	.05	.06	.05	.06	.06	.08	.06
35	.07	.07	.06	.06	.06	.10	.06	.06	.10	.10	.09	.12	.07	.05	.06	.05	.06	.06	.05	.06
36	.05	.06	.04	.04	.04	.07	.06	.06	.09	.06	.06	.29	.06	.12	.15	.12	.13	.20	.08	.14
37	.14	.16	.14	.15	.17	.18	.22	.22	.21	.15	.16	.23	.22	.13	.14	.19	.22	.39	.19	.36
38	.09	.13	.09	.07	.07	.13	.36	.33	.26	.12	.10	.20	.19	.30	.28	.50	.34	.24	.17	.24
39	.04	.06	.04	.03	.04	.06	.33	.34	.18	.05	.04	.13	.07	.09	.10	.20	.23	.07	.03	.13
40	.07	.07	.07	.08	.03	.07	.07	.04	.08	.08	.08	.15	.08	.42	.38	.37	.25	.29	.19	.14
41	.06	.08	.07	.06	.10	.08	.13	.22	.13	.06	.07	.36	.18	.10	.16	.09	.11		.0	.23
42	.14	.14	.15	.17	.21	.14	.24	.25	.23	.14	.16	.26	.25	.14	.17	.17	.16		.16	.27
43	.28	.29	.29	.26	.24	.21	.17	.19	.27	.26	.30	.17	.27	.40	.32	.36	.31		.46	.24
44	.14	.14	.14	.15	.18	.15	.16	.20	.22	.14	.17	.26	.23	.16	.17	.17	.18	.1	.24	.33
45	.15	.14	.15	.18	.21	.15	.22	.25	.29	.14	.16	.32	.42	.24	.20	.26	.27	'	.22	.22
46	.13	.12	.13	.11	.12	.12	.15	.14	.13	.12	.13	.20	.16	.49	.42	.40	.35	)	.27	.24
47	.26	.26	.27	.26	.28	.26	.21	.22	.26	.24	.26	.21	.28	.31	.30	.33	.31	.34	.41	.37
48	.00	.06	.01	.02	-.03	.11	-.10	-.14	.02	.08	.14	.06	.08	.12	.14	.10	.09	.13	.14	.07
49	-.00	-.01	.02	.20	.31	-.01	-.05	-.01	.10	.04	.22	-.12	.06	.01	.00	-.00	.00	.10	.14	.12
50	-.01	.03	-.01	-.03	-.04	.01	.21	.26	.08	-.00	-.03	.22	.05	.03	.04	.12	.12	.1	.01	.25
51	-.14	-.15	-.12	-.00	.11	.05	-.00	.05	-.02	-.09	-.06	-.16	-.12	-.16	-.20	-.10	-.05	-.07	-.05	.01
52	-.05	-.01	-.05	-.03	-.03	.02	.29	.31	.24	.02	-.01	.07	-.04	.05	.03	.15	.24	.06	-.00	.08
53	.27	.27	.27	.14	.01	.16	.09	.04	.05	.18	.16	.02	.05	.25	.22	.23	.19	.17	.23	-.01

TABLE 6 continued

	CRITERION FACTORS (MEANS)											
	61	62	63	64	65	66	67	68	69	70	71	72
1	.12	.30	.21	.08	-.13	-.13	-.01	.04	.42	.24	.51	.28
2	.12	.20	.18	.07	-.09	-.09	.02	.04	.38	.22	.44	.33
3	.09	.17	.16	.07	-.08	-.08	-.00	.04	.34	.19	.40	.28
4	.11	.24	.21	.07	-.11	-.10	-.02	.04	.42	.24	.48	.32
5	.11	.25	.21	.08	-.12	-.12	-.01	.05	.45	.26	.51	.32
6	.06	.10	.16	.06	-.03	-.04	.03	.04	.20	.10	.22	.25
7	.07	.10	.16	.06	-.04	-.04	.03	.04	.20	.10	.22	.26
8	.11	.22	.21	.08	-.09	-.08	-.01	.04	.38	.21	.45	.34
9	.08	.13	.17	.07	-.05	-.06	.02	.04	.28	.12	.28	.30
10	.09	.23	.22	.08	-.08	-.07	-.02	.05	.36	.18	.44	.34
(11)	-.04	-.10	-.14	-.04	.02	.03	.00	-.01	-.12	-.08	-.12	-.15
12	.09	.10	.15	.07	-.05	-.04	.03	.04	.20	.10	.20	.23
13	.08	.12	.16	.08	-.06	-.07	.04	.06	.27	.14	.27	.33
14	.07	.10	.13	.06	-.05	-.04	.04	.06	.30	.12	.34	.25
15	.06	.11	.16	.07	-.04	-.04	.03	.04	.18	.09	.19	.21
16	.06	.16	.13	.06	-.02	-.02	-.00	.04	.18	.11	.23	.13
17	.17	.12	.16	.10	-.16	-.13	.05	.11	.29	.14	.29	.29
18	.09	.21	.17	.07	-.08	-.07	-.01	.03	.30	.13	.36	.23
19	.11	.22	.21	.08	-.09	-.07	-.00	.02	.34	.17	.42	.33
20	.09	.13	.17	.08	-.05	-.04	.02	.02	.22	.12	.25	.23
21	.16	.34	.31	.12	-.14	-.11	-.03	.01	.47	.28	.54	.38
22	.09	.20	.15	.06	-.06	-.05	.01	.03	.30	.15	.33	.20
23	.09	.12	.16	.07	-.04	-.04	.02	.02	.21	.12	.24	.19
24	.07	.12	.22	.08	-.04	-.03	.04	.03	.20	.12	.22	.21
25	.09	.11	.21	.06	-.05	-.05	.02	.04	.21	.12	.20	.22
26	.06	.12	.20	.06	-.04	-.04	.02	.04	.16	.09	.17	.19
27	.07	.07	.18	.15	-.03	-.03	.05	.05	.11	.08	.08	.19
28	.05	.07	.11	.08	-.03	-.03	.05	.05	.08	.08	.07	.16
29	.06	.06	.12	.07	-.02	-.02	.04	.04	.08	.07	.07	.14
30	.06	.06	.14	.10	-.02	-.02	.06	.05	.09	.07	.06	.20
31	.06	.06	.07	.06	-.02	-.02	.04	.04	.09	.07	.07	.15
32	.06	.06	.10	.07	-.02	-.02	.04	.04	.07	.06	.06	.14
33	.06	.06	.07	.06	-.02	-.02	.04	.04	.09	.07	.08	.11
34	.06	.06	.10	.05	-.02	-.02	.04	.04	.07	.06	.06	.12
35	.05	.06	.12	.09	-.02	-.02	.04	.04	.07	.06	.06	.10
36	.14	.17	.31	.55	-.07	-.05	.09	.13	.18	.16	.08	.17
37	.16	.31	.29	.35	-.07	-.08	.09	.20	.27	.36	.20	.26
38	.18	.27	.19	.10	-.30	-.23	.15	.27	.20	.24	.14	.16
39	.00	.22	.08	.06	-.04	-.05	.08	.12	.09	.11	.05	.07
40	.34	.15	.11	.14	-.33	-.28	.21	.22	.18	.23	.10	.12
41	.12	.44	.23	.18	-.08	-.06	.09	.09	.16	.20	.09	.15
42	.20	.33	.25	.20	-.16	-.15	.10	.21	.22	.24	.18	.19
43	.43	.21	.20	.21	-.34	-.31	.15	.18	.30	.33	.26	.27
44	.28	.30	.30	.23	-.21	-.18	.08	.19	.27	.30	.23	.24
45	.32	.34	.38	.17	-.29	-.26	.15	.24	.26	.24	.21	.22
46	.44	.27	.17	.16	-.53	-.45	.20	.20	.27	.25	.18	.18
47	.38	.30	.23	.21	-.34	-.33	.18	.24	.36	.43	.30	.32
48	.09	.04	.15	.14	-.08	-.08	.15	.20	.11	.12	.03	.22
49	.07	-.09	.05	.02	-.01	-.04	-.03	.04	.14	.12	.25	.10
50	.03	.35	.14	.08	-.04	-.06	.05	.12	.11	.14	.07	.08
51	-.11	-.12	-.03	-.03	.12	.12	-.17	-.12	-.04	-.03	-.02	-.06
52	-.02	.16	.04	.04	-.06	-.05	.03	.08	.08	.06	.07	.06
53	.21	-.02	.06	.07	-.17	-.14	.13	.15	.14	.12	.09	.11

Table 7. Nine Factor Solution for 53 Predictors

VAR	MS	A	B	C	D	E	F	G	H	I
5	978	97*	00	14	02	10	-02	01	-05	-06
18	977	95*	03	06	21	07	09	01	07	07
21	982	95*	05	13	18	14	09	04	-04	-01
1	975	94*	03	12	-03	23	01	07	-07	-11
4	973	94*	08	14	23	-05	-02	02	04	03
10	972	93*	03	06	25	03	03	04	13	13
8	969	92*	04	07	29	-02	06	01	13	13
2	950	88*	10	14	28	-17	-09	03	06	-05
3	940	88*	14	15	29	-17	-04	02	05	04
22	956	86*	-05	07	-19	34	04	05	-16	-13
16	901	85*	-04	-07	-07	34	15	-02	-17	04
14	915	85*	11	17	03	-19	-10	25	-02	-24
19	938	83*	15	04	49	-07	01	01	16	04
23	949	75*	06	06	53*	-08	-01	25	-12	01
49	881	66*	-13	10	26	-37	-03	-05	-19	42
12	921	65*	31	09	34	-26	-11	37	04	-26
17	931	64*	34	43	20	-19	-13	-03	30	-22
			--B-							
32	939	11	92*	03	23	-05	-05	-17	03	-00
28	955	17	89*	04	23	-06	-07	12	19	14
34	910	07	87*	14	07	-01	06	25	-20	12
29	951	01	85*	04	11	-01	06	41	02	22
30	955	05	77*	-04	23	-19	18	-24	36	-27
27	816	-11	72*	-03	05	-10	36	21	25	-11
33	865	22	72*	17	-01	-11	06	11	02	-49
35	840	-07	66*	07	-07	-10	53*	25	-01	-19
31	846	31	63*	07	21	-13	-21	-31	17	-34
				--C-						
46	960	11	-00	93*	-01	22	18	02	07	04
40	931	-09	-05	93*	-07	15	15	-01	05	08
43	966	34	07	91*	07	01	10	-02	-02	-04
47	930	34	13	84*	10	09	26	-05	03	-01
53	936	03	07	79*	-21	-25	-24	09	14	-32
38	925	01	-07	72*	-10	62*	08	-03	01	01
45	930	25	18	65*	06	40	42	26	-03	02
44	956	32	08	65*	07	23	57*	02	-07	07
					--D-					
6	966	48	28	-01	79*	-14	-03	03	07	-05
7	975	52	27	00	77*	-13	-02	02	15	-01
15	935	57*	17	04	72*	-12	-03	17	03	-12
9	965	52*	33	-00	72*	-13	01	-05	23	00
11	888	-16	-05	30	-70*	23	05	-41	10	-15
20	926	51*	17	03	66*	-15	-05	10	20	08
						--E-				
39	932	08	-12	23	-16	91*	03	-07	-03	-02
50	947	-04	-21	15	-25	79*	40	05	-20	11
52	915	-01	-28	21	-33	73*	05	-05	-37	02
							--F-			
36	860	-29	14	24	-09	-00	83*	04	06	-00
37	874	16	12	45	03	23	73*	-12	06	05
41	864	08	-09	25	-05	59*	55*	13	-05	02
42	948	39*	13	50*	12	43*	53*	04	-03	-02
								--G-		
25	944	57*	33	08	24	04	07	67*	01	-02
26	920	54*	36	01	21	14	11	64*	05	-03
24	944	53*	41	00	21	-01	12	62*	09	-08
									--H-	
51	883	22	-20	-40	-00	25	-04	-31	-67*	19
13	947	44	33	-00	40	-23	01	-22	64*	08
48	946	-06	45*	14	44*	-39	05	-07	60*	08

SUM OF COMMUNALITIES = 49.346

FACTOR VARIANCES:

A:	16.067	D:	5.445	G:	2.545
B:	7.270	E:	4.579	H:	2.170
C:	6.652	F:	3.393	I:	1.221

Table 8

HIERARCHICAL MAP OF PREDICTOR SPACE

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension 5. Reading Comprehension 16. Ideational Fluency 18. Analogical Reasoning 21. Omnibus Intelligence/Aptitude 22. Word Fluency	A. Verbal Ability/General Intelligence	
4. Word Problems 8. Inductive Reasoning: Concept Formation 10. Deductive Logic	B. Reasoning	
2. Numerical Computation 3. Use of Formula/Number Problems	C. Number Ability	COGNITIVE ABILITIES
12. Perceptual Speed and Accuracy	E. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory 17. Follow Directions	J. Memory	
19. Figural Reasoning 23. Verbal and Figural Closure	F. Closure	
6. Two-dimensional Mental Rotation 7. Three-dimensional Mental Rotation 9. Spatial Visualization 11. Field Dependence (Negative) 15. Place Memory (Visual Memory) 20. Spatial Scanning	Z. Visualization/Spatial	VISUALIZATION/SPATI
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental/Information Processing	INFORMATION PROCES
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests 51. Artistic Interests (Negative)	N. Realistic vs. Artistic Interests	
28. Control Precision 29. Rate Control 32. Arm-Hand Steadiness 34. Aiming	I. Steadiness/Precision	
27. Multilimb Coordination 35. Speed of Arm Movement	D. Coordination	PSYCHOMOTOR
30. Manual Dexterity 31. Finger Dexterity 33. Wrist-Finger Speed	K. Dexterity	
39. Sociability 52. Social Interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising Interests	R. Enterprising Interest	
36. Involvement in Athletics and Physical Conditioning 37. Energy Level	T. Athletic Abilities/Energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/Self-Esteem	
40. Traditional Values 43. Conscientiousness 44. Non-delinquency 53. Conventional Interests	M. Traditional Values/Conventionality/Non-Delinquency	
46. Locus of Control 47. Work Orientation	O. Work Orientation/Locus of Control	MOTIVATION/STABIL
38. Cooperativeness 45. Emotional Stability	P. Cooperation/Emotional Stability	

Table 9. Six Factor Solution for 72 Criteria

VAR	M2	A	B	C	D	E	F
3	977	95*	04	03	20	14	11
6	981	95*	04	03	21	17	11
9	970	94*	02	04	18	20	07
1	953	94*	06	11	09	11	-17
10	973	94*	03	07	17	23	-01
12	968	94*	02	06	15	21	05
4	962	94*	07	12	10	14	-18
2	956	93*	-02	20	00	09	-21
7	966	93*	06	14	09	15	-22
15	964	93*	07	-02	29	-00	11
5	958	92*	-01	21	01	11	-22
10	967	92*	06	15	09	15	-23
13	980	92*	08	-06	31	07	14
14	972	92*	07	-07	31	07	15
16	973	92*	10	-03	33	01	13
8	956	91*	-01	21	01	13	-25
11	953	90*	-01	23	00	15	-26
18	967	87*	26	04	33	15	10
25	960	87*	24	09	37	06	-02
17	933	84*	29	01	41	22	07
72	945	83*	17	36	18	16	-15
20	951	83*	06	11	28	42	-08
30	909	81*	01	12	42	32	-19
24	928	81*	00	28	-07	01	-43
27	952	81*	31	07	42	07	11
39	796	80*	-12	18	-15	15	-25
23	897	80*	-03	25	-04	08	-43
51	874	76*	11	24	-05	45	-13
37	823	70*	26	11	23	55*	04
50	892	68*	34	18	15	58*	08
26	859	67*	-03	20	19	13	-57*
29	903	66*	03	-08	60*	27	-15
53	874	62*	03	38	25	25	-47*
71	938	62*	05	58*	-20	25	-32
69	936	61*	30	60*	-07	23	-25
46	812	60*	31	30	10	63*	13
44	956	57*	-01	49*	-13	47*	-40
---			--B-				
66	941	-09	-95*	-09	02	-10	07
65	935	-06	-95*	-09	03	-11	06
55	940	10	95*	02	13	13	00
56	918	-02	93*	16	-04	09	11
61	931	14	93*	03	13	13	-14
54	926	18	91*	07	-03	25	-03
58	863	08	89*	17	20	-04	-03
68	891	-09	88*	02	26	-12	17
57	859	-10	87*	27	-04	01	17
67	928	-09	84*	-24	35	-01	18
70	895	28	72*	51	10	01	-12
59	901	43	73*	38	09	32	-26
60	832	-02	66*	52*	20	-29	-06
---				--C-			
48	923	-04	15	93*	-12	14	-00
40	880	10	01	90*	-19	15	-08
47	835	-13	35	81*	-06	15	13
62	897	07	47	78*	21	-12	-07
49	789	47	11	68*	02	31	05
38	708	52*	-16	58*	-13	09	-23
45	926	55*	-07	58*	-20	25	-42*
---				--D-			
32	890	15	13	-01	92*	-02	-01
31	794	40	02	-21	73*	21	-07
36	959	65*	22	-12	68*	11	08
34	905	62*	10	-13	68*	20	-04
35	905	65*	07	-13	66*	14	03
33	913	62*	15	-15	65*	24	-04
28	935	62*	29	-07	63*	15	22
52	881	53	41	59*	59*	-02	-05
63	928	33	36	55*	59*	-09	-23
64	663	-02	45*	17	56*	-29	06
---						--E-	
22	872	50*	05	04	34	71*	06
41	943	51*	22	38	03	68*	-17
21	880	51*	05	31	26	64*	-18
43	938	52*	23	45*	-03	63*	-14
42	915	55*	30	37	07	61*	-13

SUM OF COMMUNALITIES = 65.675

FACTOR VARIANCES:

A:	30.571	D:	7.182
B:	11.900	E:	5.242
C:	8.133	F:	2.648

HIERARCHICAL MAP OF CRITERION SPACE

CONSTRUCTS	CLUSTERS	FACTORS
1. Inspect Mechanical Systems 2. Troubleshoot Mechanical Systems 4. Inspect Fluid Systems 5. Troubleshoot Fluids Systems 7. Inspect Electrical Systems 8. Troubleshoot Electrical Systems 10. Inspect Electronic Systems 11. Troubleshoot Electronic Systems	1. Inspect/Troubleshoot Equipment	
3. Repair Mechanical Systems 6. Repair Fluids Systems 9. Repair Electrical Systems 12. Repair Electronic Systems 19. Install Electronic Components	2. Repair/Install Equipment	
13. Repair Metal 14. Repair Plastic and Fiberglass Structures 15. Construct Wooden Buildings and Other Structures 16. Construct Masonry Buildings and Structures	3. Construction/Repair Metal, Fiberglass, Wood, Masonry	
17. Prepare Parachutes 18. Prepare Equipment and Supplies for Air Drop 25. Place Tactical Equipment and Materials in the Field	4. Parachute Preparation/Field Placement of Equipment	
20. Operate Electronic Equipment 30. Operate Fire Controls of Indirect Fire Weapons	5. Operate Electronic Equipment/Fire Control, Indirect	TECHNICAL SKILLS
23. Use Maps in the Field 24. Plan Placement or Use of Tactical Position and Features 26. Detect and Identify Targets 29. Fire Heavy Direct Fire Weapons 39. Draw Maps and Overlays	6. Battlefield Perception/Planning	
37. Reproduce Printed Materials 46. Prepare Food 50. Select, Layout & Close Medical/Dental Equipment/Supplies 51. Perform Medical Laboratory Procedures	11. Food Preparation/Medical Preparation and Laboratory Procedures	
69. Training Progress/Success 71. Performance of Theoretical, or "Classroom" Parts of Training 72. Performance of Practical, "Hands-On" Parts of Training	15. Training Performance	
53. Control Air Traffic	16. Air Traffic Control	
21. Send and Receive Radio Messages 22. Operate Keyboard Device 41. Record and File Information 42. Receive, Store, Issue Supplies, Equipment & Other Material 43. Prepare Technical Forms and Documents	10. Clerical	CLERICAL/DATA
44. Translate or Decode Data	12. Translate/Decode Data	
32. Engage in Bayonet and Hand-to-Hand Combat 52. Control Individuals and Crowds 63. Survive in the Field 64. Maintain Physical Fitness	8. Physical Combat Tasks	
27. Prepare Heavy Weapons for Tactical Use 28. Load Field Artillery or Tank Guns 31. Fire Individual Weapons 33. Operate Wheeled Vehicles 34. Operate Tract Vehicles 35. Operate Lifting, Loading, and Grading Equipment 36. Operate Power Excavating Equipment	7. Operate Heavy Artillery, Wheel and Track Vehicles	COMBAT
38. Make Movies and Videotapes 40. Write and Deliver Presentations 45. Analyze Intelligence Data 47. Receive Clients, Patients, Guests 48. Interview 49. Provide Medical and Dental Treatment 62. Emergent Leadership	9. Personal Interaction/Presentation Tasks	PERSONAL INTERACT
54. Following Regulations 55. Commitment to Army Norms 56. Cooperation with Supervisors 57. Cooperation with Other Unit Members 58. Hard Work and Perseverance 59. Attention to Detail 61. Discipline 65. Disciplinary Problems 66. Attrition 67. Reenlistment 68. Job Satisfaction/Morale	13. Commitment, Discipline, Cooperation, Perseverance	COMMITMENT/INITIA
60. Initiative 70. Effort/Motivation in Training	14. Initiative/Effort in Training	

TABLE 11

Mean (SD) of Mean Estimated Validities of
Predictor Factors for Criterion Factors

Predictor Factors	Criterion Factors: Technical Skills	Clerical Data	Combat	Personal Interaction	Commitment/ Initiative
Cognitive Abilities	.23 (.09)	.24 (.10)	.13 (.05)	.24 (.11)	.10 (.06)
Visualization/Spatial	.24 (.08)	.13 (.03)	.14 (.04)	.14 (.05)	.07 (.03)
Information Processing	.16 (.06)	.19 (.07)	.17 (.05)	.15 (.03)	.07 (.03)
Mechanical	.21 (.12)	.10 (.06)	.18 (.07)	.10 (.07)	.10 (.04)
Psychomotor	.12 (.06)	.10 (.06)	.14 (.07)	.08 (.05)	.05 (.02)
Social Skills	.06 (.04)	.03 (.02)	.06 (.05)	.19 (.11)	.08 (.06)
Vigor	.13 (.06)	.10 (.05)	.20 (.10)	.18 (.10)	.16 (.07)
Motivation/Stability	.15 (.07)	.16 (.07)	.15 (.07)	.18 (.09)	.28 (.10)

TABLE 12

Mean (SD) of Mean Estimated Validities of Predictor Clusters for Criterion Clusters

		Technical Skills									Clerical		Combat		Pers. Int.	Commitment/ Initiative		
		Criterion Cluster #:																
Predictor Cluster		1	2	3	4	5	6	11	15	16	10	12	8	5	9	13	14	
Verbal	A	.25 (.06)	.21 (.07)	.16 (.04)	.15 (.03)	.20 (.07)	.21 (.08)	.18 (.07)	.24 (.12)	.29 (.10)	.24 (.08)	.32 (.10)	.15 (.07)	.12 (.03)	.31 (.21)	.16 (.05)	.14 (.04)	
Reasoning	B	.36 (.03)	.22 (.01)	.20 (.07)	.17 (.01)	.25 (.04)	.20 (.05)	.19 (.07)	.29 (.03)	.34 (.01)	.22 (.04)	.29 (.01)	.24 (.04)	.23 (.02)	.22 (.09)	.10 (.04)	.12 (.03)	
Number	C	.32 (.04)	.23 (.05)	.22 (.01)	.17 (.02)	.23 (.05)	.22 (.04)	.23 (.08)	.26 (.05)	.26 (.03)	.25 (.03)	.28 (.04)	.22 (.04)	.24 (.02)	.21 (.08)	.10 (.04)	.14 (.03)	
Per. S & A	N	.22 (.01)	.23 (.00)	.20 (.01)	.20 (.03)	.24 (.01)	.25 (.04)	.18 (.02)	.21 (.01)	.48 —	.29 (.05)	.36 —	.24 (.04)	.22 (.04)	.24 (.05)	.06 (.03)	.10 (.01)	Cognitive Abilities
Inv. Int.	U	.22 (.02)	.03 (.01)	.01 (.01)	.01 (.00)	.04 (.01)	.13 (.06)	.02 (.09)	.16 (.04)	.06 —	.01 (.01)	.20 —	.03 (.04)	.01 (.03)	.09 (.09)	.04 (.04)	.12 —	
Memory	J	.23 (.04)	.23 (.07)	.20 (.04)	.20 (.07)	.23 (.04)	.20 (.03)	.22 (.07)	.20 (.02)	.26 (.01)	.26 (.03)	.31 (.02)	.22 (.04)	.24 (.03)	.22 (.04)	.12 (.07)	.10 (.03)	
Closure	F	.30 (.08)	.24 (.00)	.22 (.04)	.19 (.04)	.21 (.03)	.20 (.07)	.15 (.03)	.20 (.08)	.36 (.04)	.20 (.02)	.33 (.02)	.23 (.03)	.24 (.02)	.22 (.08)	.06 (.04)	.13 (.03)	
Vis./Spatial	K	.22 (.03)	.24 (.03)	.24 (.04)	.20 (.04)	.20 (.03)	.31 (.07)	.13 (.04)	.22 (.03)	.36 (.02)	.12 (.03)	.37 (.02)	.23 (.04)	.16 (.04)	.14 (.03)	.04 (.03)	.09 (.02)	Visual/Spatial
Mental/Info. Proces.	G	.19 (.01)	.14 (.02)	.12 (.01)	.11 (.01)	.20 (.01)	.20 (.04)	.12 (.02)	.20 (.03)	.42 (.03)	.19 (.03)	.23 (.03)	.12 (.06)	.12 (.03)	.13 (.03)	.02 (.02)	.09 (.03)	Info Process
Mech. Comp.	L	.42 (.03)	.29 (.03)	.23 (.01)	.23 (.02)	.21 (.03)	.16 (.02)	.15 (.02)	.29 (.03)	.19 —	.12 (.02)	.12 —	.13 (.03)	.12 (.04)	.12 (.03)	.08 (.03)	.11 (.03)	Mechanical
Mech. Int.	M	.21 (.09)	.22 (.09)	.21 (.12)	.20 (.06)	.19 (.04)	.12 (.04)	.09 (.02)	.08 (.07)	.10 (.02)	.11 (.04)	.01 (.01)	.12 (.07)	.11 (.07)	.09 (.08)	.12 (.04)	.06 (.04)	
Steadiness/Precis.	I	.10 (.03)	.13 (.05)	.12 (.04)	.09 (.02)	.10 (.05)	.12 (.01)	.11 (.03)	.09 (.05)	.13 (.04)	.09 (.05)	.02 (.01)	.10 (.03)	.16 (.08)	.09 (.03)	.03 (.02)	.07 (.01)	
Coordination	D	.08 (.02)	.11 (.04)	.14 (.03)	.11 (.03)	.13 (.04)	.09 (.03)	.10 (.01)	.10 (.04)	.10 (.03)	.09 (.04)	.02 (.01)	.10 (.03)	.20 (.09)	.08 (.03)	.05 (.02)	.07 (.01)	Psychomotor
Dexterity	K	.13 (.04)	.22 (.08)	.20 (.09)	.24 (.04)	.16 (.04)	.10 (.04)	.13 (.04)	.10 (.04)	.11 (.01)	.16 (.08)	.02 --	.10 (.03)	.11 (.04)	.09 (.04)	.05 (.02)	.02 (.01)	
Sociability	Q	.09 (.05)	.09 (.03)	.09 (.03)	.07 (.03)	.08 (.03)	.05 (.02)	.04 (.02)	.02 (.01)	.06 (.02)	.04 (.02)	.02 --	.07 (.04)	.02 (.02)	.20 (.10)	.08 (.03)	.10 (.02)	Social Skills
Enter. Int.	R	.05 (.01)	.05 (.00)	.05 (.01)	.02 (.01)	.04 (.01)	.02 (.01)	.02 (.02)	.09 (.02)	.03 —	.02 (.03)	.03 —	.12 (.06)	.01 (.01)	.19 (.11)	.02 (.04)	.20 (.04)	
Athletic Ability	T	.17 (.04)	.17 (.04)	.17 (.03)	.13 (.04)	.11 (.04)	.11 (.04)	.11 (.03)	.19 (.06)	.14 (.08)	.10 (.03)	.10 (.04)	.34 (.09)	.17 (.03)	.15 (.08)	.14 (.07)	.26 (.11)	Vigor
Dom/SE	S	.12 (.05)	.11 (.05)	.11 (.04)	.12 (.03)	.12 (.05)	.12 (.05)	.11 (.04)	.12 (.04)	.22 (.04)	.11 (.04)	.22 (.06)	.24 (.05)	.23 (.04)	.22 (.09)	.14 (.04)	.25 (.05)	
Trad. Values	N	.11 (.03)	.12 (.05)	.12 (.04)	.14 (.03)	.13 (.05)	.12 (.04)	.13 (.07)	.18 (.03)	.14 (.09)	.17 (.08)	.25 (.07)	.12 (.06)	.12 (.04)	.12 (.08)	.19 (.11)	.20 (.09)	
Work Orien./LOC	O	.21 (.03)	.20 (.04)	.20 (.04)	.21 (.03)	.19 (.04)	.20 (.04)	.20 (.05)	.20 (.05)	.24 (.01)	.19 (.04)	.21 (.04)	.23 (.04)	.19 (.05)	.23 (.04)	.22 (.10)	.24 (.01)	Motivation/ Stability
Coop./Stability	P	.12 (.04)	.12 (.04)	.12 (.03)	.14 (.04)	.14 (.03)	.13 (.03)	.13 (.02)	.20 (.04)	.21 (.12)	.13 (.03)	.11 (.04)	.21 (.11)	.13 (.03)	.24 (.08)	.22 (.09)	.24 (.01)	

Table 13

Median Validity Coefficients for Predictors of Performance

	Criterion			
Predictor	Job Knowledge	Task Performance	Global Ratings	Suitability
Aptitude Test	.14 - .58	.00 - .33	.12	.24
Biodata		-.13	.17	.29
Education	.13 - .17		.12	.36
Interest/ Attitude			.12	
Training Performance	.52	.23 - .40	.23	.29
Months on Job	.50	.43		
Concurrent Trait Rating			.71	
Age				.21
Total No. of Coefficients	110	18	225	42

From: Vineberg & Joyner, 1982.

Appendices A - L of

EXPERT JUDGMENTS OF PREDICTOR-CRITERION VALIDITY RELATIONSHIPS

are reproduced in ARI Research Note 85-14 (in press)

Covariance Analyses of Cognitive and Noncognitive Measures of Army Recruits: An Initial Sample of Preliminary Battery Data

Leaetta Hough and Marvin D. Dunnette
Personnel Decisions Research Institute

Hilda Wing
Army Research Institute

Janis Houston and Norman G. Peterson
Personnel Decisions Research Institute

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A791 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

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.

Presented at the Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Covariance Analyses of Cognitive and Noncognitive Measures in Army Recruits: An Initial Sample of Preliminary Battery Data

Since World War II, decisions regarding selection and classification of Army enlisted personnel have been based on the primary criterion of training performance and with cognitive abilities as predictors. Under the Army Research Institute's Project A (Improving the Selection, Classification, and Utilization of Army Enlisted Personnel) now underway (U.S. Army Reseach Institute 1983), the predictor-criterion space to be covered has been expanded to include noncognitive constructs of perceptual and psychomotor abilities, vocational interests and temperament. Project A is developing both new predictor and criterion measures and is improving existing ones.

The schedule for developing new predictors and relating them to new criteria has three phases. The first is centered about a Preliminary Battery (PB) of measures, administered to all entrants into four selected Army Military Occupational Specialties (MOS) during the first nine months of Fiscal Year 1984 (October 1983 - June 1984). The PB is composed of off-the-shelf paper-and-pencil measures and is designed to identify meaningful dimensions in the predictor space which are not currently included in Army selection and classification procedures. The scales include cognitive measures of spatial ability and reasoning as well as measures of vocational interest, background, and temperament. The second and third phases will build on the results of the first phase.

The selection of variables to include in the PB began with an exhaustive literature search to identify possible predictors. If a predictor appeared promising, information about it was summarized and assigned to an appropriate construct category according to a preliminary taxonomy. Third, each predictor was evaluated on several psychometric and substantive dimensions relevant to Army enlisted selection and classification, such as appropriateness of selected population and ease of administration. Predictors which were not really "off-the-shelf" were eliminated at this stage, as were those which could not be administered in a group, paper-and-pencil mode. Psychomotor measures fell into the latter category. Fourth, using information about available job performance criteria and of operational selection and classification instruments, the remaining predictors were compared with each other. The best measures covering the broadest range were selected for inclusion in the PB.

The data described in this report are from an initial sample from the larger group, those tested during the first two months of the nine-month data collection effort. Analyses were performed on the total subgroup, and of the four occupational clusters within the subgroup, to inform the selection and development of predictor measures in the second phase of Project A. Final decisions about these second phase predictor measures will be based on analyses of data from the total sample. This larger data set will include training criterion measures in addition to predictors; analyses of demographic subgroups (males, females, blacks, whites, MOS) will be possible because subgroup sizes will be sufficiently large. Such was not the case for the initial sample

described here. An interesting question will be how well analyses of the total sample, and its subgroups, will replicate the analyses reported here.

This data set will be useful in evaluating at least two sets of research questions. A first set is the extent to which prior findings are replicated. For example, Kass, Mitchell, Grafton, and Wing (1983) factor analyzed Armed Services Vocational Aptitude Battery (ASVAB) scores for over 98,000 Army applicants. The resulting four-factor solution included Verbal Ability, Speeded Performance, Quantitative Ability, and Technical Knowledge. Would these factors be found in the ASVAB scores of the initial sample? How would the cognitive measures of the PB relate to these factors? Another example concerns interest measures. Would a limited (four MOS) Army sample be sufficient to replicate a larger number of occupational scales? What evidence might be available to support Holland's (1973) six occupational themes (Realistic, Artistic, Conventional, Social, Enterprising, Investigative)? A final example concerns the biographical and temperament inventories: Would the predesignated scales be recoverable?

The greater importance of this research is, however, that it combines a broad range of cognitive or maximal performance measures with a broad range of noncognitive or typical performance measures. The research tradition has been to treat these two domains separately. The second set of research questions to be evaluated here concerns the covariation of these domains in a large, heterogeneous sample of young men and women.

Method

Research Participants

The population from which these examinees were selected consisted of those soldiers (recruits) who had entered active duty in the Regular Army and who began training in one of four MOS at one of five selected Army posts between October 1, 1983, and June 30, 1984, as follows:

MOS 05C: Radio Teletype Operator. Entered Advanced Individual Training (AIT) at Post A. Estimated population for PB: 2,150; 18% female.

MOS 19E/K: Tank Crewman. Entered One Station Unit Training (OSUT: Combined Basic Training or BT and AIT) at Post B. Estimated PB population: 2,720; 0% females in this combat occupation.

MOS 63B: Vehicle and Generator Mechanic. Entered AIT at Post C or Post D. Estimated PB population size: 4,065; 8% female.

MOS 71L: Administrative Specialist. Entered AIT at Post E. Estimated PB population: 4,575, 61% female.

The total population was initially expected to include approximately 13,500 soldiers: The number has since been revised downward to 11,000 to include approximately 3,500 cases each from 63B and 71L and smaller numbers from 05C and 19E/K. The initial subgroup consisted of those soldiers administered the PB from October 1 through December 1, 1984, whose records could be matched in our Longitudinal Research DataBase (LRDB; Wise, Wang, & Rossmeissl, 1983) constructed from existing Army files of admission and enlistment data.

Materials

ASVAB. Before entry into military service, each soldier had taken the Armed Services Vocational Aptitude Battery (ASVAB), a 3 1/2 hour test battery used for selection and classification by all the military services. Individuals who were examined at Military Entrance Processing Stations (MEPS) were given forms of the ASVAB with ten subtests. (A different version is used for a high school testing program; this data set did not include individuals taking this version.).) The following list of the ten subtests includes parenthetically the test acronym and the number of items.

Word Knowledge (WK:35), Paragraph Comprehension (PC:15),
Arithmetic Reasoning (AR:30), Numerical Operations (NO:50),
General Science (GS:25), Mechanical Comprehension (MC:25),
Math Knowledge (MK:25), Electronics Information (EI:20),
Coding Speed (CS:84), Auto-Shop Information (AS:25).

All but NO and CS are considered to be power tests; the two exceptions are speeded. Prior research (in Kass et al, 1983) has shown the reliability of the subtests to be within expectable limits for cognitive tests of this length (i.e., .78 - .92).

The participants had been selected for the Army based on their ASVAB scores as follows. All recruits had to achieve a minimum score on a composite known as the Armed Forces Qualification Test (AFQT), summed from scores on WK, PC, AR, and 1/2 NO. High school graduates had to be at or above the 21st percentile while nongraduates had to be at or above the 31st percentile, based on World War II norms. Second, each MOS has a specific composite on which a minimal score is required for entry. These composites, formed from sums of different combinations of from three to five subtest scores, are normed to the same World War II population and are highly correlated with AFQT scores in current forms of the ASVAB. For 19E/K and 63B, the cutoff was at the 26th percentile while for 05C and 71L the cutoff was at the 39th percentile.

Preliminary Battery (PB). The PB was administered to recruits during their first week of AIT following BT and a short leave, except that 19 E/K recruits were administered the PB before BT. The battery required about four hours to administer.

The PB included eight perceptual-cognitive measures (five from the Educational Testing Service or ETS French Kit (Ekstrom, French, & Harman, 1976), two from the Employee Aptitude Survey or EAS (Ruch & Ruch, 1980), one from the Flanagan Industrial Tests or FIT (Flanagan, 1965)) 18 scales from the Air Force Vocational Interest Career Examination (VOICE; Alley & Matthews, 1982); five temperament scales adapted from published scales (two from the Differential Personality Questionnaire or DPQ (Tellegen, 1982), one from the California Psychological Inventory or CPI (Gough, 1975), the Rotter I/E scale (Rotter, 1966)) and validity scales from both the DPQ and the Personality Research Form or PRF (Jackson, 1967); and Owen's (Owens & Schoenfeldt, 1979) Biographical Questionaire (BQ). The BQ could be scored for either eleven scales for males or fourteen for females based on Owen's research, or for 18 predesignated, combined sex scales developed for this research and called Rational Scales. The rational scales had no item on more than one scale; some of Owen's

scales included items on more than one scale. Items tapping religious or socio-economic status were deleted from Owens' instrument for this use, while items tapping physical fitness and vocational-technical coursework were added.

The scale names, with the number of items each included parenthetically, are as follows:

Perceptual-cognitive: ETS Figure Classification (FC: 28 items with 8 responses each); ETS Map Planning (MP: 40); ETS Choosing a Path (CP: 32); ETS Following Directions (FD: 20); ETS Hidden Figures (HF: 32); EAS Space Visualization (SV: 50); EAS Numerical Reasoning (NR:20); Flanagan Assembly (FNA: 20).

Vocational interests (VOICE): Office Administration (20); Heavy Construction (20); Electronics (20); Medical Service (20); Outdoors (15); Aesthetics (15); Mechanics (15); Food Services (15); Law Enforcement (15); Agriculture (15); Mathematics (12); Audiographics (10); Teacher/Counseling (10); Marksman (7); Drafting (7); Craftman (7); Automated Data Processing (7).

Temperament (Personnel Opinion Inventory or POI): Conscientiousness (DPQ Unlikely Virtues/PRF Infrequency: 10); Leadership (DPQ Social Potency: 26); Stress (DPQ Stress Reaction: 26); Discipline (CPI Socialization: 30); Motivation (Rotter I/E Locus of Control: 29).

Biographical Questionnaire (BQ): Scales for Males. Warmth of Parental Relationship (11); Academic Achievement (25); Social Introversion (22); Athletic Interest (10); Intellectualism (18); Aggressive/Independence (10); Parental Control vs. Freedom (11); Social Desirability (10); Scientific Interest (12); Academic Attitude (8); Sibling Friction (5).

Scales for Females. Warmth of Maternal Relationship (13); Social Leadership (22); Academic Achievement (13); Parental Control vs. Freedom (11); Cultural Literary Interests (5); Athletic Participation (9); Scientific Interest (13); Feelings of Social Inadequacy (3); Adjustment (5); Expression of Negative Emotion (4); Social Maturity (2); Popularity with Opposite Sex (4); Positive Academic Attitude (7); Warmth of Parental Relationship (5).

Rational (Combined Sex) Scales: Leadership (12); Social Confidence (4); Social Activity (11); Self Control (5); Antecedents of Self Esteem (6); Parental Closeness (13); Sibling Harmony (5); Independence (8); Academic Confidence (5); Academic Achievement (6); Positive Academic Attitude (6); Effort (4); Scientific Interests (5); Reading/Intellectual Interests (6); Athletic Interests (2); Athletic/Sports Participation (6); Physical Condition (18); Vocational-Technical Activities (4).

Analyses

Data editing procedures. The data were initially checked for consistency of Social Security Number, race, and sex within person across tests and inventories in the PB and between the PB and existing Army data files in the LRDB, which included ASVAB scores. Only those soldiers who had taken the currently operational forms of ASVAB had their data retained for these initial analyses. Those soldiers whose ASVAB scores of record were based on other forms (e.g., those used in the military high school testing program) were dropped from all subsequent analyses.

For the three noncognitive inventories there were several data quality

screens. For the VOICE and the BQ there was a three step process to eliminate records which contained too much missing data to yield interpretable scores. First, if the entire inventory had more than 10% of the items blank there were no scale scores generated for that inventory for that individual. Second, the remaining inventories were inspected scale by scale, and if more than 10% of the items on any given scale were blank, that scale score was set to blank. Finally, where data records contained only a few blanks (i.e., those which had fewer than 10% overall and fewer than 10% within scales) those blanks were set to (a) the "indifferent" response for VOICE or (b) the individual's mean response for that scale for the BQ.

The POI (temperament inventory) had a ten-item validity scale in addition to the two 10% missing data screens described above. The screens for this inventory were applied as follows. First, if the entire inventory had more than 10% of the items blank, there were no scale scores generated for that individual. Second, if the validity score scale for the individual was greater than 3, the inventory was deleted. Third, for each scale, if more than 10% of the items were blank, that scale was deleted. Last, of those records surviving the above screens but which contained a few blanks (i.e., less than 10%), each blank was set to the individual's mean response for that scale for the POI.

Thus, for item analyses purposes, the sample sizes varied across scales within inventory and across inventories.

Cognitive Measures: Item Analyses. For each of the eight cognitive measures in the PB, item validity was computed as the point biserial correlation between each item response and total score (number correct) on the test. Item difficulty was computed as the proportion of persons attempting the item who responded correctly.

Summary item statistics for each measure included the mean, standard deviation, and range for both item validities and item difficulties. Additionally, plots were prepared which showed the proportions of persons completing various proportions of items in each test, thereby illustrating the degree of speededness of each measure. Finally, mean test scores, standard deviations of scores, and Kuder-Richardson estimates (KR-20 and KR-21) of test reliability were calculated for each of the eight measures.

Cognitive Measures: Factor Analyses. The eight cognitive measures of the PB were intercorrelated with the ten ASVAB subtest scores. These intercorrelations were factor analyzed (principal factors) and rotated (varimax). The best fitting solution was selected.

Noncognitive Measures: Item Analyses. For each of the four temperament (POI) scales, 18 interest (VOICE) scales, and 18 rational biographical (BQ) scales, item validity was computed as the point biserial correlation between each item response and total (keyed) score. Summary statistics for each scale included means, standard deviations, sample sizes, median item-total correlations and ranges of item-total correlations. These statistics were examined to determine scale heterogeneity.

Noncognitive Measures: Factor Analyses. Item factor analyses were performed

for each of the three noncognitive inventories: 126 items of the BQ; 245 items of the VOICE; 121 items of the POI. Principal factors and varimax rotation were the procedures used. For the BQ, the mapping of the rational scales on the factors was evaluated. For the VOICE, two types of factor analytic solution were sought: The first was the possible mapping of the 18 VOICE occupational scales while the second was a possible mapping of the six Holland occupational themes. The POI factor analyses were to evaluate the coherence of the included four substantive scales with this subgroup of subjects.

Scale factor analyses were performed on the joint intercorrelations of the 18 VOICE scales, 18 BQ scales, and four POI scales.

Relationships between Cognitive and Noncognitive Measures. The ten ASVAB subtests, the eight PB cognitive tests and the 40 noncognitive scales were intercorrelated. For summary purposes, the ten ASVAB test scores were collapsed into four factor measures (Verbal:GS, PC, WK; Speed: CS, NO; Quantitative: AR, MK; Technical:AS, MC, EI). The noncognitive scales were collapsed into a five-factor (best-fit) solution. Then, for each of the twelve cognitive measures (four ASVAB factors and eight PB measures), the median intercorrelation across all scales included in each noncognitive factor was determined. The absolute values of correlations were used in this determination because the direction of any relation between cognitive and noncognitive measures may be quite arbitrary. These medians were inspected.

Descriptive Statistics by MOS Groups. For each of the four MOS groups, descriptive statistics were calculated for each of the eight PB cognitive tests and for five noncognitive factors. For this last analysis, three of the noncognitive factors were subdivided. Maximum effect sizes across the four MOS were calculated for each grouping of PB measures. Those PB measures with the largest maximum effect size across MOS have the greatest potential for differentiating among these four occupational groups.

Results

Research participants

Respondents consisted of 2,286 soldiers who had entered active duty in the Regular Army, who were in training in one of the four selected MOS at one of the five designated Army posts and who had been administered the PB between October 1 and December 1, 1984. Data for these examinees were checked for consistency of Social Security Number, race, and gender within person across tests and inventories of the PB and between the PB and existing Army data files as captured on the LRDB (ASVAB scores, etc.) A small number (151 or 7%) could not be matched to the LRDB and a somewhat larger but still expectable number (292 or 13%) were eliminated from further consideration here as the matched ASVAB form was not one of those currently operational. This loss rate was judged to be understandable and acceptable. The distribution of the 1,843 matched cases by sex within MOS is displayed in Table 1. Other data indicated the sample to be representative by race (63% white, 28% black, 5% Hispanic) as well as by gender but the numbers of cases were too low to permit reliable subgroup analyses.

Analyses

Data Editing Procedures. As indicated above, there was an initial 7% loss rate for failure to match existing data files and an additional loss of 13% for inappropriate ASVAB forms. It can be assumed that most of the latter group had the high school testing program form with a somewhat different and slightly larger number of subtests than the currently operational ASVAB, as described above. Such soldiers might be different from those entering via normal operating procedures; subsequent analyses of the complete PB sample as well as of subsequent Project A samples should provide more clarification.

The data quality screens applied to the noncognitive measures led to fairly small percentages of deletions from the matched sample of 1,843. For the BQ, 27 or 1.5% cases were deleted; 41 or 2.2% VOICE inventories were deleted; 140 or 7.6% POI cases were deleted, more on the basis of the validity score screen (105 or 5.7%) than on missing data only (35 or 1.9%). Such percentages are low for noncognitive measures administered to incumbents. They attest to the overall high quality of the PB administration at the five Army posts. The numbers above do not represent mutually exclusive groups of individuals but, rather, the maximum sample size available for analysis of any given scale within each inventory. For example, the maximum sample size for any scale on the BQ was 1,843 - 27, or 1,816. The smallest sample actually available for any BQ scale was 1,711; for the POI the figure was 1,697; for the VOICE it was 1,781.

Cognitive Measures: Item Analyses. Results of the item analyses for the eight cognitive tests are displayed in Table 2. The plots of the proportions of items completed against the proportion of persons completing indicated that most of these cognitive tests were speeded. For each test, the number of items completed by 80% of the group was determined. This datum permitted the estimation of the test time necessary for 80% to complete each test, included as the last column in Table 2. As can be seen, most of the tests had too stringent time limits, indicating that the Kuder-Richardson estimations of reliability may be inappropriate. The test of Following Directions (FD) appeared to have nearly adequate time limits while the Hidden Figures (HF) test appeared to have been too difficult rather than too speeded. For this test, some sizeable proportion of these examinees appeared to be unable to cope with the requirements of the embedded figures item format.

Cognitive Measures: Factor Analyes. The correlation matrix formed from the 18 cognitive tests (ten ASVAB and eight PB scores) was factored (principal factors) and rotated orthogonally (varimax). The correlations between the ASVAB and the PB tests are displayed in Table 3. The best factor analytic solution appeared to have five factors and is displayed in Table 4.

All of the PB tests have their highest loadings on the same first factor, a factor most clearly defined by tests presumed to be measuring the construct of Space Visualization (Space Visualization or SV; Flanagan Assembly or FNA) and Spatial Scanning (Map Planning or MP, Choosing a Path or CP). All of the ASVAB subtests have their highest loadings on one of the remaining factors, although the Mechanical Comprehension (MC) subtest also shows a high loading

on Factor 1. The pattern of loadings of the ASVAB subtests is very similar to the structure reported by Kass et al (1983).

The major conclusion is that the PB cognitive tests do not duplicate measures already contained in the ASVAB. This conclusion must be tempered somewhat by the fact that the unrotated factor solution yielded a large general factor with high loadings from both ASVAB and PB subtests. As can be seen in Table 3, the correlations among the two sets of subtests are generally positive and mostly moderate in size or larger.

Noncognitive Measures: Item Analyses. Means, standard deviations, sample sizes, median item-total correlations, and ranges of item-total correlations are presented in Table 5 for each noncognitive scale. An examination of the median item-total correlations indicates that the interest scales are most homogeneous: The median of these medians is .71 with a range of .57 to 83. Similar in level of scale homogeniety are the BQ rational scales: The median of the median item-total correlations is .68 with a range of .41 to .89. In comparison with these relatively high levels of scale homogeniety, the temperament scales have a median item-total scale correlation of .39 (.28 -.54), indicating that the items in each scale appear to be tapping more diverse areas. As will be shown in the item factor analyses of each of these three inventories, fewer factors (in comparison with the number of original scales) are needed to account adequately for the common variance in the interest and biographical item pools while more factors are required in the temperament item pool.

Noncognitive Measures: Factor Analyses. Item factor analyses of the Biographical Questionnaire (BQ) were based in the intercorrelations of the 126 BQ items with a principal factors approach rotated to simple structure (varimax). The 15-factor solution was most psychologically meaningful. Table 6 presents the factor names, defining items, factor loadings of the defining items and variance accounted for by the factors, as well as how the rational scales map onto the 15-factor solution. The 17 rational scales are quite similar to the factors. Most of the rational scales map onto their own unique factor although five of the factors have two or more rational scales. Three rational scales split into two factors but such splits appeared sensible. That is, Physical Condition splits into the factors Physical Ability and Exercise; Positive Academic Achievement splits into Academic Achievement and Intellectualism factors; Parental Closeness splits into Maternal and Parental Closeness factors. One factor had no counterpart in the rational scales. It appears to be the remnants of a socioeconomic status scale incompletely deleted here from Owen's original instrument.

Item factor analyses of the 245-item interest inventory (VOICE) yielded a 17-factor solution as the most psychologically meaningful. Table 7 presents the factor names, defining items, factor loadings of the defining items, variance accounted for by the factors, and how the VOICE scales map onto the factors. The VOICE scales are very similar to the factors found here. The Electronics and Mechanics scales of the VOICE merged to form a broad Construction/Repair factor. The VOICE Food Services scale became a Food and Clothing Preparation factor because the items "Sew clothes from patterns" and "Tailor" loaded on that factor rather than the Craftsman factor. The VOICE Craftsman

scale, in fact, was the only VOICE scale which did not remain essentially intact as items from this scale loaded more heavily on a number of other factors. For example, "Jeweler" loaded on the Office Work factor, "Shoe repairman" loaded on the Heavy/Physical Work factor, "Printer" loaded on the Food and Clothing factor.

Interest inventories are often described in terms of Holland's (1973) six occupational themes: Realistic, Investigative, Artistic, Conventional, Social, Enterprising. The six-factor (principal factors, varimax rotation) solution of the VOICE items was examined to determine how similar these six factors would be to Holland's six themes. Table 8 presents the factor names, defining items, and variance accounted for by the factors. Examination reveals that the forcing yields at best only a moderate relationship to Holland's six themes. The 17-factor solution presents factors which do seem to be more similar to the basic interest scales, however. The six-factor solution did provide the insight that the VOICE Craftsman scale items load on a factor made up of Domestic or what is often thought of as feminine interests. An inspection of the Craftsman items reveals that most of them (except for "Steam fitter," which loaded on the Realistic-Construction/Repair factor) consist of fine precision, detail work, such as "Jeweler," "Watchmaker," "Tailor."

Factor analyses of the 121 items of the temperament inventory (POI) indicated that the six-factor solution was the most meaningful. Table 9 presents the factor names, defining items, factor loadings of defining items, variance accounted for by the factors, and how the four temperament scales mapped onto the factors. A comparison of the scales and factors suggests that the Stress Reaction and Social Potency scales are each unidimensional while the other two are more heterogeneous. The Socialization scale split into three components: Emotional Closeness to Family and Friends; Rule-Making Behavior; items loading on the Stress Reaction factor. The Locus of Control scale split into a Just World/Nonrational World factor and a Predictable World/Personal Effort Makes a Difference factor. Prior research (Collins, 1974; Kaemmerer & Schwebel, 1976) had suggested that this latter scale might consist of four or five components. The items for one of these components, Belief in a Politically Responsive World, had not been included in this research. The remaining four components split into the two factors named above.

The item factor analyses indicated that most of the 40 scales included in the initial assembly of the PB had retained their integrity in the responses of this sample, so factor analyses of the 40 scales were completed. An item factor analyses of the combined 492 items would provide an arduous task of both computation and comprehension, a job better postponed until the complete sample of over 11,000 cases is available. The five factor solution for the scale factor analyses was most psychologically meaningful. Table 10 presents the factor names, scales loading on the factors, factor loadings of the scales, and the variance accounted for by the factors. Two interest factors emerged, a Realistic Interest factor and a Non Realistic Interest factor. The latter consisted of all the remaining (not Realistic) VOICE scales except for Science, which loaded on the third factor of Scientific/Intellectual Orientation. This third factor also included six biographical scales. The remaining two factors combined biographical and temperament scales, Potency-Athletic and Social and Personal Well-Being. These analyses suggest that interests should

be measured separately while biographical and temperament constructs can be combined and measured in a single inventory.

Relationships between Cognitive and Noncognitive Measures. The intercorrelation matrix of the ten ASVAB subtests and the eight PB cognitive measures with the 40 noncognitive scales indicated minimal overlap between the two domains. A summary of these intercorrelations is displayed in Table 11, which presents median intercorrelations (ignoring signs) of the cognitive measures (four ASVAB factor scores, eight PB tests) with the noncognitive domain (five factors). As can be seen, the three biographical/temperament factors show minimal overlap with the cognitive measures, ranging from .01 to .14 with the grand median of .05. There is somewhat more overlap between the two interest factors and the cognitive domain with a grand median of .095 and a range of .04 to .32. In general, however, the correlations are quite small, indicating the statistical as well as conceptual independence of the two domains.

Descriptive Statistics by MOS Groups. In this initial sample, only the dimension of MOS yielded sufficiently large subsample sizes for further investigation. Table 12 presents descriptive statistics by MOS for the eight PB cognitive measures while Table 13 presents these for ten combined noncognitive measures. Each table includes a bottom row labeled "Maximum effect size" which indicates how much the four MOS groups differ on the measure(s) in the associated column.

For the cognitive domain, three PB tests show greater potential for differentiating among the MOS, all three being measures of spatial visualization or scanning. In each case MOS 71L, Administrative Specialist, has the lowest score. Two of the other three MOS have the highest scores on these three tests, CP, FNA, and SV. These results are only suggestive, as the maximum effect sizes are insufficiently large to draw any definitive conclusion about the use of these cognitive tests in military placement.

In the noncognitive domain, the two interest factors appear to have great potential for differential placement among MOS. Again, MOS 71L is at the extreme: It shows the lowest average for the Realistic Interests factor and the highest for the Nonrealistic. To what extent this difference may be linked to the greater percentage of women in this MOS is not known. The third noncognitive factor, Scientific-Intellectual, shows the same differentiation as the ASVAB cut scores for these MOS: 71L and 05C have higher cutoff scores and also show higher averages on measures for this factor. Again, such results are merely suggestive and require the analyses of the complete data set for adequate verification.

Discussion

The factor structure of the 18 cognitive measures (including both the ten ASVAB subtests and the eight PB tests) confirm both prior factor analyses of ASVAB (e.g., Kass et al., 1983) and the independence of the PB tests. Measures of spatial visualization and spatial scanning most clearly define the common factor variance contained among the cognitive measures of the Preliminary Battery. While there is some overlap of these PB measures with the ASVAB, the data indicate that measures of spatial abilities could be profita-

bly investigated for their use in differential classification among MOS. Two of the PB tests, Hidden Figures and Figure Classification, showed low correlations across all ASVAB subtests. This may indicate their potential for providing variance unique from measures currently used, or may reflect important psychometric and/or conceptual difficulties of these measures. The data collected here do not permit any conclusion.

The item factor analyses of the noncognitive measures essentially confirmed the conceptual integrity of the scales which went into them. The biographical inventory split into factors easily mappable into the predetermined rational scales; the interest inventory split rather nicely into most of the 18 occupational areas for which it was constructed; the temperament inventory consisted of scales more multidimensional than those of the other two inventories, but still explicable. While the failure to retrieve the six Holland occupational themes from the VOICE was somewhat disappointing it was not surprising. Enlisted military occupations, from and for which the VOICE was developed, are not uniformly distributed over occupational theme space but tend to emphasize realistic endeavors.

Factor analyses of the 40 noncognitive scales showed that interests should be considered a separate domain but that biographical and temperament items are tapping essentially the same domain. The two interest factors split into two occupational theme areas: Realistic and Nonrealistic. The first biodata/temperament factor, Scientific-Intellectual Orientation, may be a bridge between the two clear interest factors and the two temperament factors of Potency and Personal Well Being.

Also, these data indicate minimal overlap between cognitive and noncognitive domains. What overlap is found makes conceptual sense, in that the interest and, possibly, the intellectual orientation factors show slightly more relation to traditional cognitive measures than do the two more strictly temperament factors. However, this is at best tentative. A more finegrained analysis of the total PB sample, estimated to exceed 11,000 cases, may provide more definition. It will be very important to specify hypotheses in detail and in advance. The large number of variables, even with this much larger sample, will provide massive opportunity for capitalization on chance.

One cannot overstress the limited nature of the valid inferences to be drawn from these analyses, primarily based on the limitations of this initial sample. Multivariate analyses of individual differences on such a large variety of measures requires even larger samples than this for adequate stability. While we anticipate that the findings reported here will replicate in the larger sample, there were many questions we did not ask. Comparison among the various demographic subgroups is an obvious omission but a totally appropriate one, we believe, based on subgroup sizes. Also curtailed were more detailed analyses of measure interrelationships because the item data on which such analyses are based would be insufficiently stable. Another concern is the characteristics of those soldiers who had to be excluded from the initial sample. We hope to explore such issues more completely with the complete sample.

All problems will not be eliminated with the larger group, however. The com-

plete sample will be restricted in range, in mostly unknown ways. The explicit selection on ASVAB will be reflected in the variances and covariances of the PB cognitive measures. And while soldiers are currently not selected either explicitly or implicitly on interests or on biodata/temperament measures to any great degree, it is not at all clear exactly what is the population of healthy young men and women to which we may safely generalize our results. While this sample of soldiers is heterogeneous across MOS there are other MOS, not included here, which might easily display different profiles of both cognitive and noncognitive measures, Subsequent phases of Project A research will add such different occupational groups to the research domain so that our inferences may be more broadly drawn.

These data and analyses have been useful, however. Based on them we are developing measures for the next phase of Project A, scheduled for a massive data collection effort during the summer of 1985. New cognitive measures are being tested which will, hopefully, avoid some of the psychometric difficulties of the PB measures which were not developed with the Regular Army soldiers in mind. An expanded version of the VOICE has been developed which will include more Army activities as well as provide more adequate coverage of the six Holland occupational themes. A combined biodata/temperament inventory has been developed to tap some of the important and heterogeneous dimensions in this domain. Final development and selection of measures for next summer's use will depend on more complete analyses of a more complete PB data set. The analyses reported here provided a very important head start on this process.

References

Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination. Journal of Psychology, 112, 169-193.

Collins, B. E. (1974). Four Components of the Rotter Internal-External Scale: Belief in a difficult world, a just world, a predictable world, and a politically responsive world. Journal of Personality and Social Psychology, 29, 381-391.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Flanagan, J. C. (1965). Flanagan Industrial Test Manual. Chicago: Science Research Associates.

Gough, H. G. (1975). Manual for the California Psychological Inventory. Palo Alto, CA: Consulting Psychologists Press.

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

Jackson, D. N. (1967). Personality Research Form Manual. Goshen, NY: Research Psychologists Press.

Kaemmerer, W. F. & Schwebel, A. I. (1976). Factors of the Rotter Internal-External Scale. Psychological Reports, 39, 107-114.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10: 1981 Army applicant sample. Educational and Psychological Measurement, 43, 1077-1088.

Owens, W. A., & Schoenfeldt, L. F.(1979). Toward a classification of persons. Journal of Applied Psychology Monographs, 64, 569-607.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80, (1, Whole No.609).

Ruch, F. L., & Ruch, W. W. (1980). Employee Aptitude Survey: Technical Report. Los Angeles: Psychological Services.

Tellegen, A. (1982). Brief Manual for the Differential Personality Questionnaire, Unpublished manuscript, University of Minnesota.

U.S. Army Research Institute (1983). Improving the Selection, Classification and Utilization of Army Enlisted Personnel. Project A: Research Plan. (Research Report 1332). Alexandria, VA: Author.

Wise, L. L., Wang, M., & Rossmeissl, P. G. (1983). Development and validation of Army selection and classification measures. Project A: Longitudinal research database plan (Research report 1356). Alexandria, VA: U.S. Army Research Institute.

Table 1. Numbers of Soldiers with Preliminary Battery Data According to MOS and Gender
[tested upon entry into Advanced Individual Training (AIT) for MOS]

MOS	N		
	Male	Female	Total
05C Radio Teletype Operator	144	28	172
19E/K Tank Crewman	309	0	309
63B Vehicle and Generator Mechanic	592	41	633
71L Administrative Specialist	351	378	729
TOTALS	1397	446	1843

Table 2. Item/Test Analyses Statistics for Eight Cognitive Measures of Preliminary Battery

Initial Sample (N=1,843)

Test	Mean	S.D.	KR20	KR21	Validity Mean	Validity S.D.	Validity Range	Difficulty Mean	Difficulty S.D.	Difficulty Range	No. Items	Time Limit	*Proposed Time
Numerical Reasoning (NR)	10.08	3.07	.77	.49	.40	.17	.08-.63	.67	.25	.23-.99	20	5	10
Spatial Visualization (SV)	22.60	10.19	.94	.90	.50	.13	.25-.70	.73	.14	.39-.95	50	5	10
Following Directions (FD)	6.01	1.91	.58	.38	.45	.12	.19-.60	.64	.24	.24-.99	10	7	8
Flanagan Assembly (FNA)	10.03	3.82	.80	.69	.45	.11	.16-.62	.61	.21	.22-.91	20	10	17
Map Planning (MP)	18.12	6.49	.91	.78	.42	.20	.13-.74	.81	.12	.57-.99	40	6	15
Choosing a Path (CP)	4.96	3.35	.81	.74	.51	.12	.20-.64	.53	.13	.28-.71	16	8	24
Hidden Figures (HF)	5.16	3.35	.77	.73	.47	.10	.22-.59	.52	.12	.26-.70	16	14	?
Figure Classification (FC)	52.72	15.12	.94	.89	.36	.15	.11-.66	.73	.22	.06-.98	112	8	18

*Estimated Time for 80% to Complete Test

Table 3. Correlations Betwen ASVAB Tests and Tests in the Preliminary Battery (N=1843)

	Preliminary Battery Test							
ASVAB Tests	CP	FD	FC	HF	MP	FNA	SV	NR
General Sciences (GS)	31	41	20	23	30	41	30	34
Arithmetic Reasoning (AR)	31	48	27	27	37	43	45	55
Word Knowledge (WK)	17	43	14	21	22	27	29	31
Paragraph Comprehension (PC)	18	38	19	17	25	24	27	29
Numerical Operations (NO)	-05	10	07	03	07	-07	-02	15
Coding Speed (CS)	-02	15	09	10	15	00	05	12
Auto/Shop Information (ASI)	37	21	18	15	29	43	43	22
Mathematics Knowledge (MK)	30	44	24	32	33	40	38	50
Mechanical Comprehension (MC)	47	35	30	29	41	57	56	38
Electronics Information (EI)	36	23	17	17	26	40	38	22

CP: Choosing a Path
FD: Following Directions
FC: Figure Classification
HF: Hidden Figures
MP: Map Planning
FNA: Flanagan Assembly Test
SV: Space Visualization
NR: Numerical Reasoning

Table 4. Rotated Orthogonal Factor Solution for Five Factors Extracted from 18 x 18 Correlation Matrix (ten ASVAB tests and eight preliminary battery tests)

(N=1843)

Test	I	II	III	IV	V	h^2
ASVAB General Science (GS)		67	38			69
ASVAB Arithmetic Reasoning (AR)	35	42			54	66
ASVAB Word Knowledge (WK)		77				66
ASVAB Paragraph Comprehension (PC)		66				50
ASVAB Numerical Operations (NO)				66		47
ASVAB Coding Speed (CS)				63		43
ASVAB Auto/Shop Information (AS)			61			63
ASVAB Mathematics Knowledge (MK)		41			52	64
ASVAB Mechanical Comprehension (MC)	48		54			68
ASVAB Electronics Information (EI)			64			60
PB Choosing a Path (CP)	56					41
PB Following Directions (FD)	42	41				42
PB Figure Classification (FC)	52					29
PP Hidden Figures (HF)	41					22
PB Map Planning (MP)	67					50
PB Planagan Assembly (FNA)	67					57
PB Space Visualization (SV)	73					63
PB Numerical Reasoning (NR)	47				42	48
Variance	3.29	2.40	1.60	1.17	1.01	

Factor loadings of .35 are shown in the above matrix. Decimals have been omitted.

Table 5. Item Analyses for TEMPERAMENT, INTERESTS, and BIOGRAPHICAL Scales

	Full Scale				Item Total Correlations	
Temperament	N	Mean	Standard Deviation	# Items in Scale	Median r	Range
Stress Reaction	1702	11.2	6.6	26	.54	.37 to .66
Social Potency	1703	11.8	5.8	26	.47	.33 to .65
Locus of Control	1698	18.6	3.3	29	.28	.02 to .43
Socialization	1697	19.1	4.1	30	.31	.04 to .50
Interests						
Electronics	1796	39.4	12.6	20	.77	.66 to .83
Science	1784	37.1	11.8	20	.71	.61 to .84
Medical Services	1801	35.8	10.9	20	.70	.41 to .78
Outdoors	1790	37.3	6.0	15	.59	.37 to .67
Audiographics	1801	22.0	5.3	10	.71	.53 to .79
Teacher/Counselor	1797	21.9	5.5	10	.72	.59 to .75
Marksman	1797	14.9	4.6	7	.77	.62 to .85
Drafting	1802	13.6	4.3	7	.78	.53 to .81
Craftsman	1801	10.3	2.9	7	.64	.50 to .73
Automated Data Processing	1801	14.8	4.2	7	.81	.60 to .83
Office Administration	1797	38.4	12.4	20	.77	.59 to .84
Heavy Construction	1802	34.8	11.0	20	.72	.56 to .79
Aesthetics	1797	27.4	7.9	15	.68	.46 to .75
Mechanics	1800	32.1	10.2	15	.83	.64 to .88
Food Service	1784	24.5	7.4	15	.69	.48 to .78
Law Enforcement	1799	28.7	6.8	15	.57	.30 to .71
Agriculture	1796	28.3	7.4	15	.61	.53 to .75
Mathematics	1781	22.4	6.7	12	.71	.50 to .80

Table 5 Continued

Biographical	Full Scale N	Mean	Standard Deviation	# Items in Scale	Item Total Correlations Median r	Range
Leadership	1813	22.6	6.7	12	.59	.38 to .78
Social Confidence	1804	12.7	2.9	4	.68	.40 to .79
Social Activity	1798	35.5	6.6	11	.51	.18 to .65
Self-Control	1805	17.5	3.3	5	.65	.57 to .66
Antecedents of Self-Esteem	1815	20.6	4.3	6	.68	.55 to .71
Parental Closeness	1804	39.5	9.1	13	.62	.36 to .71
Sibling Harmony	1793	15.5	4.1	5	.64	.63 to .72
Independence	1812	26.6	4.1	8	.46	.30 to .62
Academic Confidence	1796	15.5	3.4	5	.71	.61 to .77
Academic Achievement	1795	16.7	5.1	6	.74	.69 to .77
Positive Academic Attitude	1815	20.2	4.6	6	.71	.57 to .77
Effort	1816	12.5	2.9	4	.67	.51 to .69
Scientific Interest	1791	12.5	4.1	5	.72	.53 to .78
Reading/Intellectual Interests	1816	15.4	3.7	6	.58	.52 to .66
Athletic Interests	1807	7.2	2.2	2	.89	.88 to .90
Athletic/Sports Participation	1815	19.9	4.6	6	.70	.19 to .86
Physical Condition 1	1750	25.7	4.0	18	.41	.06 to .64
Vocational/Technical Activities	1711	9.9	3.5	4	.74	.44 to .84

TABLE 6

Factor Analysis
Biographical Questionnaire

	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR I: Physical Activity/Condition	6.85	11.2%
Defining Item: In the three months prior to joining the Army, how physically active were you? (-.71)		
Consists of 22 items, including 5 of the 6 items from Athletics/Sports Participation scale, 16 of the 18 items from Physical Condition scale, and one from the Leadership scale.		
FACTOR II: Leadership	5.57	9.1%
Defining Item: While in high school, how many times were you chairman/chairwoman of an important committee? (.68)		
Consists of 16 items, including 10 of the 12 items from the Leadership scale, 2 items from the Social Activity scale, one item each from the Scientific Interests, Reading/Intellectual Interests, and Athletic/ Sports Participation scales, plus one unscaled item.		
FACTOR III: Academic Achievement	4.94	8.1%
Defining Item: What was your approximate standing in your high school class? (-.67)		
Consists of 13 items, including 4 of the 6 items from the Academic Achievement scale, 3 of the 6 items from Positive Academic Attitude, 3 of the 4 items from Effort, one from Academic Confidence, one from Independence, plus one unscaled item.		
FACTOR IV: Maternal Closeness	4.43	7.3%
Defining Item: In high school, how close were you to your mother? (.79)		
Consists of 9 items, including 6 items from Parental Closeness, 2 from Antecedents of Self-Esteem, plus one unscaled item.		

TABLE 6Continued

Factor Analysis
Biographical Questionnaire

	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR V: Paternal Closeness	4.36	7.2%
Defining Item: In high school, how close were you to your father? (.84)		
Consists of 8 items, including 6 items from the Parental Closeness scale, one from the Antecedents of Self-Esteem scale, plus one unscaled item.		
FACTOR VI: Science Orientation	4.24	7.0%
Defining Item: How well do you think you did in biological sciences relative to other students with about the same ability at your high school? (.70)		
Consists of 9 items, 3 of the 5 items from Scientific Interests scale, 4 of the 5 items from Academic Confidence, and 2 remaining Academic Achievement scale items.		
FACTOR VII: Popularity	4.18	6.9%
Defining Item: How often did you go on dates during high school? (.59)		
Consists of 10 items, including 7 of the 11 items from the Social Activity scale, one from each of the Social Confidence and Independence scales, plus one unscaled item.		
FACTOR VIII: Adjustment	3.96	6.5%
Defining Item: During high school, how much did you wish you could become more socially acceptable? (.56)		
Consists of 16 items, including 4 of the 5 items from the Self-Control scale, 3 of the 4 items from Social Confidence, 2 items each from the Social Activity and Antecedents of Self-Esteem scales, one item from each of the Parental Closeness and Independence scales, and 3 unscaled items.		

TABLE 6 Continued

Factor Analysis
Biographical Questionnaire

	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR IX: Intellectualism	3.46	5.7%
Defining Item: In high school, how much did you enjoy discussion courses? (.61)		
Consists of 11 items, including 3 of the 6 items from the Reading/Intellectual Interests scale, 3 of the 6 items from Positive Academic Attitude, 2 items from the Independence scale, one item from each of Effort and Leadership, plus one unscaled item.		
FACTOR X: Parental Control	3.27	5.4%
Defining Item: In high school your parents were: very strict / strict / about average / lenient / very lenient (-.72)		
Consists of 7 items, including 3 of the items from the Independence scale, one item from the Self-Control and Antecedents of Self-Esteem scales, plus 2 unscaled items.		
FACTOR XI: Vocational/Technical Orientation	2.23	3.7%
Defining Item: How often have you repaired electrical or mechanical devices or machines in the past four years? (.65)		
Consists of 4 items, all the items from the Vocational/ Technical Activities scale, and one item from the Scientific Interest scale.		
FACTOR XII: Athletic/Sports Interests	1.84	3.0%
Defining Item: During high school how often did you watch each of the following types of television programs? (.57)		
Consists of 2 items, which are also all the items from the Athletic Interests scale.		

TABLE 6 Continued

Factor Analysis
Biographical Questionnaire

	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR XIII: Sibling Relationshio	1.61	2.6%
Defining Item: How much younger than you is your nearest younger brother or sister? (.60)		
Consists of 5 items, which is also the entire set of items from the Sibling Harmony scale.		
FACTOR XIV: Remainder of Items from Owens' Socioeconomic Status Factor	1.54	2.5%
Defining Item: When you were growing up, about how many books were around the house? (.33)		
Consists of 2 items. These are items that are part of a socioeconomic factor of Owens' original Biographical Questionnaire; however, the other items from this scale were eliminated from the preliminary battery.		
FACTOR XV: Moderate Exercise	1.15	1.9%
Defining Item: You got moderate exercise at work or school and some other exercise (sports, jogging, etc.) too. (.41)		
Consists of 2 items, both of which are from the Physical Condition scale.		
TOTAL	53.63	88.1%

NOTE:-45.8% of the variance in the matrix is common variance.

TABLE 7

Factor Analysis
Vocational Interest Career Examination

Factors	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR I: Construction/Repair	24.5	15.6
Defining Item: Repair small electric motors. (.79)		
Consists of 42 items, including all 20 items of the Electronics scale, all 15 items of the Mechanics scale, 5 of the 20 items of the Heavy Construction scale, one item from Automated Data Processing (Perform maintenance on a computer) and one item of the Marksman scale (replace defective parts on a rifle).		
FACTOR II: Office Work	18.6	11.8
Defining Item: Help prepare the payroll for a business. (.78)		
Consists of 28 items, including all 20 items of the office Administration scale, 3 Mathematics items, 2 Automated Data Processing items, 2 Law Enforcement scale items (investigate insurance claims and customs agent), and one Craftsman scale item (jeweler).		
FACTOR III: Science	14.6	9.3
Defining Item: Work in a scientific laboratory. (.81)		
Consists of 23 items, including all 20 items of the Science scale, plus one item from the Mathematics scale (use of a slide rule), one item from the Teacher/Counselor scale (solve problems by analyzing them logically) and one item from the Agriculture scale (experiment on plants with different types of fertilizer).		
FACTOR IV: Medical Service	9.7	6.2
Defining Item: Give injections to people for immunizations. (.72)		
Consists of 19 items, all of which are on the Medical Service Scale.		

TABLE 7 Continued

Factors	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR V: Heavy/Physical Work	9.1	5.8
Defining Item: Pour concrete for highway construction. (.67)		
Consists of 19 items, including 15 items of the 20 items on the Heavy Construction scale (the remaining five loaded on Factor I: Construction/ Repair), 2 items of the Craftsman scale (steam-fitter and shoe repairman), and 2 items of the Agriculture scale (drive a tractor on a farm and mow lawns, clip hedges, and bushes, and trim trees).		
FACTOR VI: Food and Clothing Preparation	7.8	4.9
Defining Item: Baker. (.73)		
Consists of 17 items, including all 15 items of the Food Service scale, plus 2 items from the Craftsman scale (sew clothes from patterns and tailor).		
FACTOR VII: Outdoors	6.4	4.1
Defining Item: Go canoeing. (.60)		
Consists of 17 items, including 14 of the 15 items of the Outdoors scale (exercising for physical fitness loaded on FACTOR XIII, Teacher/Counselor), plus 3 items of the Marksman scale (teach marksmanship, collect rifles and pistols, and belong to a gun club).		
FACTOR VIII: Aesthetics	6.1	3.9
Defining Item: Go to a symphony concert. (.70)		
Consists of 15 items, all of which are from the Aesthetics scale.		
FACTOR IX: Law Enforcement	5.4	3.4
Defining Item: Police Officer (.75)		
Consists of 13 items, all of which are from the Law Enforcement scale.		

TABLE 7 Continued

Factors	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR X: Audiographics	5.0	3.2
Defining Item: Photographer. (.69)		
Consists of 10 items, 9 of which are from the Audiographics scale, plus one item from the craftsman scale (printer).		
FACTOR XI: Agriculture	4.2	2.7
Defining Item: Gardener (.56)		
Consists of 11 items, all of which are from the Agriculture scale.		
FACTOR XII: Mathematics	4.2	2.7
Defining Item: Solve arithmetic problems (.68)		
Consists of 8 items, all of which are from the mathematics scale.		
FACTOR XIII: Teacher/Counselor	3.6	2.3
Defining Item: Teach someone to read. (.50)		
Consists of 10 items, 9 of which are from the Teacher/Counselor scale, and one of which is from the Outdoors scale (exercise for physical fitness).		
FACTOR XIV: Drafting	3.3	2.1
Defining Item: Draw Bridge blueprints. (.66)		
Consists of 7 items, all of which are from the Drafting scale.		
FACTOR XV: Computer Programming	2.6	1.6
Defining Item: Computer Programmer (.64)		
Consists of 4 items, all 4 of which are from the Automated Data Processing scale.		
FACTOR XVI: Uninterpretable	2.3	1.5
Defining Item: None		

TABLE 7 Continued

Factors	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR XVII: Craftsman	2.1	1.3
Defining Item: Gunsmith		
Consists of 2 items, both of which are from the Craftsman scale.		
TOTALS	129.5	82.4%

NOTE: 64.1% of the variance in the matrix is common variance.

TABLE 8

Six-Factor Solution
Vocational Career Examination

	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR I: Realistic-Construction/Repair	27.02	17.2%
Defining Item: Repair small electrical motors. (.79)		
Consists of 54 items, including all 20 items of the electronics scale, all 15 items of the mechanics scale, 15 of the 20 items of the Heavy Construction scale, and one item from each of the Audiographics, Marksman, Automated Data Processing, and Craftsman scales.		
FACTOR II: Investigative/Artistic	19.73	12.6%
Defining Item: Devise special scientific equipment for an experiment. (.75)		
Consists of 45 items, including all 20 items of the Science scale, 10 of the 15 items of the Aesthetics scale, 5 of the 7 items of the Drafting scale, 2 Audiographic items, 3 mathematics items,, 3 Teacher/Counselor items, one Agriculture item, and one Outdoors item.		

TABLE 8 Continued

	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR III: Conventional/Investigative	17.42	11.1%
Defining Item: Prepare a monthly financial statement. (.73)		
Consists of 39 items, including all 20 items of the Office Administration scale, 9 of the 12 items of the Mathematics scale, 6 of the 7 items of the Automated Data Processing scale, 3 of the Teacher/Counselor items, and one Law Enforcement item.		
FACTOR IV: Domestic and Skilled Crafts/Art	13.83	8.8%
Defining Item: Decorate cakes. (.69)		
Consists of 40 items, including all 15 items of the Food Service scale, 6 of the 7 items from the Craftsman scale, 6 of the 10 items on the Audiographics scale, 5 items from the Aesthetics scale, 5 items from the Agriculture scale, one item each from the Drafting, Outdoors, and Teaching/ Counseling scales.		
FACTOR V: Realistic-Outdoors/Nature/Adventure	13.17	8.4%
Defining Item: Fight a forest fire. (.62)		
Consists of 41 items, including 13 of the 15 items of the Outdoors scale, 11 of the 15 items of the Law Enforcement scale, 9 of the 15 items of the Agriculture scale, 6 of the 7 items on the Marksman scale, and 5 Heavy Construction items.		
FACTOR VI: Investigative	11.70	7.4%
Defining Item: Take human blood samples. (.65)		
Consists of 26 items, including 19 of the 20 Medical Service items, 3 Teaching/Counseling items, 3 Law Enforcement items, and one Audiographics item.		
TOTALS	102.93	65.5%

NOTE: 64.1% of the variance in the matrix is common.

TABLE. 9 .

Factor Analysis
Personal Opinion Inventory

	Variance Explained	Amount of Common Variance Accounted For
FACTOR I: Stress Reaction	8.62	28.7%
Defining Item: I often find myself worring about something. (.65)		
Consists of 36 items, including all 26 items from the Stress Reaction scale, 8 items from Socialization, and 2 items from Locus of Control scale.		
FACTOR II: Social Potency	5.51	18.3%
Defining Item: I am very good at influencing people. (.64)		
Consists of 27 items, including all 26 items from the Social Potency scale and one item from the Socialization scale.		
FACTOR III: Emotional Closeness to Family and Friends	2.16	7.2%
Defining Item: My home life was always very pleasant. (.61)		
Consists of 7 items, all of which are from the Socialization scale.		
FACTOR IV: Just World/Nonrational World	2.12	7.0%
Defining Items: People's misfortunes result from the mistakes they make. (.42) Without the right breaks one cannot be an effective leader. (.38)		
Consists of 16 items, 14 of which are from the Locus of Control scale, and 2 of which are from the Socialization scale.		
FACTOR V: Rule Abiding	2.09	6.9%
Defining Item: In school I was sometimes sent to the principal for cutting up. (.43)		
Consists of 9 items, all of which are from the Socialization scale.		

TABLE 9 Continued

Factor Analysis
Personal Opinion Inventory

	Variance Exolained	Amount of Common Variance Accounted For
FACTOR VI: Predictable World/Personal Effort Makes A Difference	2.01	6.7%
Defining Item: Becoming a success is a matter of hard work, luck has little or nothing to do with it. (.50)		
Consists of 16 items, 13 of which are from the Locus of Control scale and 3 of which are from the Socialization scale.		
TOTAL	22.51	74.8%

NOTE: 25% of the variance in the matrix is common variance.

TABLE 10

Five-Factor Solution of the
Interest, Biographical, and Temperament Scales

	Variance Accounted For	Percent of Common Variance Accounted For
FACTOR I: Non-Realistic Interests	4.93	24.2%
Consists of the following scales in order of factor loading magnitude:		
(I) Teacher/Counselor (.74)		
(I) Office Administration (.74)		
(I) Medical Services (.72)		
(I) Food Service (.67)		
(I) Aesthetics (.67)		
(I) Craftsman (.66)		
(I) Audiographics (.59)		
(I) Mathematics (.53)		
(I) Automated Data Processing (.50)		
(I) Drafting (.37)		
FACTOR II: Realistic Interests	4.34	21.3%
Consists of the following scales in order of factor loading magnitude:		
(I) Mechanics (.82)		
(I) Heavy Construction (.80)		
(I) Marksman (.78)		
(I) Electronics (.73)		
(I) Outdoors (.67)		
(B) Vocational/Technical Activities (.59)		
(I) Agriculture (.55)		
(I) Law Enforcement (.42)		
FACTOR III: Scientific/Intellectual Orientation	3.46	17.0%
Consists of the following scales in order of factor loading magnitude:		
(B) Scientific Interest (.74)		
(B) Academic Confidence (.69)		
(I) Science (.60)		
(B) Academic Achievement (.56)		
(B) Positive Academic Attitude (.46)		
(B) Reading/Intellectual Interests (.41)		
(B) Effort (.41)		

Note: I signifies interest scale
B signifies biographical scale
T signifies temperament scale

TABLE 10 Continued

		Variance Accounted For	Percent of Common Variance Accounted For
FACTOR IV: Potency Athletic and Social		2.84	13.9%
Consists of the following scales in order of factor loading magnitude:			
(B)Athletic/Sports Participation (.76)			
(B)Athletic Interests (.60)			
(B)Social Activity (.58)			
(B)Leadership (.56)			
(B)Physical Condition (.55)			
(T)Social Potency(.40)			
(B)Independence (.31)			
FACTOR V: Personal Well-Being		2.39	11.7%
Consists of the following scales in order of factor magnitude:			
(B)Antecedents of Self-Esteem (.73)			
(T)Socialization (.66)			
(B)Self-Control (.61)			
(B)Parental Closeness (.60)			
(T)Stress Reaction (.43)			
(B)Social Confidence (.35)			
(B)Sibling Harmony (.27)			
(T)Internal-External Locus of Control (.15)			
	TOTALS	18.0	88.1%

NOTE: 67.9% of the varinace in the matrix is common variance.

TABLE 11

Median[1] Correlations Between Cognitive and Noncognitive Measures

Cognitive Measure	Noncognitive Measures: Realistic Interests	Non-Realistic (Other) Interests	Scientific/ Intellectual Orientation	Potency (Athletic & Social)	Personal Well-Beir
ASVAB: Verbal Ability	.07	.05	.14	.10	.05
ASVAB: Speeded Performance	.21	.13	.09	.05	.03
ASVAB: Quantitative Ability	.05	.06	.14	.10	.05
ASVAB: Technical Knowledge	.32	.16	.09	.04	.05
Figure Classification	.07	.07	.04	.01	.02
Map Planning	.13	.12	.05	.05	.03
Hidden Figures	.04	.04	.07	.04	.02
Following Directions	.05	.05	.06	.10	.03
Choosing a Path	.18	.13	.05	.04	.02
Spatial Visualization	.21	.16	.05	.07	.07
Numerical Reasoning	.06	.05	.06	.05	.04
Assembly Test	.19	.14	.07	.09	.05

[1]The absolute values were used to compute the median correlations; i.e., the signs of the correlations were ignored.

Table 12 Means and Standard Deviations of PB Tests According to MOS

	Preliminary Battery Test							
05C: Radio Teletype Operator (N=172)	CP	FD	FC	HF	MP	FNA	SV	NR
M	5.5	6.5	54.1	4.7	18.5	11.2	22.8	11.0
SD	3.5	1.9	15.6	3.1	6.2	3.5	9.8	2.8
19E/K: Tank Crewman (N=309)								
M	5.7	5.9	52.6	5.9	19.1	10.8	25.9	10.6
SD	3.5	2.0	16.3	3.7	6.4	3.8	9.6	3.1
63B: Vehicle & Generator Mechanic (N=633)								
M	5.6	5.9	54.2	5.1	18.8	10.8	24.5	9.9
SD	3.5	1.9	14.0	3.3	5.8	3.6	9.2	3.0
71L: Administrative Specialist (N=729)								
M	3.9	6.0	51.2	5.1	17.0	8.8	19.5	9.8
SD	2.8	1.9	15.4	3.3	7.0	3.8	10.5	3.2
Maximum Effect Size	0.51	0.31	0.20	0.36	0.33	0.65	0.65	0.40

CP: Choosing a Path
FD: Following Directions
FC: Figure Classification
HF: Hidden Figures
MP: Map Planning
FNA: Flanagan Assembly
SV: Space Visualization
NR: Numerical Reasoning

TABLE 13

Differential Usefulness of Non-Cognitive Measures for Placement Decisions

	I Realistic Interests	II Non-Realistic Interests	III Scientific Intellectual		IV Potency Athletic/Social			V Personal Well Being		
	1	2	3	4	5	6	7	8	9	10
05C Radio/Teletype (N = 170)										
M	33.8	36.2	16.1	18.0	7.5	35.6	23.4	20.5	18.6	12.3
SD	10.6	10.4	3.4	5.2	2.3	6.4	6.9	4.3	4.3	6.7
19 E/K Tank Crewman (N = 305)										
M	37.9	32.9	15.3	15.5	7.2	34.8	22.4	19.7	18.9	9.4
SD	9.4	10.0	3.2	4.7	2.1	6.8	6.4	4.2	4.0	6.5
63B Vehicle Mechanic (N = 625)										
M	41.2	32.2	15.0	15.2	7.4	36.4	21.4	20.8	18.7	11.3
SD	9.7	10.4	3.4	4.8	2.1	6.3	6.1	4.1	4.2	6.5
71L Administrative Specialist (N = 695)										
M	27.9	46.9	16.0	18.2	7.0	35.0	23.6	20.7	19.7	11.7
SD	8.6	10.4	3.4	5.1	2.4	6.6	7.0	4.5	4.1	6.6
Maximum Effect Size	1.39	1.43	.32	.61	.22	.25	.33	.26	.27	.44

Note: Numbered columns refer to the following variables:

1 = Heavy Construction Interest
2 = Office Administration Interest
3 = Academic Confidence
4 = Academic Achievement
5 = Athletic Interest
6 = Social Activity
7 = Leadership
8 = Antecedents of Self Esteem
9 = Socialization
10 = Stress Reaction

Meta-Analysis: Procedures, Practices, Pitfalls
Introductory Remarks

Hilda Wing
Army Research Institute

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A791 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

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.

Presented at the Annual Convention of the American Psychological Association, Toronto, Ontario, Canada. Chair, Symposium on Meta-Analysis: Procedures, Pitfalls.

Meta-Analysis: Procedures, Practices, Pitfalls

Introductory Remarks

Welcome to the Division 5 Symposium on Meta-Analysis: Procedures, Practices, Pitfalls. The term "meta-analysis" is less than a decade old (Glass, 1976) yet the processes it describes, for combining the results of research from different studies, have quickly come to be considered "an innovation whose time has come" with "extensive publications" (Schoenfeldt, 1984, p. 79). Most members of this audience are probably aware of this brief meteoric history but many of you may be skeptical about how much and how well meta-analyses can deliver. The purpose of this symposium is to raise such concerns, explore and evaluate them, so that we all may have a more accurate understanding of what meta-analysis techniques can and cannot accomplish.

Procedures for making statistical inferences from two or more data sets were available before the birth of meta-analysis. I followed such procedures myself (Wing, 1982) in 1975 as an inhouse expert witness for the U.S. Civil Service Commission in an adverse impact charge brought against the White House Fellows program. Data for selection rates were available for 1973, 1974, and 1975; part of my testimony was based on combining these three data sets. For each year considered separately, adverse impact (against women) was not statistically significant although the probabilities were low (.17, .25, .07). To estimate the probability for the three years combined, I followed procedures suggested by Winer (1981) and came up with an overall probability level between .10 and .20, which was nonsignificant. Pooling the data into one large data set, as my colleague Frank Schmidt suggested, does some injustice to the actuality that each year's candidates were competing only against each other. It also reduced the probability somewhat, to .07. Still nonsignificant.

Plaintiff's experts used a different procedure to combine the separate probabilities of these independent events: They multiplied them. This led to an overall probability level of .005 which is statistically significant, and also totally wrong. Our resident intelligent layman, the Commission's General Counsel, was quick to determine this. He noted that if one flips a coin 120 times, the probability of 60 or fewer heads is .50. One can also consider this as six groups of 20 coin tosses each, however, each with a separate probability of 10 or fewer heads. Following the same logic as the plaintiff's experts, multiplying the six separate probabilities, leads to an overall probability of less than .05. We referred to this lapse as an example of the gambler's fallacy.

I would hope that the recent explosion in meta-analysis research would serve to eliminate such errors which at the very least are embarrassing to those who catch themselves in them. However, meta-analysis has its own hazards for the unwary, and we hope to identify many of them today.

I would like to give much of the credit, and none of any possible blame, for this symposium to Tom Cook, who provided many ideas and impetus. He was a great help.

The first speaker, Laurel Oliver, will describe the general methodology of meta-analysis. Ken Pearlman will provide a more detailed description of one specific and powerful form of meta-analysis, validity generalization. Penny Hauser-Cram will discuss the relative strengths and weaknesses of non-quantitative and quantitative or meta-analytic procedures for combining research results. Norm Miller will present a case study of meta-analytic procedures which may prove to be a classic: The NIE-sponsored research on the effects of school integration on the achievement of black children. Finally, Bob Linn will make sense of it all.

Thank you for coming.

References

Glass, G. V. Primary, secondary and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Schoenfeldt, L. F. The status of test validation research. In B. S. Plake (Ed.). Social and technical issues in testing: Implications for test construction and usage. Hillsdale, NJ: Lawrence Erlbaum Associates, 1984, 61-86.

Winer, B. J. Statistical procedures in experimental design (2nd. ed.) New York: McGraw-Hill, 1971.

Wing, H. Statistical hazards in the determination of adverse impact with small samples. Personnel Psychology, 1982, 35, 153-163.

ARI Technical Report 648*
VERBAL INFORMATION PROCESSING PARADIGMS.
A REVIEW OF THEORY AND METHODS

Karen J. Mitchell
(September 1984)

The theory and research methods of selected verbal information processing paradigms are reviewed. Work in factor analytic, information processing, chronometric analysis, componential analysis, and cognitive correlates psychology is discussed. The definition and measurement of performance on verbal test items and test-like tasks is documented. Portions of the reviewed verbal processing paradigms are synthesized and a general model of text processing presented. The model was used as a conceptual framework for subsequent analyses of the construct and predictive validity of the verbal subtests of the Armed Services Vocational Aptitude Battery (ASVAB) 8/9/10.

---

* In press. To be available from the Defense Technical Information Center, 5010 Duke Street, Alexandria, VA, 22314. Phone: (202) 274-7633. The paper was published in the FY83 annual report (ARI Research Note 83-37) prior to publication as a Technical Report.

IV. VALIDATION

Paul G. Rossmeissl and Lauress L. Wise

During Project A's second year, the Longitudinal Research Database (LRDB) was expanded dramatically to provide a firm basis for validation research. The first major validation research effort was carried out using information on existing predictors and criteria in the expanded LRDB. The initial validation research led to proposed improvements in the Army's existing procedures for selecting and classifying new recruits. The proposed improvements were adopted after thorough review and are to be implemented at the beginning of FY85. In addition, a number of smaller research efforts were supported with the expanded LRDB.

In describing validation research results during FY84, we turn first to an overview of the growth of the LRDB. Next, we summarize the ASVAB Aptitude Area Composite research that was based on the expanded LRDB. We conclude with a brief desription of other supporting analytic activities.

Growth of the LRDB

FY84 saw three major LRDB expansion activities. These were:

- The expansion of the FY81/82 cohort data files.
- The establishment of the FY83/84 cohort data files.
- The addition and processing of pilot and field test data files for different predictor and criterion instruments.

Each of these activities is described briefly.

Expansion of the FY81/82 Cohort Data Files. During FY83, we had accumulated application/accession information on all Army enlisted recruits who were processed in FY81 or FY82, and we had processed data from Advanced Instructional Training (AIT) courses on their success in training. During FY84, we added SQT data providing information on the first-tour performance of these soldiers subsequent to their training. SQT information was found for a total of 63,706 soldiers in this accession cohort, notwithstanding the fact that many of the soldiers in this cohort were not yet far enough along to be tested in this time period and others were in MOS which were not tested at all during this period.

In addition to SQT information, administrative information from the Army's Enlisted Master File (EMF) was added to the FY81/82 data base. Key among the variables culled from the EMF were those describing attrition from the Army, including the cause recorded for each attrition, and those describing the rate of progress of the remaining soldiers. Records were found for a total of 196,287 soldiers in this cohort. While the major source of administrative information was the FY83 year-end EMF files, information on progress and attrition was added from March and June 1984 quarterly EMF files.

Establishment of the FY83/84 Cohort Data Files. During FY84, application and accession information was assembled on recruits processed during FY83 and FY84. This cohort is of particular importance to Project A since it is the cohort to be tested in the concurrent validation effort. In addition to accession information, administrative data on the progress of this cohort also were extracted from annual and quarterly EMF files.

With the FY83/84 cohort, we began to include data collected on new instruments developed by Project A. Preliminary Test Battery information was collected on more than 11,000 soldiers in four different military occupational specialties. For three of these specialties (05C/31C, Radio/Teletype Operator; 71L, Administrative Clerk; and 63B, Light Wheel Vehicle Mechanic), data were collected at the beginning of AIT. In the fourth MOS (19E/K, Armor Crewman), data were collected at the beginning of combined Basic and AIT, generally within the first two weeks after accession. Data collected on these soldiers are described in Hough et al. (see Chapter III).

During FY84 we also collected data on success in AIT for soldiers in four MOS to which the Preliminary Battery was administered. At the end of FY84, data were still being added on soldiers who had taken the Preliminary Battery at the beginning of their training. The data collected included both written and hands-on performance measures administered at the end of individual modules as well as more comprehensive end-of-course measures. Table 6 shows the number of soldiers for whom Preliminary Battery information is available, the number of soldiers for whom training performance information is available, and the number of soldiers for whom both types of information are available.

Creation of Pilot and Field Test Data Files. During FY84, a great deal of information was collected in conjunction with the development of new instruments to be used in the FY85 concurrent validation. The largest accumulation of such information resulted from the Batch A combined criterion field test. (Batch A refers to the first four MOS of the nine MOS for which comprehensive performance measures are being developed.) In this effort, 548 soldiers in four different MOS each completed 2.5 days of testing. The tests administered included hands-on performance tests, job knowledge tests (both the task-specific version and the comprehensive tests being developed for use during training), and a wide range of rating data. (See Chapter II.) The combined information led to over 3,000 analysis variables for each of the soldiers tested.

A second major field test effort during FY84 was the Pilot Trial Battery field tests. These tests included both paper-and-pencil measures of aptitudes, interests, and background and the new computerized battery of perceptual and psychomotor tests. Scheduling conflicts postponed the data collection effort until the very end of the fiscal year, so initial processing of these data has only begun.

In addition to the major field tests of predictor and criterion instruments, data from a number of other efforts were incorporated into the LRDB. These included ratings of task and item importance, pilot tests on

Table 6. FY83/84 Soldiers With Preliminary Battery and Training Data

MOS	TOTAL PB CASES	TOTAL* TRAINING CASES	TOTAL CASES WITH BOTH PB & TRAINING DATA		
				%PB	%TR
05C/31C	2,411	1,971	833	(37)	(45)
19E/K	2,617	2,749	1,809	(69)	(66)
63B	3,245	1,959	1,223	(38)	(62)
71L	3,039	4,654	2,079	(68)	(45)
Total	11,312	11,313	5,944		

*As of FY84 year-end.

trainees of the comprehensive job knowledge tests intended for training use, and data gathered during the exploratory round of utility workshops.

ASVAB Area Composite Validation

As a first step in its continuing research effort to improve the Army's selection and classification system, Project A completed a large-scale investigation of the validity of Aptitude Area Composite tests used by the Army as standards for the selection and classification of enlisted personnel. This research had three major purposes: to use available data to determine the validity of the current operational composite system, to determine whether a four-composite system would work as well as the current nine-composite system, and to identify any potential improvements for the current system.

The Armed Services Vocational Aptitude Battery (ASVAB) is the primary instrument now used by the Armed Services for selecting and classifying enlisted personnel. The ASVAB is composed of ten cognitive tests or subtests, and these subtests are combined in various ways by each of the services to form Aptitude Area (AA) Composites. It is these AA composites that are used to predict an individual's expected performance in the service. The U.S. Army uses a system of nine AA composites to select and classify potential enlisted personnel: Clerical/Administrative (CL), Combat (CO), Electronics Repair (EL), Field Artillery (FA), General Maintenance (GM), Mechanical Maintenance (MM), Operators/Food (OF), Surveillance/Communications (SC), and Skilled Technical (ST).

The criterion measures used as indices of soldier performance in these analyses were end-of-course training grades and SQT scores. While both of these measures have some limitations, they were the best available measures of soldier performance. These two criterion measures were first standardized within MOS, and then combined to form a single index of a soldier's performance in his or her MOS.

One unique aspect of the composite development research was the large size of the samples used in the analyses. The sample sizes in the validity analyses for each of the AA composites are shown in Figure 19. The total sample size of nearly 65,000 soldiers renders this research one of the largest (if not the largest) validity investigations conducted to date.

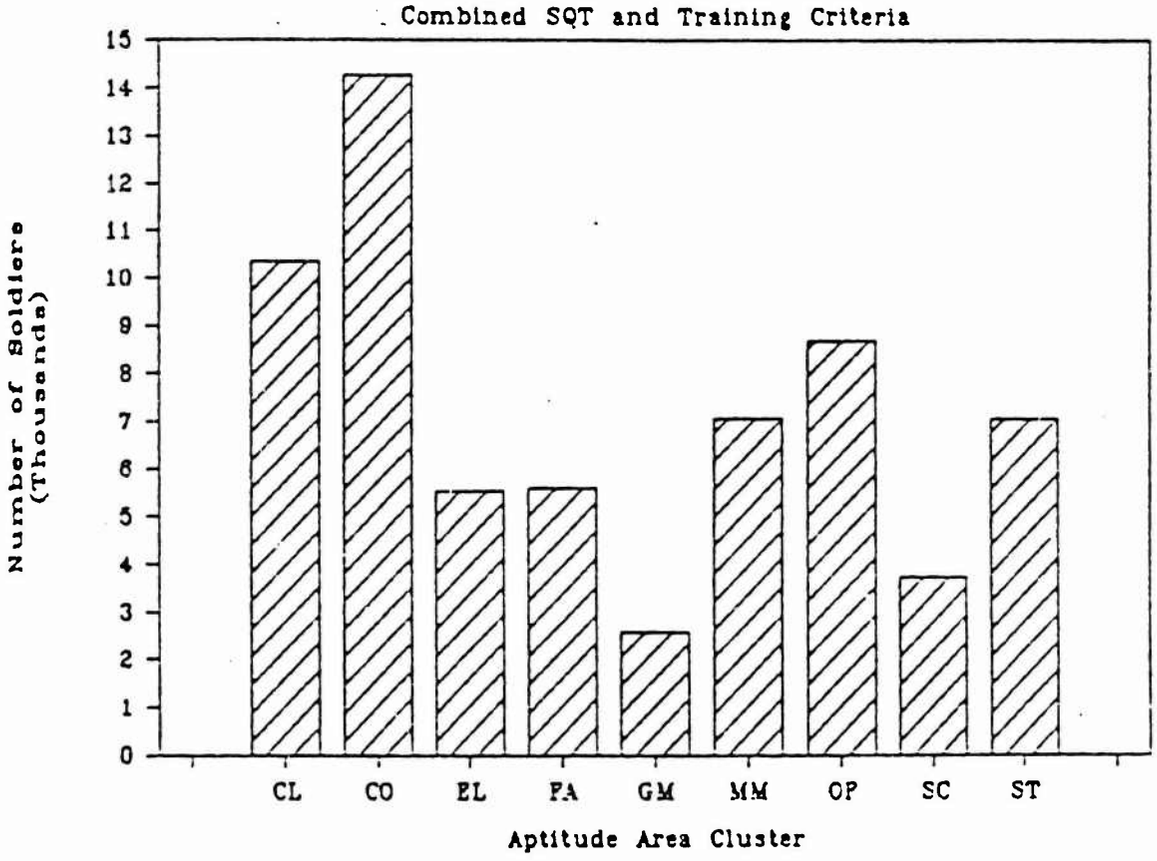


Figure 19. Validity Analyses Sample Sizes

The validities obtained in this research for the current nine AA composites are given in Figure 20. As can be seen, the existing composites are very good predictors of soldier performance. The composite validities ranged from a low of .44 to a high of .58, with the average validity being about .48. These numbers are high as test validities go.

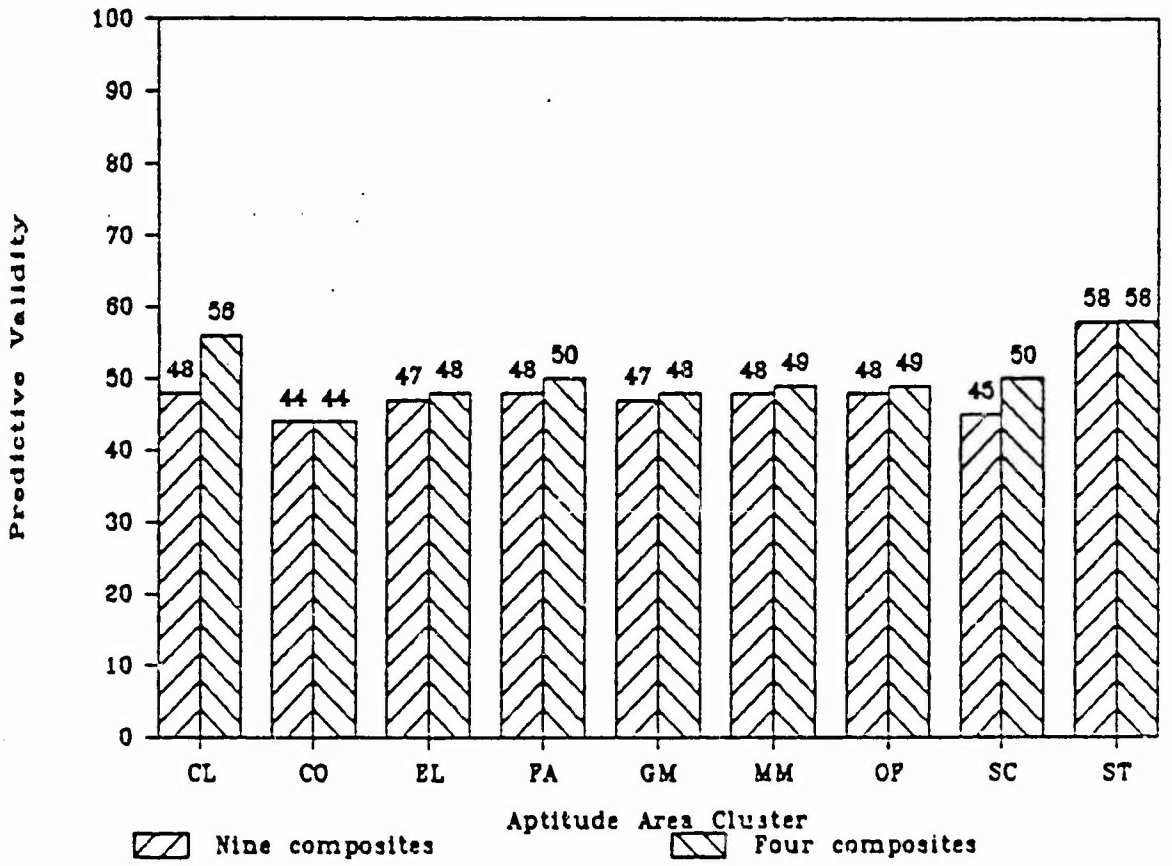


Figure 20. Predictive Validities Systems for Nine and Four Composites

A second finding of this research was that despite the high validities of the existing composites, a set of four newly defined AA composites could be used to replace the current nine without a decrease in composite validity. This set of four alternative composites included: a new composite for the CL cluster of MOS; a single new composite for the CO, EL, FA, and GM MOS clusters; a single new composite for the GM, MM, OF, and SC MOS clusters; and a new composite for the ST cluster of MOS. Figure 20 also shows the test validities (corrected for range restriction) for this four-composite system when it is used to predict performance in the nine clusters of MOS defined by the current system. In all cases the four-composite solution showed test validities equal to or greater than the existing nine-composite case.

A corollary finding of the investigation into the four-composite solution was that the validities for two of the nine composites could be substantially improved without making major changes to the entire system. This improvement was accomplished by dropping two speeded subtests (numerical operations and coding speed) from the CL and SC composites and replacing them with the arithmetic reasoning and mathematical knowledge subtests for the CL composite and the arithmetic reasoning and mechanical comprehension subtests for the SC composite. Figure 21 compares the old and new forms for the CL and SC composites. This simple substitution of different subtests was able to improve the predictive validity of the CL composite by 16 percent and of the SC composite by 11 percent.

Based upon these data the Army has decided to implement the proposed alternative composites for CL and SC, effective 1 October 1984. Using the techniques developed by Hunter and Schmidt (1982) (which assume that an

	Current Composite		Proposed Composite	
Clerical/Administrative MOS	(VE+NO+CS)	.48	(VE+AR+MK)	.56
Surveillance/Communications MOS	(VE+NO+CS+AS)	.45	(VE+AR+MC+AS)	.50

Figure 21. A Comparison of Current and Alternative Composites

individual's salary provides an approximation of that individual's worth to the organization), it can be estimated that these changes could lead to increased performance in the CL and SC MOS worth approximately $5 million per year. A fuller discussion of the research entailed in the development and validation of the AA composites can be found in McLaughlin, Rossmeissl, Wise, Brandt, and Wang (1984).

LRDB Support Activities

The expanded LRDB was also used in support of a number of other analytic activities. One such activity was the creation of an initial workfile containing Preliminary Battery data from tests administered through December 1983. Analyses based on this file were used to inform the development of the Trial Battery as well as to preview results for the Preliminary Battery.

EMF information being added to the LRDB was also used in support of ARI efforts to analyze the effects of alternative criteria for second-tour reenlistment eligibility.

A number of analysis files were provided to ARI staff in support of in-house research. These include a MAP data workfile, a Transportation School criterion data workfile, SQT information for addition to cohort files, and a workfile containing data from the Work Environment Questionnaire.

Associated Reports and Papers

We have divided Project A reports and papers associated with validation into three categories. Those dealing with operational research activities are presented first, those dealing with methodological issues are presented second, and a paper dealing with utility is presented last.

Reports and papers dealing with operational research activities

Six reports dealing with specific aspects of the ASVAB validation process were issued during Project A's second year. This included preparation as an ARI Technical Report of the comprehensive report on the validation analyses for the FY81/82 enlisted accessions.

(1) An evaluation of the ASVAB 8/9/10 Clerical (CL) composite for predicting performance in training, prepared in FY83 by Weltin and Popelka, was issued as ARI Technical Report 594. The composite showed high validity (r = .68) as a predictor, but an alternative version had even higher validity (r = .74).

(2) The factors that enter into the grouping of military occupations into clusters for prediction purposes were considered by Wise, McLaughlin, Rossmeissl, and Brandt. An initial investigation of several alternative clustering algorithms was carried out, using the Skill Qualification Test scores as the criterion.

(3) The validity of the ASVAB composites in connection with the assignment of soldiers to specific MOS was analyzed by McLaughlin. The differential validity was estimated for (a) unconstrained assignment, using a formula solution, and (b) assignment constrained by Army operational considerations (i.e., a certain number of recruits are needed for each MOS, and the MOS compete with each other for the more highly qualified applicants), using a representative assignment procedure. Current composites performed less well than the alternatives in which fewer composites were used.

(4) Two uses of repeated replication methods to assess the stability of sample statistics in ASVAB validation work were described in a paper by Brandt, McLaughlin, Wise, and Rossmeissl.

(5) Several analyses aimed at determining possible subgroup bias were conducted by Rossmeissl and Brandt for the current and proposed alternative ASVAB Aptitude Area composites. Both sets of composites showed small differences in predictive validity as a function of race or gender, but it was judged that either set could be used to select and classify enlisted personnel without resulting in increased bias against blacks or women.

(6) The results of the validation of current and alternative ASVAB composites, based on training and SQT information for FY81/82 enlisted accessions, are being presented in ARI Technical Report 651 by McLaughlin, Rossmeissl, Wise, Brandt, and Wang (in press).

ARI Technical Report 594*
EVALUATION OF THE ASVAB 8/9/10 CLERICAL COMPOSITE
FOR PREDICTING TRAINING SCHOOL PERFORMANCE

Mary M. Weltin and Beverly A. Popelka
(October 1983)

The composite of Armed Services Vocational Aptitude Battery (ASVAB) subtests used to select applicants for entry-level training in Army clerical schools was evaluated by correlating composite scores with training performance scores. The clerical composite (CL) had high validity (r=.68) for this criterion, but an alternate composite of Arithmetic Reasoning, Paragraph Comprehension, and Mathematics Knowledge scores produced from multiple regression analyses had even higher validity (r.=.74). Differential prediction for classification purposes is discussed.

---

* Available from Defense Technical Information Center, 5010 Duke Street, Alexandria, VA 22314. Phone: (202) 274-7633. Order Document No. ADA143235. The report was included in the FY83 annual report (ARI Research Note 83-37) prior to publication as a Technical Report.

BLANK PAGE

Clustering Military Occupations in Defining Selection and Classification Composites

Lauress L. Wise and Donald H. McLaughlin
American Institutes for Research

Paul G. Rossmeissl
Army Research Institute

David A. Brandt
American Institutes for Research

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A791 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

The work described here was conducted under contract No. MDA903-82-C-0531 to the U.S. Army Research Institute for the Behavioral and Social Sciences.

Paper presented at the Annual Convention of the American Psychological Association in Toronto, Canada.

CLUSTERING MILITARY OCCUPATIONS IN DEFINING SELECTION AND CLASSIFICATION COMPOSITES

The work presented here was part of a larger effort to investigate the validity of the Armed Services Vocational Aptitude Battery (ASVAB) for selecting and classifying potential Army recruits. The ASVAB is a written test battery that includes ten subtests. These ten subtests are frequently divided into four groups based on the results factor analyses. These groupings are:

- Verbal Skills and Knowledge
  1. Paragraph Comprehension* (PC)
  2. Word Knowledge* (WK)
  3. General Science (GS)

- Mathematical Skills and Knowledge
  4. Arithmetic Reasoning (AR)
  5. Mathematics Knowledge (MK)

- Technical Skills and Knowledge
  6. Mechanical Comprehension (MC)
  7. Auto Shop Information (AS)
  8. Electronics Information (EI)

- Processing Speed
  9. Coding Speed (CS)
  10. Numerical Operations (NO)

*These two subtests are frequently combined into a single measure called Verbal Skills (VE).

The military currently uses a combination of four of these subtests as an overall determinant of qualification for military service. This composite, known as the Armed Forces Qualifying Test (AFQT), combines WK, PC, AR, and NO, giving only one-half weight to NO. In addition to AFQT, each service uses additional ASVAB composites to qualify recruits for particular military occupational specialties (MOS). For nearly all of the Army MOS, a soldier must have a passing score on one (or in a few cases 2) of nine different ASVAB composite scores. These composites were developed in a previous analyses of the relationship of of ASVAB scores to criterion measures (Maier & Grafton, 1981). They are:

• Clerical/Administrative	CL	VE+NO+CS
• Combat	CO	AR+CS+AS+MC
• Electronics Repair	EL	GS+AR+MK+EI
• Field Artillery	FA	AR+CS+MK+MC
• General Maintenance	GM	GS+AS+MK+EI
• Mechanical Maintenance	MM	NO+AS+MC+EI
• Operators/Food	OF	VE+NO+AS+MC
• Surveillance/Communications	SC	VE+NO+CS+AS
• Skilled Technical	ST	GS+VE+MK+MC

Each of these nine composites may be thought of as identifying a cluster of occupational specialties for which a common predictor of future performance is used. Setting aside those few specialties, such as Army Band, that require auditions rather than passing composite scores, we have an effective partitioning of all Army MOS into nine different clusters. One part of our investigation was to examine alternative clusterings of the entry level Army MOS for purposes of defining common predictor composites. This paper describes the steps taken in identifying alternative MOS clusters. In particular, results of comparisons among different clustering algorithms are reported.

DATA

The persent validation was based on information on the 274,220 recruits who entered by Army between 1 October 1980 and 30 September 1982 (FY81 and FY82). ASVAB scores, along with basic demographic information, were assembled and edited for all of these recruits. In addition, subsequent performance data were collected for samples of these recruits. The analyses reported here are part of a much larger effort to develop an improved predictor battery and improved measures of job performance. The analyses reported here were necessarily restricted to the use of available predictor and criterion measures. The results provide a baseline against which improvements due to better measurement instruments can be gauged.

Two basic kinds of performance information were collected. The first consisted of test scores from advanced instructional training (AIT) courses. Data for 172 MOS trained at 23 different schools were collected by the Army Research Institute during 1981. The type of test score varied qualitatively between courses, but most frequently reflected a trend toward criterion-referenced testing during training. This means that the scores did not always show great variability and were frequently negatively skewed indicating ceiling effects. (After all, if the course was successful, everyone should pass.) In

analyzing training information, "training cells" were defined by individual MOS courses for which the score variables were comparably scaled. We used cells where scores on at least 100 different soldiers were available. There were 98 such cells, covering 88 different MOS.

The second type of criterion information was Skill Qualification Test (SQT) scores. Since 1977, the Army has administered SQTs to enlisted soldiers to assess individual qualifications for promotion and to evaluate the overall effectiveness of Army training programs. Each year, a separate SQT is constructed for each MOS and skill level (with the exception of a very few exempted MOS). In some cases, alternative forms are constructed for the same MOS and skill level corresponding to different "tracks" within that MOS. An SQT "track" corresponds to a specialization within an MOS, most commonly to specialization on a particular type of equipment.

The SQT data used here consisted of written tests designed to assess a soldier's knowledge of a sample of the tasks listed in the Soldier's Manual for that soldier's specialty. For each task, a number of multiple choice items (from 2 or 3 up to 9 or 10) were constructed. An overall total score is computed averaging the percentage of items passed for each task. "SQT cells" were defined for each different SQT form used during 1982 and 1983 testing. A total of 113 cells containing at least 100 observations were analyzed. These covered 68 different MOS.

In the overall validation effort, we analyzed training and SQT measures separately and then performed a "combined" analyses, pooling analyses across all training and SQT cells for a particular MOS. In this paper, we focus on an initial investigation of alternative clustering algorithms which was carried out using the SQT data as criterion measures. The SQT data were used since there were significantly more data points and there was believed to be significantly fewer distributional problems in comparison to the training data.

METHOD

Criterion for the Evaluation of Alternative Cluster Solution

The first step in the identification of alternative MOS clusters was a careful definition of the goal of this activity. The reasons for limiting the number of different predictor composites, and hence the reasons for identifying clusters of MOS that will use common composites, flow from operational

constraints. Army systems for scoring the ASVAB, counseling potential recruits, and maintaining score records are all designed around the current limited set of predictor composites. While it might be possible to alter current systems to accommodate a much larger number of predictor composites up to a separate composite for each of the approximately 250 entry-level occupations, the cost of such changes would be very significant. In fact, one goal of the overall effort reported here is to determine whether an increased number of predictor composites would lead to any significant improvement in subsequent occupational performance. Conversely, we were also seeking to determine whether the Army could use a reduced number of separate predictor composites without significiant loss in prediction accuracy.

For each occupational specialty for which performance measures were available, regression analysis can be used to identify an "optimum" composite of the ASVAB subtests for predicting those measures. We began by estimating such optimal regression functions. In doing so, we computed ridge regression coefficients which have been found to be better estimators of optimum prediction coefficients for the whole population rather than computing ordinairy least-squares (OLS) estimates which give optimum prediction coefficients for the analysis sample. Predictions generated from the ridge regression equations then represented the best that the ASVAB could do in predicting performance.

The correlation between the predicted scores for two different specialties is one measure of the extent to which a single composite could be used for both specialties without losing predictive information. If the predictions for two specialties were perfectly correlated, then a single composite could be used without losing any predictive information. Note that this is true even if the proportion of total variance accounted for in performance measures for the two specialties is quite different. In the present context we can only be concerned with those aspects of job performance that are predictable from the ASVAB. For this reason, the correlation of predicted scores was viewed as an appropriate measure of MOS similarity in evaluating the effects of clustering. (In any event, we have no basis for estimating correlation of the performance measures that is independent of ASVAB predictions, since the available performance measures are on different individuals for the different MOS.)

We chose to use the correlation rather than the squared correlation. Any two separate prediction functions can be

partitioned into a common and unique part. The correlation between the two original functions is the product of the correlations of each of these functions with the underlying "common" function. Assuming appropriate standardization, the observed correlation is, in fact, the square of the correlation of each observed measure with the underlying common measure. The simple correlation between the observed measures thus gives the measure of variance retained in substituting the common prediction function for the two individual functions.

In evaluating clusters containing more than two specialties, we chose to examine the average correlation of the predicted scores for each pair of MOS in the cluster. This measure was chosen primarily for reasons of computational efficiency over the primary alternative which was the sum of squared correlations of each separate prediction function with a single common function such as the first principal component of the set of separate predictor functions defined by the cluster. The latter definition (subtracted from the number of MOS in the cluster) gives an exact measure of the loss of the prediction function variance accounted for when the common function is substituted for each of the separate functions. This latter measure is still not fully precise because of a desire to weight each MOS by some measure of size and by the proportion of total performance variance accounted for in deriving a single common composite. In the special case where the common function is equally correlated with each of the separate functions, this latter measure and our proposed measure are identical. Otherwise, they will still be very closely related given the very restricted dimensionality of the present predictor battery.

In the final definition of the criterion used for an initial evaluation of alternative clustering solutions, each of the corelations being averaged was weighted by the sum of a size measure for the two MOS involved. The size measure used for each MOS was the total number of enlistments into the MOS for the two years on which the sample was based (FY81 and FY82). The use of a size measure reflected the view that a fixed degree of loss in predictive power would be more serious if a large number of classification decisions were involved than a smaller number of decisions were involved.

Clustering Algorithms Examined

In general, we considered algorithms that operated on similarity data rather than distance data. Most of the available procedures are step-wise, leaf-to-stem procedures. This means

that they begin with each entity in a separate cluster and then continue to combine two clusters at a time until all entities are in a single cluster. At each stage, the two clusters combined satisfy some criteria for maximizing within cluster similarity. The three common criteria are known as single-linkage which is based on the maximum similarity (or minimum distance) between each pair of clusters, complete-linkage which examines the minimum similarity (or maximum distance) between each pair of clusters, and average linkage which examines the average similarity between each pair of clusters. Prior investigations (e.g., Milligan, 1980) have generally found the average linkage approach to be somewhat more effective.

Three available procedures were examined. These include BMDP1M from the BMDP package, PROC VARCLUS from the Statistical Analysis System (SAS), and the OCLINK subroutine from the IMSL library. These three routines represent a good deal of variation among available approaches. PROC VARCLUS (SAS Institute, Inc., 1982) was the only stem-to-leaf approach examined. At each step, the cluster with the greatest heterogeneity is identified and split into two separate clusters. The method used, a type of oblique component analysis, is a form of factor analysis.

BMDP1M implements a leaf-to-stem approach designed by Hartigan (1975). We used the average linkage rule. The OCLINK routine from IMSL is similar, except that here a complete linkage rule was used.

In addition to these available procedures, we programmed and investigated two additional approaches. The first, WTDLINK, is identical to the BMDP1M approach, except that each entity is weighted by a size measure when computing combined similarities. The second approach, labelled CONLINK, is identical to WTDLINK except that the hierarchy of clusters is constrained to pass through a particular set of clusters. In this case, we constrained the results to include the nine MOS clusters defined by the current ASVAB composites. This means that solutions involving more than nine clusters reflected further partitioning of the existing nine clusters while solutions with fewer than nine clusters reflected a combining of the current nine clusters. One reason for this approach was the belief that the current clusters reflect commonalities that are otherwise obscured in our analyses due to imcomplete coverage of the criterion space. Another was that, if combinations of the current nine clusters did not perform appreciably worse than other clustering solutions, it would greatly simplify the problem of assigning predictor composites to those specialties where adequate criterion information was not available (i.e., we would keep such MOS together with their current clustermates).

One additional issue to be investigated is the similarity of results from alternative procedures independent of the overall evaluation criterion. In assessing similarity, we used a symetric information measure that reflects the extent to which knowing where each MOS is classified in one set of clusters reduces the "uncertainty" as to the assignment of MOS to some other set of clusters. This measure is defined as

$$.5*(U[x]+U[y]-U[xy])/(U[x]+U[y])$$

where U[x] "uncertainty" for a set of categories defined as the sum over the categories of p*ln(p) where p is the proportion of entities in the category and ln(p) is the base 2 logarithm of p. Here U[x] corresponds to one set of clusters and U[y] to the other. U[xy] corresponds to the set defined by the cross-tabulation of the two categories. This measure has the value of 1.0 if there is a perfect correspondence between the two cluster sets and a value of 0.0 if there is total independence.

RESULTS

For each clustering method, a hierarchy of clustering solutions was obtained. In investigating the results, we chose to focus initially on solutions with 20 or fewer clusters. In every instance, improvements in the criterion became negligible with far fewer clusters. For PROC VARCLUS, we only examined solutions up to eight clusters, the default stopping point of this algorithm. Table 1 shows the weighted average within cluster similarity (predicted score correlation) for each solution and method. The average similarities shown as entries in this table are all exceedingly high. In the end, we determined that the differences in similarity were not sufficiently stable to yield any significant cross-validation beyond the two or three cluster solutions (see Brandt et al., one of the companion papers). Nonetheless, several conclusions can be drawn about the results of the different algorithms for a fixed set of similarities. Except for the VARCLUS procedure, which is affected by the absolute size of the correlations, the results would be comparable for any set of similarities linearly related to those used.

The first conclusion to be drawn from the information in Table 1 is that the weighted linkage procedure (WTDLINK) does, in fact, lead to the best solutions when a weighted average criterion is used. The cluster sets resulting from this procedure have uniformly higher weighted average within cluster similarity in comparison to each other procedure. Thus, the use

of weights in clustering and in the evaluation of clusters did make some difference.

The second finding was that the VARCLUS routine produced better solutions for small numbers of clusters than did either of the two unweighted leaf-to-stem procedures. This result may result from the fact that step-wise procedures tend to show greater departures from optimality in later steps than they do in earlier steps. VARCLUS is at the beginning of its stepwise iterations for the smaller number of clusters while BMDP1M and OCLINK are at the end of a rather large number of steps.

A third conclusion is that the differences between BMDP1M and OCLINK did not make appreciable differences for these data. There were small differences favoring the average linkage approach (BMDP1M) for solutions with larger numbers of clusters, but not for most of the solutions presented here.

A final conclusion is that the constrained solutions, while showing uniformly poorer fit in comparison to the WTDLINK procedure, gave results that were not appreciably worse than the BMDP1M or OCLINK solutions in many cases. Predictably, the CONLINK procedure performed worst in comparison to the other procedures near the point of maximum constraint, the nine-cluster solution (which was totally fixed in advance). Even at this point, however, the results show an improvement over random assignment (as evidenced by the one cluster solution).

Table 2 shows the similarity of the solutions resulting from the different procedures. The entries are information measures described above. At each level, as defined by the number of clusters in the solution, the BMDP1M and OCLINK procedures gave the most similar results. This was not particularly surprising since both algorithms were leaf-to-stem procedures and neither differentially weighted the MOS being clustered. By the time they both reached the numbers of clusters indicated, there were a few large clusters and a number of clusters with individual outliers. The three-cluster solutions produced by these two algorithms were identical with two clusters each consisting of a single MOS and the third cluster consisting of all other MOS.

What was somewhat more surprising, was that the relatively high similarity of the solutions form WTDLINK and VARCLUS. These two procedures were quite different with VARCLUS being a stem-to-leaf decomposition algorithm and WTDLINK being a leaf-to-stem composition algorithm. Further, VARCLUS did not weight the MOS differentially while WTDCLUS did. The reason for their greater similarity was that neither paid as much attention

to the smaller more "deviant" MOS in comparison to the other procedures. The three-cluster solution from VARCLUS had 48, 31, and 34 MOS and the corresponding clusters from WTDLINK contained 62, 32, and 19 MOS, respectively. Further study of this property would be useful.

The final point to be made from examiniation of the similarity of the solutions of the different algorithms is that the constrained solutions generated by CONLINK were more similar to the WTDLINK and VARCLUS solutions than they were to the other solutions. One of the nine constrained clusters, Field Artillery, tended to stay by itself, while the remaining eight clusters sorted themselves into two groups.

CONCLUSION

In our current investigation, lack of stability in the similarity measures led us to abandon the attempt to cluster MOS on a purely empirical basis. We were further constrained in the present context by a desire to identify unit-weight composites for reasons of computational efficiency and because unit-weight prediction equations typically show less shrinkage on cross-validation. In the end, we also performed a more-or-less manual search for an optimal clustering of the existing nine clusters using a measure of loss of variance accounted for through substitution of the best unit-weight composite for each cluster. In future investigations, with more complete criterion information and a broader predictor base, we will develop an automated version of this search procedure. In addition, we want to expand the definition of similarity of predictor measures to include the similarity of subgroup (minorities and women) differences in the predictive relationships.

TABLE 1

Weighted Average Within Cluster Similarities
by Method and Number of Clusters

No. of Clusters	METHOD				
	VARCLUS	BMDP1M	OCLINK	WTDLINK	CONLINK
1	.916	.916	.916	.916	.916
2	.932	.922	.922	.934	.922
3	.936	.927	.927	.943	.922
4	.944	.928	.928	.945	.931
5	.949	.929	.929	.950	.929
6	.952	.931	.930	.953	.929
7	.954	.931	.931	.958	.926
8	.954	.933	.933	.959	.923
9	.	.933	.933	.963	.925
10	.	.937	.935	.964	.927
11	.	.937	.936	.964	.929
12	.	.937	.938	.964	.938
13	.	.937	.938	.964	.934
14	.	.937	.938	.964	.934
15	.	.937	.938	.964	.937
16	.	.937	.939	.964	.941
17	.	.939	.939	.964	.941
18	.	.939	.940	.964	.946
19	.	.954	.940	.964	.947
20	.	.954	.941	.964	.946

TABLE 2

Agreement Between Alternative Clustering Solutions
As Measured by Reduction in Uncertainty

First Method	Second Method	Number of Clusters			
		3	6	9	12
WTDLINK	CONLINK	.127	.184	.240	.333
	VARCLUS	.328	.392	.466	.456
	BMDP1M	.060	.104	.207	.287
	OCLINK	.060	.096	.197	.256
CONLINK	VARCLUS	.189	.181	.222	.263
	BMDP1M	.017	.076	.165	.185
	OCLINK	.017	.071	.125	.197
VARCLUS	BMDP1M	.031	.193	.304	.321
	OCLINK	.031	.174	.232	.320
BMDP1M	OCLINK	.000	.745	.628	.584

REFERENCES

Dixon, W. J., & Brown, M. B. (1979). BMDP-79 biomedical computer progress P-score. Berkeley, CA: University of California Press.

Hartigan, J. A. (1975). Clustering algorithms. New York: John Wiley & Sons, Inc.

IMSL, Inc. (1980). ISML LIBRARY Reference Manual. Houston: ISML, Inc.

Maier, M. H., & Grafton, F. C. (1981). Effectiveness of selection and classifiction testing (Research Report 1179). Arlington, VA: U.S. Army Research Institute for the Behaviorial and Social Sciences.

Milligan, G. W. (1980). "An examination of the effect of six types of error perturbation on fifteen clustering algorithms." Psychometrika, 45, 325-342.

SSAS Institute, Inc. (1982). SAS user's guide: Statistics, 1982 edition. Cary, NC: SAS Institute, Inc.

Differential Validity of ASVAB for Job Classification

Don McLaughlin

American Institutes for Research

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A791 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Presented at the Annual Convention of the American Psychological Association, in Toronto, Canada.

The overall performance of the Army depends on how well the skills of recruits can be matched to the requirements of the MOS they enter. Therefore, a set of composites must be evaluated in terms of its differential validity.

In theory, the differential validity of a set of composites is based on the correlation of the best predictor of differences between MOS with the actual differences that one would observe. The practical problem is that, in general, it is infeasible to collect criterion data from the same individual in all jobs. One cannot ordinarily observe the criterion needed for estimating differential validity. Fortunately, Horst (1954) developed a method for measuring the crucial part of the differential validity of a test battery without the necessity of these observations.

Actual use of a set of composites for classification of recruits into MOS is a complex process, however, and an abstract measure of differential validity can only approximate the relative value of one set of composites, compared to another. A more accurate comparison of composites would involves simulation of the constrained assignment process, and work is progressing on the development of the appropriate simulation algorithm. In the present report, we have estimated differential validity both for the case of unconstrained assignment, using the procedure outlined by Horst (1954), and for the case of constrained assignment using a representative assignment algorithm.

Unconstrained Assignment, Formula Solution

The starting point for this measurement is the work of Horst (1954). Horst demonstrated that one could compute the ordinary least squares (OLS) linear predictor of the difference between two criteria for an individual without actually having measurements of both criteria for any single individual. Using this result, he proposed a Classification Efficiency index equal to the average (over all pairs) of the variances of the predictors of the differences. In addition, he showed that this index could be elegantly represented in terms of the variances and covariances of the predictors of single criteria.

The formula for Horst's Classification Efficiency index which we used is:

$$(1) \qquad H^2 = \text{Average}(\hat{y}_{ik} - \hat{y}_{jk})^2/2,$$

where the $\hat{y}_{ik}$ and $\hat{y}_{jk}$ are the OLS estimates of standardized criteria i and j for individual k, and the average is over all i, j, and k, such that i does not equal j.

Horst pointed out the problem in using H as a direct measure of differential validity; namely, that the maximum value it can take on, if the predictors are perfectly accurate, is not unity. The maximum value is 1 minus the average intercorrelations among the criteria. Unfortunately, these intercorrelations cannot be measured without observing multiple criteria for single individuals. However, because the intercorrelations of criteria will

be the same no matter what the predictors, the values of H for different sets of predictors of the same criteria can be compared.

Brogden (1959) proposed a measure of differential validity similar to the measure derived by Horst. Brogden's measure, which we shall call D, is the product of (a) the average absolute predictive validity of the predictors and (b) the square root of one minus the average of the intercorrelations of the predictors. When these intercorrelations are equal, it is related to Horst's measure by the following equation:

$$(2) \quad H^2 = D^2 + (1 + g/(p-1)) \times \text{Variance(validity coefficients)},$$

where the variance is between criteria,
g is the intercorrelation of the predictors, and
p is the number of predictors.

That is, when all the criteria are equally well predicted by the composites, H is equal to D, and in any other case, H includes a component of differential validity due to the variation in predictability of the criteria.

A corollary of equation (2) is that a battery can possess differential validity, as measured by H, even though the predictors are all perfectly correlated with each other. In that case, D is equal to zero, but H can be greater than zero, if some criteria are more predictable than others. Thus, as pointed out by Maier (1982), examination of the intercorrelations of predictors is insufficient for estimation of the differential validity of a battery.

Although it seems counter-intuitive at first that a set of perfectly correlated composites could possess differential validity, the following example makes clear that they can. Consider the case of a single composite. Suppose that the composite measures a set of skills that account for much of the variation in performance in MOS A but very little of the variation in performance in MOS B. (Perhaps some unmeasured skill accounts for most of the variance in MOS B.) Then it makes sense to assign individuals with higher values of the composite to MOS A and individuals with lower scores to MOS B. Although the skill measured by the composite is related to performance in both MOS, its relation is much stronger in MOS A. Thus, a single composite has differential validity for classification among MOS.

The present problem is somewhat different from that addressed by Horst. His objective was to measure the performance of an entire battery in predicting differences between criteria, while the objective of the present analyses is to compare the performance of different sets of composites based on the same (ASVAB) battery. As noted by Maier (1982), Horst's derivation is based on the assumption that the predictors are based on the full battery; i.e., that they are the multiple regression vectors for predicting the criteria from the ASVAB. Thus, while computation of H for the 98 separate MOS regression vectors provides the maximum achievable differential validity of the ASVAB, the computation of the differential validity of a particular set of composites involves more than merely applying Horst's formula to the covariance matrix of the composites.

Each MOS is associated with a single composite, so the comparison of expected performance between two MOS is associated with a pair of composites (although in many cases, they are the same composite). To assess the differential validity of alternative sets of composites, then, we applied the formula in equation (1), where the predictors for each pair were limited to the one or two composites associated with the pair. Specifically, we obtained the least squares predictor of the difference in criteria between each pair of MOS and then averaged the squares of these over all pairs of MOS.

The measure of composite differential validity we used was:

(3) $M^2 = \text{Average}(B_{(ij)}C_{ijk})^2/2,$

where C_{ijk} is the pair of composite values associated with MOS i and j for individual k,
$B_{(ij)}$ is the regression vector for predicting the difference $y_{ik}-y_{jk}$ based on the two composites, and
the average is over all i, j, and k, with i and j not equal.

As Horst had noted, one can estimate the required regression coefficients, even though no individual case has more than a single criterion score. Because the estimation for different pairs is based on different sets of composites, however, the elegant solution which Horst discovered is not available. Nevertheless, the computations were straightforward, though somewhat expensive in computer time.

The computations were carried out for two separate cases: (1) pairs of MOS associated with the same composite, and (2) pairs of MOS whose composites were not perfectly correlated. In each case, the critical assumption, which also underlies Horst's derivation, is that the regression of criterion on composite is the same in both the selected and unselected groups. To simplify the derivation, we assumed that all variables were standardized for the group for which they are available.

Case 1: both MOS i and MOS j use the same composite

The objective is to select b_{ij} to minimize

(4) $\text{Average}(y_{ik}-y_{jk} - b_{(ij)}c_k)^2,$

where the average is over all accessions, and
c is the common composite for both MOS_i and MOS_j.

The solution can be shown to be

(5) $b_{(ij)} = b_i - b_j.$

That is, the result is simply the difference between the regression coefficients for predicting the criteria in the two MOS separately.

Case 2: MOS i and MOS j use different composites

The objective is to minimize

(6) $$\text{Average}\left(y_{ik}-y_{jk} - (b_{(ij)i}c_{ik}+b_{(ij)j}c_{jk}) \right)^2,$$

where the average is over all accessions,
c_{ik} and c_{jk} are the two composite values associated with MOS i and j, for individual k, and
$b_{(ij)i}$ and $b_{(ij)j}$ are the associated regression coefficients for predicting the difference.

The joint solution for $\underline{B}_{(ij)}' = (b_{(ij)i}, b_{(ij)j})$ turns out to be

(7) $$\underline{B}_{(ij)} = \underline{B}_{(i)} - \underline{B}_{(j)},$$

where $\underline{B}_{(i)}$ and $\underline{B}_{(j)}$ are the regression vectors for predicting the available criteria in MOS i and j each using the pair of composites. Note that the values of $\underline{B}_{(i)}$ and $\underline{B}_{(j)}$ depend on the particular pairing of i and j.

Thus, in both Case 1 and Case 2 we obtain computable estimates of the regression coefficients; and from these it is straightforward to obtain the measure defined in equation (3). The maximum value for this statistic, for any set of linear composites based on the ASVAB, is the value of H^2 in equation (1).

The failure of a set of composites to possess differential validity, therefore, can be divided into two parts: (1) failure of the ASVAB as a battery to measure skill components that differ between MOS, and (2) failure of the particular set of composites to capture the potential differential validity of the ASVAB. We can assess the extent to which the composites capture the differential validity possessed by the ASVAB as the ratio of M to H.

Although all of the MOS differences are important for some decisions, it is plausible to assign greater weight to the valid estimation of differences that are involved in the most frequent decisions. Therefore, we computed H and M, weighting the entry for each pair by the product of the numbers of the accessions in the two MOS. The data base consisted of Army enlisted accessions for FY81/82 for whom SQT or training scores were available.

The results are contained in Table 1. Generally, the unit-weight composites yield differential validity estimates from 55% to 68% of the potential differential validity in the ASVAB. The solutions with fewer composites, as expected, yielded slightly lower estimates of differential validity, although there was virtually no difference between the 2, 3, and 4 composite alternatives. Use of a single composite resulted in noticeably lower differential validity.

The comparison between the operational composites and the alternative set of nine composites, in which the Clerical & Administrative (CL) and Surveillance

& Communications (SC) composites were replaced, yielded no noticeable difference. Thus, the significant increase in overall predictive validity achieved by introduction of these two changes is not at the cost of decrease in differential validity.

Table 1

Differential Validity Estimates for Alternative Sets of Composites (Weighted)

Composites	Average Squared Difference (H^2 or M^2)*	Root Mean Square Difference (H or M)	Relative Efficiency (H or M)/H
Full linear model (98 composites)	.046	.214	(100%)
Current 9 composites	.021	.146	68%
Revised 9 composites (CL and SC changed)	.020	.142	66%
Alternative 4 composites	.016	.125	59%
Alternative 3 composites	.014	.120	56%
Alternative 2 composites	.016	.125	58%
Alternative 1 composite	.011	.106	50%
GEM**	.014	.117	55%

* Note: the measure of differential validity is H for the full 98-composite alternative and M for the other alternatives.

** Three of the four "MAGE" composites. G is used for CL and OF clusters of MOS; E is used for EL, SC, and ST clusters of MOS; and M is used for CO, FA, GM, and MM clusters of MOS.

The results presented in this section must therefore be interpreted with caution. One set of composites might be measured as possessing greater differential validity than another, even though the other set would lead to a more valuable increase in overall performance of enlisted personnel. Four aspects of the practical application of composites for classification of Army recruits are particularly important to consider in interpreting the results.

(1) The constraints on numbers of recruits needed in each MOS severely restrict the assignment process, so that many recruits must be assigned to MOS for which they are not optimally matched.

(2) Recruits are free to make choices and cannot be summarily assigned to the MOS that the composites identify as optimal.

(3) The appropriate criteria are not expected performance differences but the relative utility of those differences; however, the utility scales are not yet available.

(4) Current practice mixes selection and classification, yet we are addressing the questions of validity for selection and classification separately.

Each of these factors would affect the measurement of overall performance of any set of composites, and to the extent that the effects are the same for all composites, the general results can be meaningfully interpreted. Factors that would affect one set of composites more than another, however, will require further investigation. For example, if the source of differential validity in one set of composites lies primarily in comparisons between "high payoff" MOS, the real value of that set of composites would be relatively higher than its measured validity.

Constrained Assignment, Simulation Solution

We turn now to a method that addresses the problem of selecting optimal composites in the Army's context of constrained assignment, where a certain number of recruits are needed for each MOS, and the MOS compete with each other for the more highly skilled applicants. Composites must be evaluated in terms of expected increase in performance, given assignments within constraints. The expected gains are a complex function of (a) the validities of predictors, (b) the correlations among predictors, (c) the distribution of requirements, and (d) the amount of information available about other applicants when each classification is made. The choice of composites will affect both validities and intercorrelations, but the distribution of requirements will determine the relative importance of the various validities and intercorrelations, and the value of the composites must be shown to be robust across levels of information prior to classification.

In the context of constrained assignment, the level of information available at the time each assignment is made becomes important, and it is not necessarily the case that the composites which perform best in the context of one assignment procedure will also work best in other assignment methods. While algorithms exist for the identification of optimal assignments when information is simultaneously available on the entire cohort of candidates, the employment of these algorithms is both exceedingly costly and, to the extent that they disregard the need to make selection and classification decisions sequentially, unrealistic. Efforts being undertaken in Project B aim to create an assignment system which makes use of as much simultaneous cohort information as possible while modeling the sequential nature of the actual

assignment process. In any case, the employment of simultaneous assignment algorithms do serve to show the upper limits that must be placed on expectations for improvement in aggregate performance based on improved assignment tools and procedures.

In order to assess the discriminant validity of alternative composite sets, we selected a criterion and implemented a representative assignment algorithm using alternative composite sets. The results were then evaluated to determine which sets of composites possessed the most effective discriminating power.

Discriminant Validity Criterion. In order to measure the discriminant validity of a set of composites, in the context of an assignment procedure, a sample of individuals already assigned to MOS were drawn and hypothetically reassigned, using the procedure and the composites. The expected performance of each individual was then computed for his or her revised assignment, and the results were aggregated across the entire sample. The results can be stated in standard deviation units of expected performance improvement in comparison with random assignment.

For the purposes of this assessment, the 64,907 individuals in the 98-MOS data base used for combined criterion validation were hypothetically reassigned, with the constraint that the total number assigned to each MOS be the same as the current assignment. For the purposes of sequential assignment, these individuals were sorted in order of date of entry into the service.

The standard deviation of performance was defined by the variation of the criterion measure in the validation analyses. As noted earlier, the criterion measures were scaled so that the standard deviation was equal to 20 for every cell in the analysis. As a result, the assignment algorithms operated as if the value of an increment in performance were the same in every MOS. While this is most surely not accurate, any improvement in this assumption requires data on the relative utility of performance increments in different MOS, data to be gathered later in Project A.

An assignment without constraints provided an extreme upper limit on expected gains from improved assignments, producing an expected gain of .54 standard deviations over random assignment, as shown in Table 2. The resulting assignment, of course, was highly unrealistic, shifting large numbers of personnel into MOS whose performance is closely associated with skills measured by the ASVAB. The greatest gain obtained for a constrained assignment, an approximation to a simultaneous assignment of all individuals, using the ridge regression vectors themselves as composites, was .21 standard deviations.

As a realistic baseline for comparison of sequential assignments using different sets of composites, the current assignments have an average expected performance gain, compared to random assignment, of .04 standard deviations (see Table 2). The cluster means range from .39 sd for Skilled Technical MOS to -.13 sd for Clerical/Administrative MOS. That is, some MOS are getting better recruits than a random draw, others worse, in terms

Table 2

Expected Gains in Performance, for Alternative Baseline Assignments, in Performance Standard Deviations

Random Assignment	Current Assignment	An Effective Simultaneous Constrained Assignment	Best Unconstrained Assignment
.00	.04	.21	.54

of skills measured by the ASVAB. This is to be expected, because for many MOS the critical skills are not those measured by the ASVAB, so it is not important that soldiers in these MOS have high ASVAB scores. It is important, however, that the overall mean be greater than zero.

Parallel assignments can be optimized through linear programming, but only at substantial computational costs. As an inexpensive substitute, we employed the procedure of sorting the soldiers so that those for which the expected performance depended most on the assignment to a particular MOS appeared early in a sequential assignment. In particular, we employed a variation of the procedure proposed by Ward (1958), in which we took as the index of dependence the difference between the highest composite and the second highest composite. Using this procedure, we could ensure that the large majority of soldiers for whom the expected difference was large would be assigned optimally.

Assignment Procedure. The sequential assignment procedure, shown in Figure 1, identified the best MOS in each cluster for each soldier and then selected among this small number of MOS. The composites were calculated for each soldier for each MOS. Then a "best" MOS in each cluster was selected for the individual. An increment was added to each composite in proportion to its lag* in being filled, relative to other clusters. The addition of the

*If the desired proportions in clusters 1,..., m were p_1,..., p_m, the increment was defined as

$$\text{delta} \times (n_r p_i - n_{ir}),$$

where n_r is the number already assigned, n_{ir} is the number already assigned to cluster i, and delta is selected to ensure appropriate sensitivity. In fact, results were generally independent of delta, over a range from .05 to 1.00, and the value .20 was used in most cases.

increment served to distribute the deviations from the maximum throughout the process. Without this increment, all soldiers in some clusters would be selected from the first small fraction of the file. The addition of the increment is in lieu of the procedure of modifying requirements and cutoffs for MOS from month to month.

The choice of the "best" MOS in a cluster was made in a manner similar to that proposed by Cronbach and Gleser (1965) for placement with fixed quotas. The MOS within a cluster were rank-ordered in terms of the estimated validity of the composite for the MOS. Then, the distribution of composite scores for the population (i.e., the entire sample on which the simulation was run) was partitioned so that the highest scores could be assigned to the MOS with the highest r-squared, and so forth. The assignment for an individual soldier would be determined by the place in the distribution associated with his or her composite value, as shown in Figure 2. When an MOS was filled, the assignment was to an adjacent MOS.

The choice among this tentative set of best MOS for each cluster was on the basis of an estimated performance score:

$$100 + r_i\,(c_{ij} - 100),$$

where the mean for composite i is 100, r_i is the validity for composite i in the selected MOS, and c_{ij} is the value of composite i for soldier j.

Results

The overall mean gains are shown in Table 3. The current composites clearly perform more poorly than the alternatives considered in this investigation. Indeed, the assignments actually made by counselors were significantly better than assignments based purely on the existing composites. This result was replicated for all four assignment procedures and is large in comparison with chance variations that might occur. Replacement of three of the current composites (CL, SC, and FA) with composites identified from the validation analyses significantly improved the expected performance gain (e.g., to .03 standard deviations for the sequential implementation of Procedure B), but the results were still inferior to the performance of the two-, three-, and four-composite solutions. Although the performance of these latter alternatives were nearly indistinguishable, the three-composite solution was slightly more powerful than the others.

Table 3

Gains in Aggregate Expected Performance,
For Alternative Composite Sets

	2	3	4	9(Rev.)	9
Expected Gain	.08	.08	.07	.03	.01

Using the two-, three-, or four-composite solution, the gain of .07 or .08 standard deviations in aggregate performance is substantial. Roughly, it corresponds to a shift in performance of one person in every eight from the mean level to a level better than 3 out of 4 soldiers. In comparison with the revised nine-composite solution, the gain from using the three-composite solution is about .05 standard deviations. It should be noted, however, that these gains are based on assignment to a particular MOS, essentially a "two-sided" cutoff. Gains can be expected to be only half as large if cutoffs are strictly one-sided, with MOS choices made randomly from among MOS for which one's scores are high enough.

Table 4

Gains in Aggregate Expected Performance, for Sequential Assignment and Different Composites, By MOS Group

	Number of Composites					
	2	3	4	9(Rev.)	9	Current
Total	.08	.08	.07	.03	.01	.04
CL	.29	.17	.16	.27	-.14	-.13
ST	.04	.19	.21	.19	.27	.39
SC	.21	-.01	-.11	-.07	-.09	.05
MM	.21	.23	.02	.05	.09	.15
OF	.02	.10	-.06	-.29	-.18	-.07
CO	-.18	-.15	-.11	.01	.02	.06
FA	.06	.08	.29	.09	.11	-.02
GM	.03	.06	.07	-.12	-.11	-.10
EL	.22	.22	.33	.07	.12	.02

The gains were not uniform across MOS, as shown in Table 4. Compared to current assignments, the simulated two-composite assignments tended especially to improve the expected performance of MOS in the CL cluster, while detracting from the average expected performance in the ST and CO clusters. As might have been expected, the simulation using the current composites matches most closely the actual current assignments.

The general trend is that fewer composites perform better. The clusters were constrained to keep intact the current composite clusters, which in many cases did not serve to optimize assignments; when there were only two

or three clusters, the barriers created by the current nine clusters were largely non-operational. In other words, the current nine clusters are not optimal for prediction of the criteria used in this investigation. Compared to the current clusters, validities of 46 of the 98 MOS could be increased by .02 or more by switching to a different one of the current composites.

One notable aspect of the data summarized in Table 4 is that the average gain for the CO job cluster is simulated to be lower with simulated assignments than it is in actual assignments. This is due at least in part to the greater utility assigned to performance in these jobs. To assess the importance of differential utility on the optimal assignment process, we replicated the three-composite assignments, assigning several increments in utility of expected performance to the CO jobs. The results, shown in Table 5, suggest that an increment of roughly 20% in the utility of performance would offset the lower predictability of the ASVAB in these jobs in

Table 5

Gains in Aggregate Expected Performance for Sequential Assignment and Different Utility Increments, By MOS Group

	Performance Utility Increment for CO jobs			
	0%	10%	37%	100%
Total	.08	.08	.05	.04
CL	.17	.16	.12	.03
ST	.19	.17	.11	.01
SC	-.01	-.05	-.14	-.20
MM	.23	.19	.01	-.04
OF	.10	.11	-.13	-.08
CO	-.15	-.09	.27	.43
FA	.08	.03	-.12	-.23
GM	.06	.01	-.14	-.14
EL	.22	.23	.08	-.12

the performance of the assignment algorithm. While these values cularly dependent on the specific nature of this assignment al apparent that any successful employment of automated classific dures must take into account differential utility of performan

It was also apparent from these analyses that the setting of c critical importance in maximizing expected performance. Sever

simulations in which choices of individual MOS were made randomly, within the cluster identified with the maximum composite value, resulted in performance no better than random assignment. Furthermore, one-sided cutoffs generally resulted in gains half the size of gains from assignments based on two-sided cutoffs. While it is perhaps unreasonable to screen individuals out of MOS because they are "over-qualified," counseling which guides recruits toward MOS where they are not over-qualified can have significant effects on aggregate average performance expectations.

References

Brogden, E. H. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. Educational and Psychological Measurement, 19, 181-190.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions (2nd ed.). Urbana: University of Illinois Press.

Horst, P. (1954). A technique for the development of a differential prediction battery. Psychological Monographs: General and Applied, 68, No. 9.

Maier, M. H. (1982, December). Issues for Defining ASVAB 11/12/13/14 Aptitude Composites (A Briefing Presented to the ASVAB Working Group) (Center for Naval Analyses No. 82-3199). Alexandria, VA: Marine Corps Operations Analysis Group.

Ward, J. H., Jr. (1958). The counseling assignment problem. Psychometrika, 23, 55-56.

Complex Cross-Validation of the Validity of a Predictor Battery

David Brandt
Don McLaughlin
Laurie Wise

American Institutes for Research

Paul Rossmeissl

Army Research Institute

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

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.

Presented at the Annual Convention of the American Psychological Association, Toronto, Ontario, Canada.

Although the research reported at this symposium was conducted on the largest database yet assembled for ASVAB validation, in many cases sample sizes for individual military occupational specialties (MOS) were barely adequate for estimation of parameters. Because much of the crucial work depended on these parameter estimates, it was necessary to employ some method of assessing the stability of these statistics. This paper discusses two methods that we employed, conventional cross-validation and the bootstrap, to assess the stability of sample statistics.

Several methods exist for determining the sampling error of a statistic. First, the standard errors of many commonly used statistics can be computed using formulas that make some distributional assumptions. Almost without exception, formula standard errors are based on standard normal theory. Its chief weakness,of course, is that it is limited to statistics for which standard normal theory is available. Also, the data at hand must conform to these assumptions reasonably well.

Regrettably, these conditions are not met in much of our work. In ASVAB validation research, an important criterion variable has been training success. Because Army training schools have, for the large part, adopted criterion-referenced tests, the usual assumption of a normal distribution of errors is sometimes not plausible. Pronounced ceiling effects were found in many MOS. The distributions of our other major criterion, Skill Qualification Tests (SQT), while better behaved, were also negatively skewed. Furthermore, some of the key statistics in validation research do not have formula standard errors. One such statistic is the validity coefficent after correction for restriction of range.

The most commonly used alternative to formula standard errors is some form of cross-validation. In its simplest form, cross validation consists of dividing the sample into halves and performing the same analysis in both halves. The researcher then evaluates the similarity of the findings to determine whether the result has "cross-validated." We used this method to evaluate the stability of our emprical clustering of MOS. In the second part of this paper, we will describe our use of this procedure.

The cross-validation methodology does not dictate that only two subsamples should be obtained. One can define many divisions of the sample. It is often advantageous to define many subsamples and repeat the parameter estimation procedure in each one. This procedure produces a sampling distribution of the statistic of interest. From this distribution, an empirical estimate of the sampling distribution of the statistic in the population can be obtained. Such estimates are valid regardless of the shape of the population distribution.

One particularly elegant method of generating multiple subsamples has been developed by McCarthy (1976). He uses the method of Balanced Half-Sample Replications (BHS) to divide up the full sample. This method is based on a set of special design matrices first published by Plackett and Burman (1946). These matrices have the property of orthogonal balance. Plackett-Burman design matrices up to order 92 are available.

The Plackett-Burman design matrices define a series of orthogonal partitionings of the sample. Because of the structured way of defining the partitionings, the BSR method has been shown to be optimal in simple cases. In more complex cases, they are believed to be highly efficient. Perhaps because of this efficiency, BHS has become the prefered repeated replication method of estimating standard errors from clustered and/or stratified sample surveys. OSIRIS IV (1981) contains a procedure for performing these calculations (&REPERR), and Wise (1983) has implemented a similar procedure, PROC BRRVAR, into SAS. Both programs use the Plackett-Burman design matrices to define partitionings of the sample, repeat a specified calculation using eachsubsample, and compute standard errors from the resulting distributions.

A less elegant method of defining multiple subsamples has been labeled the "jackknife" by John Tukey (Mosteller & Tukey, 1977, p. 133ff). The jackknife defines replicates by removing one observation at a time from the original data and calculating the statistic of interest from each of the resulting datasets. The variability of the statistic across the "jackknifed" datasets can then be described.

Although it is less elegant than the BHS method, the jackknife is more general and can be more readily applied. Tukey (Mosteller & Tukey, 1977) stated that:

> The name "jackknife" is intended to suggest the broad usefulness of a technique as a substitute for specialized tools that may not be available, just as the Boy Scout's trusty tool serves sovariedly. The jackknife offers ways to set sensible confidence limits in complex situations. The basic idea is to assess the effect of each of the groups into which the data have been divided, not by the result for that group alone, but rather through the effect upon the body of data that results from omitting that group. (p. 133)

In the ASVAB validation, we chose to use a more primitive but more general method of obtaining subsamples. This method, known as the "bootstrap," was invented by Bradley Efron (Efron, 1979). The name was chosen to suggest the idea of "picking yourself up by your bootstraps." A "bootstrap" replication is simply a subsample of size N drawn with replacement from the sample of size N. By comparison, a "jackknife" replication is a sample of size N minus 1 drawn without replacement. From a technical point of view, Efron (1979) showed that the jackknife can be thought of as a linear expansion method for approximating the bootstrap.

Unlike the jackknife and BHS methods, the bootstrap method does not determine the number of replications to be generated. When the sample size is small, many more replications than the number of data points are typically defined. For example, in their Scientific American article on the bootstrap, Diaconis and Efron (1983) illustrated the method by estimating the standard error of a correlation of average LSAT score and freshman GPA from a sample of fifteen law schools. To obtain their estimate of the standard error, they generated 1000 bootstrap replications.

The repeated replication methods of obtaining variance estimates can be thought of as ways of substituting "brute force" computing power for statistical theory. Among the repeated replication methods currently available, the bootstrap makes the heaviest demands on computing power. To a large extent, the work in the 1960's and early 1970's on BHS techniques was motivated by a desire to use computer-intensive methods at a time when computing power was considerably more expensive. Today, the cost of performing the calculations reported by Diaconis and Efron is trivial. In fact, those calculations can easily be done on microcomputers that are readily available for less than $200. In the ASVAB validation, we used the bootstrap for a much more massive task. We estimated the standard errors of each element of two 98 x 22 matrices of validity coefficients, and each bootstrap replicate consisted of nearly 65,000 observations. A major point of this paper is to illustrate how computer-intensive methods can be used to provide reasonable answers to questions that could not be addressed a few years ago. As computing power continues to become cheaper, these methods can be used more and more widely.

We first describe our use of the bootstrap to estimate the standard errors of corrected validity coefficients and then present work on the (conventional) cross-validation of our measure of similarity.

Case 1: The Bootstrap

Method

We used the bootstrap method to estimate the standard errors of validity coefficients that have been adjusted for restriction of range. Before discussing the bootstrap method, we will first present some background information on our use of the correction for restriction of range.

In the Army ASVAB validation research, we analyzed the relations between ASVAB and the training or SQT criterion in each MOS cell for which we had at least one hundred observations. An MOS cell can be thought of as the largest unit of analysis for which we had a unique outcome measure. For the training criterion, we used the unique combination of MOS, training school, and course. For SQT scores, we used combinations of MOS, year (FY 81 or FY 82), and "track," where a track denotes a specialized job within an MOS. In order to obtain a broader coverage of all MOS in the Army, we relied primarily on a "combined" criterion score. This was defined as either the standardized training or SQT score (if only one was available for a given soldier) or the higher of the two if both were available. For each analysis cell, as defined above, we estimated sample and corrected validity coefficients for the nine ASVAB composites currently used by the Army, the four alternative composites identified by this research, the four MAGE composites used by the Air Force, and the five High School composites.

The adjusted validities were computed by a program written in SAS PROC MATRIX that implemented the classical multivariate correction due to Lawley (1943; See Lord & Novick, 1968). This program takes as input the population

covariance matrix among the ASVAB subtests, and the sample covariance matrices for each of the 98 anlysis cells (i.e., MOS) in our sample of Army MOS. The major output of the program is the matrix of corrected validity coefficients. The rows of the matrix correspond to analysis cells and the columns correspond to ASVAB composites. The program also produces the matrix of unadjusted validities, i. e., sample correlations. Sample and adjusted validities were obtained for the nine composites currently in use by the Army, the four "optimal" composites identified by this research, the four MAGE composites, and the five High School composites. There were 64,907 observations in the 98 analysis cells.

The bootstrap estimates of the standard errors of these validities were obtained in the following way.

First, a bootstrap sample was drawn for each of the 98 analysis cells. From this sample, the covariance matrix among the ASVAB subtests and the combined criterion was computed. This matrix, together with the population ASVAB covariance matrix, was input to the PROC MATRIX restriction of range program. The sample and corrected validities were saved for each bootstrap replicate. This entire process was repeated one hundred times. In effect, one hundred replicates of N=64,907 were drawn, and sample and adjusted validities were obtained. After the hundred matrices of validity coefficients were obtained, the standard deviation across the hundred replicates was computed for each element in the 98 by 22 array. This process was repeated for the uncorrected validity coefficients. The cost of carrying out all of these calculations on the IBM 3081 at the National Institutes of Health was approximately $1,500.

Results

For the sake of clarity, we only report the findings for the composites currently in use by the Army. Tables 1 and 2 present the sample and corrected validities of these composites, together with their standard errors. In each table, the validity estimates for the nine composites are paired with the bootstrap estimate of the standard error. The MOS are sorted within MOS cluster and the clusters are presented in alphabetical order. The last column on the right is the approximate large sample standard error of a correlation, 1 divided by the square root of N. It is expected that the bootstrap estimates of the standard errors of uncorrected validities will be of the same magnitude as these numbers. Standard errors of adjusted validities are expected to be larger.

The sample validities and standard errors in Table 1 provide a baseline for evaluating the bootstrap estimates of standard errors of the corrected validity coefficients. First, if the bootstrap technique is behaving properly, we would expect that these standard errors are of the same magnitude as the classical estimate of the standard error of a correlation. Table 1 indicates that this is, indeed, the case. Standard errors of individual composites are generally within plus or minus .02 of the classical estimates, and frequently the agreement is even better. As would be expected, agreement is best for very large MOS (e.g., 63B, 13F, 71L), and poorest for

Table 1

SAMPLE VALIDITIES FOR CURRENT COMPOSITES
COMBINED CRITERION

MOB	AA	N	CL	SE	CO	SE	EL	SE	FA	SE	GM	SE	MM	SE	OF	SE	SC	SE	ST	SE	NORM SE
71D	CL	114	0.21	0.09	0.33	0.06	0.36	0.09	0.37	0.07	0.33	0.08	0.27	0.07	0.29	0.07	0.27	0.06	0.37	0.08	0.09
71L	CL	2782	0.33	0.02	0.38	0.02	0.43	0.02	0.43	0.02	0.38	0.02	0.31	0.02	0.37	0.02	0.36	0.02	0.42	0.02	0.02
71M	CL	182	0.32	0.07	0.33	0.06	0.34	0.06	0.34	0.06	0.33	0.06	0.32	0.06	0.37	0.05	0.35	0.06	0.38	0.06	0.07
71N	CL	173	0.36	0.06	0.36	0.06	0.38	0.08	0.42	0.06	0.33	0.07	0.30	0.07	0.32	0.07	0.37	0.06	0.35	0.07	0.08
71C	CL	478	0.46	0.03	0.53	0.04	0.53	0.03	0.55	0.03	0.49	0.03	0.47	0.04	0.52	0.03	0.52	0.03	0.52	0.03	0.05
75B	CL	920	0.29	0.03	0.38	0.03	0.42	0.02	0.41	0.03	0.37	0.03	0.32	0.02	0.35	0.03	0.34	0.03	0.39	0.02	0.03
75C	CL	317	0.27	0.05	0.45	0.05	0.51	0.04	0.48	0.04	0.48	0.05	0.40	0.05	0.45	0.05	0.37	0.06	0.50	0.04	0.06
75D	CL	801	0.19	0.03	0.34	0.03	0.41	0.03	0.38	0.03	0.37	0.03	0.30	0.03	0.32	0.03	0.27	0.03	0.39	0.03	0.04
75E	CL	417	0.31	0.05	0.41	0.04	0.45	0.04	0.43	0.04	0.42	0.04	0.38	0.05	0.41	0.04	0.39	0.05	0.45	0.04	0.05
75F	CL	137	0.37	0.08	0.51	0.07	0.55	0.08	0.57	0.07	0.49	0.08	0.45	0.08	0.47	0.07	0.43	0.07	0.53	0.08	0.09
76C	CL	1296	0.18	0.03	0.26	0.02	0.32	0.03	0.28	0.03	0.29	0.02	0.26	0.02	0.28	0.02	0.24	0.03	0.29	0.02	0.03
76P	CL	559	0.26	0.04	0.34	0.03	0.40	0.04	0.44	0.03	0.31	0.04	0.23	0.04	0.27	0.04	0.26	0.04	0.38	0.04	0.04
76V	CL	214	0.20	0.06	0.29	0.06	0.34	0.06	0.28	0.06	0.31	0.06	0.22	0.06	0.26	0.06	0.25	0.06	0.35	0.05	0.07
76W	CL	684	0.21	0.04	0.35	0.04	0.34	0.03	0.32	0.04	0.34	0.03	0.30	0.04	0.31	0.04	0.30	0.04	0.31	0.04	0.04
76X	CL	158	0.07	0.06	0.10	0.06	0.11	0.08	0.14	0.06	0.07	0.08	-0.03	0.06	-0.00	0.07	0.06	0.06	0.07	0.08	0.08
76Y	CL	1134	0.19	0.03	0.25	0.03	0.29	0.03	0.29	0.03	0.25	0.03	0.19	0.03	0.22	0.03	0.23	0.03	0.27	0.03	0.03
11B	CO	5761	0.22	0.01	0.31	0.01	0.31	0.01	0.31	0.01	0.31	0.01	0.29	0.01	0.31	0.01	0.27	0.01	0.33	0.01	0.01
11C	CO	1482	0.25	0.03	0.33	0.02	0.33	0.02	0.33	0.02	0.33	0.02	0.32	0.02	0.33	0.02	0.30	0.03	0.34	0.02	0.03
11H	CO	948	0.22	0.03	0.26	0.03	0.27	0.03	0.27	0.03	0.26	0.03	0.25	0.03	0.26	0.03	0.24	0.03	0.27	0.03	0.03
12B	CO	2411	0.19	0.02	0.33	0.02	0.32	0.02	0.31	0.02	0.32	0.02	0.32	0.02	0.32	0.02	0.25	0.02	0.32	0.02	0.02
12F	CO	224	0.08	0.05	0.21	0.05	0.16	0.06	0.15	0.06	0.18	0.07	0.21	0.06	0.20	0.06	0.15	0.05	0.18	0.07	0.07
19D	CO	1035	0.36	0.03	0.35	0.03	0.36	0.03	0.34	0.03	0.35	0.03	0.33	0.03	0.35	0.03	0.31	0.03	0.36	0.03	0.03
19E	CO	2322	0.30	0.02	0.42	0.02	0.42	0.02	0.38	0.02	0.43	0.02	0.42	0.02	0.44	0.02	0.38	0.02	0.44	0.02	0.02
19F	CO	83	0.24	0.09	0.40	0.09	0.37	0.09	0.32	0.09	0.39	0.09	0.39	0.09	0.40	0.09	0.33	0.09	0.37	0.08	0.11
17K	EL	179	0.21	0.06	0.34	0.07	0.26	0.07	0.33	0.06	0.26	0.07	0.27	0.07	0.32	0.06	0.28	0.06	0.32	0.07	0.07
26Q	EL	442	0.24	0.08	0.26	0.07	0.38	0.06	0.31	0.07	0.36	0.06	0.24	0.08	0.25	0.06	0.27	0.07	0.35	0.06	0.08
27E	EL	305	0.33	0.06	0.32	0.05	0.32	0.05	0.34	0.05	0.30	0.05	0.28	0.05	0.34	0.05	0.36	0.05	0.32	0.05	0.06
31J	EL	130	0.17	0.07	0.29	0.07	0.26	0.08	0.29	0.07	0.21	0.08	0.17	0.09	0.24	0.08	0.21	0.07	0.33	0.07	0.09
31M	EL	1858	0.17	0.02	0.35	0.02	0.36	0.02	0.31	0.02	0.37	0.02	0.34	0.02	0.35	0.02	0.27	0.02	0.36	0.02	0.02
31N	EL	193	0.18	0.08	0.09	0.06	0.05	0.06	0.04	0.07	0.03	0.06	0.07	0.06	0.11	0.07	0.17	0.07	0.05	0.07	0.07
31V	EL	650	0.22	0.04	0.31	0.04	0.33	0.03	0.32	0.03	0.31	0.04	0.30	0.04	0.29	0.04	0.26	0.04	0.29	0.03	0.04
35K	EL	121	0.34	0.08	0.46	0.06	0.41	0.06	0.42	0.06	0.42	0.06	0.44	0.06	0.46	0.06	0.44	0.07	0.39	0.06	0.09
36C	EL	374	0.12	0.05	0.13	0.05	0.10	0.04	0.14	0.05	0.07	0.05	0.10	0.05	0.10	0.05	0.14	0.05	0.03	0.04	0.05
36K	EL	1581	0.16	0.03	0.30	0.02	0.25	0.03	0.26	0.02	0.27	0.02	0.28	0.02	0.29	0.02	0.23	0.02	0.27	0.02	0.03
11B	FA	4778	0.25	0.01	0.38	0.01	0.37	0.01	0.35	0.01	0.38	0.01	0.38	0.01	0.38	0.01	0.33	0.01	0.37	0.01	0.01
11F	FA	824	0.28	0.03	0.42	0.02	0.36	0.03	0.39	0.03	0.36	0.03	0.37	0.03	0.39	0.03	0.35	0.03	0.37	0.03	0.03
41C	GM	103	0.25	0.11	0.26	0.10	0.25	0.07	0.30	0.08	0.21	0.07	0.26	0.08	0.23	0.09	0.27	0.10	0.15	0.08	0.10
43E	GM	99	0.22	0.10	0.18	0.11	0.25	0.08	0.20	0.09	0.25	0.08	0.21	0.09	0.23	0.09	0.24	0.11	0.22	0.07	0.10
44B	GM	137	0.19	0.08	0.30	0.09	0.26	0.08	0.26	0.07	0.25	0.09	0.30	0.09	0.29	0.09	0.24	0.08	0.23	0.09	0.09
45K	GM	228	0.26	0.06	0.35	0.06	0.24	0.07	0.33	0.06	0.25	0.06	0.34	0.05	0.34	0.06	0.30	0.07	0.28	0.07	0.07
51B	GM	195	0.21	0.07	0.27	0.06	0.25	0.06	0.28	0.07	0.27	0.05	0.26	0.05	0.30	0.05	0.26	0.06	0.27	0.06	0.07
51K	GM	167	0.17	0.08	0.22	0.07	0.25	0.07	0.23	0.07	0.26	0.06	0.24	0.06	0.22	0.07	0.22	0.07	0.17	0.09	0.08
52D	GM	176	0.24	0.07	0.33	0.07	0.27	0.06	0.32	0.06	0.27	0.06	0.34	0.06	0.37	0.06	0.28	0.06	0.33	0.06	0.08
55B	GM	366	0.20	0.06	0.24	0.05	0.28	0.05	0.27	0.05	0.26	0.05	0.21	0.05	0.23	0.05	0.21	0.06	0.25	0.05	0.05
57E	GM	126	0.15	0.09	0.29	0.08	0.01	0.09	0.31	0.07	0.02	0.08	0.08	0.09	0.10	0.08	0.16	0.08	0.06	0.11	0.09
57H	GM	224	0.15	0.07	0.14	0.06	0.11	0.05	0.14	0.06	0.14	0.06	0.08	0.06	0.17	0.07	0.16	0.08	0.24	0.05	0.07
62E	GM	230	0.30	0.06	0.43	0.05	0.40	0.05	0.39	0.05	0.42	0.05	0.45	0.05	0.44	0.05	0.38	0.06	0.40	0.05	0.07
62F	GM	200	0.26	0.07	0.32	0.06	0.29	0.06	0.27	0.07	0.33	0.06	0.39	0.06	0.37	0.06	0.31	0.07	0.31	0.07	0.07
68J	GM	188	0.15	0.07	0.34	0.06	0.26	0.07	0.26	0.06	0.33	0.06	0.34	0.07	0.35	0.07	0.25	0.07	0.30	0.07	0.07
68M	GM	132	0.31	0.08	0.47	0.05	0.30	0.06	0.43	0.05	0.33	0.06	0.42	0.05	0.40	0.06	0.40	0.07	0.28	0.05	0.09

(Cont'd)

SAMPLE VALIDITIES FOR CURRENT COMPOSITES (Cont'd)

COMBINED CRITERION

MOS	AA	N	CL	SE	CO	SE	EL	SE	FA	SE	GM	SE	MM	SE	OF	SE	SC	SE	ST	SE	NORM SE
12C	MM	355	0.19	0.06	0.31	0.05	0.27	0.05	0.30	0.05	0.26	0.05	0.28	0.05	0.28	0.05	0.24	0.06	0.28	0.05	0.05
61B	MM	183	0.46	0.05	0.62	0.04	0.59	0.03	0.60	0.04	0.57	0.04	0.55	0.04	0.66	0.04	0.58	0.04	0.65	0.04	0.07
61C	MM	136	0.25	0.07	0.47	0.06	0.44	0.05	0.50	0.06	0.36	0.06	0.35	0.07	0.35	0.06	0.30	0.07	0.38	0.06	0.09
62B	MM	355	0.25	0.04	0.44	0.04	0.39	0.04	0.40	0.04	0.41	0.04	0.40	0.04	0.41	0.04	0.36	0.04	0.40	0.05	0.05
63B	MM	1818	0.14	0.02	0.31	0.02	0.30	0.02	0.27	0.02	0.32	0.02	0.33	0.02	0.31	0.02	0.24	0.02	0.28	0.02	0.02
63D	MM	342	0.10	0.07	0.13	0.06	0.06	0.06	0.07	0.06	0.11	0.06	0.10	0.05	0.13	0.05	0.16	0.07	0.08	0.06	0.05
63G	MM	161	0.01	0.08	0.21	0.06	0.09	0.07	0.10	0.07	0.11	0.08	0.18	0.07	0.16	0.08	0.10	0.09	0.07	0.07	0.08
63H	MM	781	0.14	0.04	0.32	0.03	0.33	0.03	0.30	0.03	0.34	0.03	0.33	0.03	0.31	0.03	0.23	0.03	0.33	0.03	0.04
63N	MM	509	0.06	0.04	0.19	0.04	0.12	0.04	0.15	0.04	0.15	0.04	0.21	0.03	0.17	0.03	0.12	0.04	0.10	0.04	0.04
63W	MM	527	0.10	0.04	0.18	0.04	0.19	0.04	0.16	0.04	0.22	0.04	0.17	0.04	0.19	0.05	0.15	0.04	0.22	0.05	0.04
63Y	MM	238	0.09	0.06	0.18	0.07	0.14	0.07	0.17	0.07	0.17	0.06	0.20	0.07	0.21	0.06	0.14	0.07	0.17	0.06	0.06
67N	MM	471	0.22	0.05	0.34	0.04	0.36	0.04	0.35	0.04	0.34	0.04	0.31	0.04	0.33	0.04	0.26	0.05	0.36	0.04	0.05
67T	MM	124	0.44	0.06	0.59	0.07	0.65	0.05	0.59	0.06	0.62	0.05	0.47	0.06	0.53	0.06	0.49	0.05	0.65	0.05	0.09
67U	MM	278	0.20	0.05	0.41	0.04	0.35	0.05	0.39	0.04	0.31	0.05	0.31	0.06	0.31	0.05	0.25	0.05	0.34	0.05	0.06
67V	MM	310	0.30	0.06	0.38	0.06	0.37	0.06	0.37	0.06	0.34	0.06	0.34	0.06	0.36	0.06	0.33	0.06	0.36	0.06	0.06
67Y	MM	241	0.17	0.06	0.25	0.07	0.27	0.06	0.18	0.07	0.32	0.06	0.33	0.07	0.32	0.07	0.25	0.07	0.26	0.06	0.06
68D	MM	121	0.29	0.08	0.55	0.05	0.45	0.06	0.53	0.05	0.44	0.07	0.47	0.06	0.46	0.06	0.38	0.07	0.43	0.06	0.09
68G	MM	121	0.08	0.12	0.24	0.11	0.37	0.07	0.27	0.10	0.34	0.07	0.22	0.10	0.27	0.12	0.14	0.12	0.36	0.08	0.09
15D	OF	406	0.12	0.05	0.26	0.05	0.20	0.05	0.24	0.05	0.20	0.05	0.22	0.04	0.25	0.04	0.18	0.05	0.21	0.04	0.05
15E	OF	280	0.09	0.05	0.19	0.04	0.16	0.05	0.17	0.05	0.18	0.05	0.19	0.05	0.22	0.05	0.15	0.05	0.19	0.05	0.06
16B	OF	288	0.16	0.06	0.30	0.05	0.28	0.05	0.29	0.05	0.27	0.05	0.25	0.05	0.26	0.05	0.23	0.05	0.28	0.05	0.06
16C	OF	118	0.17	0.07	0.29	0.07	0.25	0.09	0.30	0.07	0.21	0.09	0.17	0.09	0.22	0.07	0.23	0.07	0.21	0.09	0.09
16D	OF	112	0.32	0.09	0.48	0.09	0.38	0.07	0.40	0.08	0.46	0.07	0.47	0.09	0.46	0.09	0.45	0.09	0.37	0.07	0.09
16H	OF	104	0.23	0.09	0.44	0.08	0.44	0.08	0.42	0.09	0.43	0.08	0.36	0.09	0.41	0.09	0.31	0.09	0.46	0.08	0.10
16J	OF	119	0.34	0.08	0.34	0.09	0.43	0.06	0.40	0.09	0.37	0.07	0.30	0.07	0.36	0.07	0.36	0.08	0.42	0.07	0.09
16R	OF	404	0.12	0.04	0.20	0.04	0.17	0.04	0.18	0.04	0.19	0.04	0.17	0.04	0.20	0.03	0.18	0.04	0.18	0.04	0.05
16S	OF	592	0.11	0.04	0.30	0.04	0.30	0.03	0.27	0.04	0.32	0.03	0.31	0.03	0.30	0.03	0.21	0.04	0.29	0.03	0.04
44C	OF	2959	0.13	0.02	0.31	0.02	0.30	0.02	0.27	0.02	0.31	0.02	0.29	0.02	0.31	0.02	0.23	0.02	0.30	0.01	0.02
94B	OF	3322	0.15	0.02	0.33	0.02	0.35	0.02	0.30	0.02	0.34	0.02	0.29	0.02	0.32	0.02	0.25	0.02	0.34	0.02	0.02
05B	SC	890	0.09	0.04	0.25	0.03	0.26	0.03	0.24	0.04	0.26	0.03	0.24	0.03	0.27	0.03	0.21	0.03	0.28	0.03	0.03
05C	SC	1971	0.10	0.02	0.37	0.02	0.38	0.02	0.35	0.02	0.37	0.02	0.35	0.02	0.36	0.02	0.26	0.02	0.37	0.02	0.02
05G	SC	119	0.26	0.10	0.43	0.07	0.27	0.09	0.34	0.09	0.30	0.08	0.36	0.07	0.41	0.08	0.39	0.08	0.33	0.09	0.09
17C	SC	187	0.06	0.07	0.22	0.06	0.22	0.07	0.19	0.06	0.21	0.07	0.18	0.07	0.22	0.07	0.20	0.05	0.23	0.07	0.07
72E	SC	562	0.12	0.04	0.46	0.03	0.49	0.03	0.42	0.03	0.51	0.03	0.51	0.03	0.49	0.03	0.34	0.04	0.47	0.03	0.04
05H	ST	110	0.49	0.07	0.43	0.07	0.43	0.06	0.51	0.05	0.37	0.08	0.33	0.07	0.42	0.07	0.51	0.07	0.45	0.06	0.10
13E	ST	678	0.19	0.03	0.29	0.03	0.31	0.03	0.33	0.03	0.28	0.03	0.22	0.03	0.24	0.03	0.23	0.03	0.30	0.03	0.04
54E	ST	270	0.32	0.05	0.36	0.05	0.36	0.05	0.39	0.05	0.33	0.05	0.33	0.05	0.34	0.05	0.34	0.05	0.36	0.05	0.06
74D	ST	52	0.32	0.10	0.34	0.09	0.37	0.09	0.40	0.09	0.30	0.08	0.22	0.08	0.29	0.09	0.34	0.09	0.37	0.10	0.10
82C	ST	536	0.32	0.04	0.43	0.03	0.43	0.03	0.46	0.03	0.38	0.03	0.38	0.03	0.37	0.04	0.35	0.04	0.36	0.04	0.04
91C	ST	233	0.37	0.05	0.38	0.05	0.39	0.05	0.38	0.05	0.35	0.05	0.34	0.05	0.36	0.05	0.41	0.05	0.35	0.04	0.07
91E	ST	301	0.20	0.05	0.25	0.05	0.32	0.05	0.31	0.05	0.27	0.05	0.21	0.05	0.22	0.05	0.21	0.05	0.32	0.04	0.06
91P	ST	159	0.25	0.08	0.24	0.08	0.24	0.07	0.27	0.08	0.21	0.08	0.25	0.08	0.24	0.08	0.25	0.08	0.22	0.08	0.08
91R	ST	145	0.38	0.07	0.31	0.06	0.26	0.07	0.31	0.07	0.24	0.07	0.23	0.07	0.27	0.07	0.40	0.06	0.24	0.07	0.08
92B	ST	364	0.15	0.04	0.29	0.05	0.34	0.05	0.32	0.05	0.30	0.05	0.24	0.05	0.25	0.05	0.20	0.05	0.32	0.05	0.05
93H	ST	114	0.09	0.09	0.27	0.08	0.30	0.09	0.30	0.09	0.24	0.09	0.24	0.08	0.24	0.08	0.14	0.09	0.31	0.09	0.09
95B	ST	3695	0.21	0.02	0.31	0.01	0.33	0.01	0.31	0.01	0.31	0.01	0.28	0.02	0.30	0.02	0.27	0.02	0.32	0.01	0.02
96B	ST	172	0.32	0.06	0.35	0.07	0.48	0.06	0.41	0.07	0.39	0.07	0.28	0.07	0.32	0.07	0.32	0.07	0.44	0.06	0.08
98C	ST	186	0.30	0.06	0.39	0.06	0.36	0.06	0.52	0.05	0.25	0.07	0.16	0.06	0.25	0.06	0.32	0.07	0.39	0.06	0.07

Table 2

ADJUSTED VALIDITIES FOR CURRENT COMPOSITES
COMBINED CRITERION

MOS	AA	N	CL	SE	CO	SE	EL	SE	FA	SE	GM	SE	MM	SE	OF	SE	SC	SE	ST	SE	NORM SE
71D	CL	114	0.26	0.25	0.37	0.18	0.39	0.16	0.38	0.19	0.34	0.15	0.29	0.18	0.32	0.21	0.29	0.24	0.40	0.18	0.09
71L	CL	2782	0.56	0.02	0.55	0.02	0.58	0.02	0.59	0.02	0.52	0.02	0.49	0.02	0.54	0.02	0.56	0.02	0.58	0.02	0.02
71M	CL	182	0.57	0.10	0.54	0 08	0.51	0.07	0.55	0.08	0.49	0.06	0.52	0.08	0.57	0.08	0.58	0.10	0.55	0.07	0.07
71N	CL	173	0.70	0.06	0.44	0.05	0.67	0.05	0.71	0.05	0.60	0.05	0.41	0.05	0.45	0.05	0.70	0.06	0.46	0.05	0.06
71C	CL	478	0.64	0.04	0.45	0.03	0.45	0.02	0.48	0.03	0.40	0.03	0.41	0.03	0.66	0.03	0.48	0.04	0.65	0.03	0.05
75B	CL	920	0.49	0.05	0.52	0.03	0.55	0.03	0.56	0.03	0.51	0.03	0.49	0.03	0.52	0.04	0.52	0.04	0.54	0.03	0.03
75C	CL	317	0.50	0.09	0.59	0.07	0.63	0.04	0.62	0.07	0.40	0.04	0.56	0.07	0.59	0.08	0.55	0.09	0.63	0.06	0.06
75D	CL	801	0.40	0.03	0.47	0.05	0.53	0.05	0.51	0.06	0.48	0.05	0.43	0.06	0.45	0.06	0.43	0.07	0.51	0.05	0.04
75E	CL	417	0.53	0.09	0.55	0.07	0.58	0.06	0.58	0.07	0.54	0.06	0.52	0.07	0.54	0.07	0.55	0.08	0.58	0.06	0.05
75F	CL	137	0.57	0.16	0.61	0.11	0.64	0.10	0.44	0.12	0.57	0.10	0.55	0.11	0.59	0.12	0.58	0.15	0.63	0.11	0.09
76C	CL	1294	0.44	0.05	0.49	0.03	0.52	0.03	0.50	0.04	0.50	0.03	0.48	0.03	0.50	0.04	0.48	0.04	0.51	0.04	0.03
76P	CL	559	0.55	0.05	0.57	0.04	0.42	0.04	0.45	0.04	0.55	0.04	0.50	0.05	0.55	0.05	0.55	0.05	0.42	0.04	0.04
76V	CL	214	0.47	0.10	0.49	0.08	0.52	0.08	0.50	0.08	0.49	0.07	0.44	0.08	0.48	0.09	0.49	0.10	0.53	0.08	0.07
76W	CL	484	0.45	0.09	0.58	0.06	0.58	0.06	0.56	0.07	0.58	0.05	0.55	0.06	0.56	0.07	0.52	0.08	0.56	0.06	0.04
76X	CL	158	0.31	0.28	0.30	0.23	0.31	0.22	0.33	0.23	0.28	0.21	0.26	0.23	0.28	0.25	0.31	0.27	0.30	0.23	0.08
76Y	CL	1134	0.39	0.04	0.41	0.05	0.43	0.04	0.45	0.05	0.40	0.04	0.36	0.05	0.39	0.05	0.40	0.06	0.43	0.05	0.03
11B	CO	5741	0.34	0.02	0.41	0.02	0.41	0.01	0.41	0.02	0.41	0.02	0.40	0.02	0.41	0.02	0.38	0.02	0.42	0.01	0.01
11C	CO	1482	0.34	0.04	0.44	0.03	0.43	0.03	0.43	0.03	0.43	0.03	0.43	0.03	0.44	0.03	0.41	0.04	0.44	0.03	0.03
11H	CO	948	0.34	0.04	0.38	0.04	0.38	0.04	0.38	0.03	0.37	0.04	0.37	0.04	0.38	0.04	0.36	0.04	0.38	0.04	0.03
12B	CO	2411	0.31	0.03	0.44	0.02	0.43	0.02	0.42	0.02	0.43	0.02	0.44	0.02	0.43	0.02	0.38	0.02	0.43	0.02	0.02
12F	CO	224	0.25	0.09	0.36	0.09	0.31	0.09	0.32	0.09	0.33	0.10	0.36	0.09	0.35	0.09	0.31	0.09	0.32	0.10	0.07
19D	CO	1035	0.40	0.04	0.47	0.04	0.47	0.03	0.46	0.03	0.46	0.04	0.45	0.04	0.47	0.04	0.45	0.04	0.48	0.03	0.03
19E	CO	2322	0.45	0.02	0.55	0.02	0.54	0.02	0.52	0.02	0.55	0.02	0.55	0.02	0.57	0.02	0.52	0.02	0.56	0.02	0.02
19F	CO	83	0.37	0.14	0.50	0.13	0.49	0.12	0.45	0.13	0.51	0.12	0.50	0.13	0.51	0.13	0.45	0.14	0.49	0.11	0.11
17K	EL	179	0.47	0.10	0.53	0.10	0.49	0.08	0.52	0.08	0.47	0.10	0.49	0.11	0.52	0.11	0.50	0.11	0.52	0.08	0.07
26Q	EL	142	0.42	0.10	0.47	0.09	0.54	0.09	0.49	0.09	0.53	0.09	0.46	0.08	0.47	0.09	0.46	0.10	0.52	0.10	0.08
27E	EL	305	0.52	0.07	0.52	0.04	0.52	0.06	0.53	0.04	0.51	0.05	0.50	0.05	0.54	0.06	0.55	0.06	0.53	0.06	0.06
31J	EL	130	0.46	0.12	0.58	0.11	0.58	0.13	0.59	0.12	0.55	0.13	0.51	0.12	0.55	0.12	0.51	0.13	0.61	0.12	0.09
31M	EL	1858	0.45	0.03	0.58	0.02	0.40	0.02	0.56	0.02	0.60	0.02	0.57	0.02	0.58	0.02	0.52	0.02	0.59	0.02	0.02
31N	EL	193	0.36	0.11	0.30	0.11	0.29	0.12	0.28	0.11	0.28	0.11	0.31	0.11	0.34	0.12	0.36	0.11	0.30	0.12	0.07
31V	EL	650	0.42	0.05	0.52	0.04	0.54	0.04	0.52	0.04	0.53	0.04	0.52	0.04	0.51	0.04	0.47	0.04	0.52	0.04	0.04
35K	EL	121	0.60	0.09	0.46	0.07	0.44	0.08	0.65	0.07	0.44	0.07	0.46	0.07	0.67	0.07	0.66	0.08	0.63	0.08	0.09
36C	EL	374	0.24	0.09	0.25	0.09	0.24	0.10	0.26	0.09	0.22	0.09	0.23	0.09	0.25	0.09	0.27	0.09	0.21	0.10	0.05
36K	EL	1581	0.30	0.04	0.43	0.04	0.41	0.05	0.40	0.04	0.43	0.04	0.42	0.04	0.42	0.04	0.37	0.04	0.42	0.04	0.03
11B	FA	4778	0.35	0.02	0.44	0.02	0.45	0.02	0.44	0.02	0.44	0.01	0.45	0.02	0.46	0.02	0.41	0.02	0.45	0.02	0.01
13F	FA	824	0.56	0.03	0.67	0.02	0.63	0.02	0.66	0.03	0.62	0.02	0.63	0.03	0.65	0.03	0.62	0.03	0.64	0.02	0.03
41C	GM	103	0.34	0.20	0.38	0.14	0.36	0.16	0.39	0.16	0.33	0.14	0.37	0.16	0.35	0.18	0.36	0.20	0.31	0.17	0.10
43E	GM	99	0.37	0.14	0.36	0.17	0.41	0.15	0.36	0.17	0.41	0.15	0.38	0.15	0.39	0.15	0.40	0.16	0.39	0.15	0.10
44B	GM	137	0.33	0.13	0.41	0.14	0.39	0.14	0.38	0.13	0.38	0.15	0.41	0.14	0.41	0.15	0.38	0.14	0.37	0.15	0.09
45K	GM	228	0.49	0.08	0.56	0.08	0.51	0.09	0.54	0.08	0.50	0.09	0.55	0.08	0.54	0.09	0.53	0.08	0.51	0.09	0.07
51B	GM	195	0.27	0.10	0.37	0.08	0.35	0.08	0.36	0.08	0.38	0.08	0.37	0.09	0.39	0.09	0.34	0.10	0.38	0.08	0.07
51K	GM	167	0.33	0.12	0.44	0.11	0.45	0.12	0.43	0.11	0.45	0.12	0.44	0.11	0.42	0.12	0.40	0.12	0.41	0.12	0.08
52D	GM	176	0.31	0.09	0.42	0.09	0.38	0.09	0.40	0.08	0.39	0.10	0.43	0.09	0.44	0.09	0.37	0.09	0.42	0.09	0.08
55B	GM	366	0.46	0.07	0.56	0.06	0.60	0.04	0.57	0.06	0.58	0.07	0.54	0.06	0.55	0.06	0.51	0.07	0.59	0.06	0.05
57E	GM	126	0.32	0.24	0.38	0.27	0.26	0.30	0.38	0.27	0.23	0.30	0.28	0.28	0.30	0.28	0.33	0.26	0.26	0.30	0.09
57H	GM	224	0.38	0.13	0.41	0.14	0.42	0.15	0.42	0.13	0.42	0.15	0.39	0.15	0.43	0.14	0.41	0.14	0.44	0.14	0.07
62E	GM	230	0.49	0.07	0.60	0.06	0.58	0.05	0.56	0.06	0.60	0.06	0.62	0.06	0.61	0.06	0.56	0.07	0.58	0.05	0.07
62F	GM	200	0.45	0.09	0.54	0.08	0.52	0.09	0.49	0.09	0.55	0.09	0.59	0.08	0.57	0.09	0.52	0.09	0.53	0.09	0.07
68J	GM	188	0.45	0.10	0.61	0.09	0.57	0.09	0.55	0.08	0.61	0.09	0.62	0.09	0.61	0.10	0.55	0.10	0.59	0.09	0.07
68M	GM	132	0.54	0.09	0.65	0.07	0.57	0.09	0.62	0.07	0.59	0.09	0.63	0.08	0.62	0.08	0.61	0.09	0.57	0.08	0.09

(Cont'd)

ADJUSTED VALIDITIES FOR CURRENT COMPOSITES
COMBINED CRITERION

MOS	AA	N	CL	SE	CO	SE	EL	SE	FA	SE	GM	SE	MM	SE	OP	SE	SC	SE	ST	SE	NORM SE
12C	MM	355	0.38	0.08	0.46	0.07	0.42	0.06	0.45	0.07	0.42	0.06	0.43	0.07	0.44	0.07	0.42	0.07	0.44	0.07	0.05
61B	MM	183	0.61	0.06	0.68	0.05	0.66	0.04	0.66	0.04	0.65	0.04	0.64	0.06	0.71	0.05	0.66	0.06	0.71	0.04	0.07
61C	MM	136	0.53	0.07	0.69	-0.09	0.66	0.08	0.70	0.07	0.64	0.09	0.62	0.10	0.62	0.10	0.59	0.09	0.65	0.08	0.09
62B	MM	355	0.41	0.05	0.56	0.04	0.51	0.05	0.52	0.05	0.54	0.05	0.53	0.05	0.53	0.05	0.50	0.05	0.52	0.05	0.05
63B	MM	1818	0.31	0.03	0.45	0.03	0.43	0.02	0.41	0.03	0.46	0.02	0.47	0.03	0.45	0.03	0.39	0.03	0.42	0.02	0.02
63D	MM	342	0.28	0.12	0.34	0.11	0.28	0.11	0.28	0.11	0.32	0.11	0.32	0.12	0.34	0.12	0.34	0.13	0.29	0.11	0.05
63Q	MM	161	0.23	0.13	0.38	0.13	0.30	0.12	0.30	0.12	0.34	0.13	0.38	0.14	0.36	0.14	0.31	0.14	0.29	0.13	0.08
63H	MM	783	0.34	0.05	0.48	0.04	0.48	0.04	0.45	0.04	0.49	0.04	0.48	0.04	0.47	0.04	0.41	0.04	0.48	0.04	0.04
63N	MM	509	0.18	0.07	0.33	0.07	0.28	0.07	0.28	0.07	0.32	0.07	0.35	0.07	0.32	0.07	0.26	0.08	0.27	0.07	0.04
63W	MM	527	0.23	0.07	0.29	0.07	0.29	0.08	0.27	0.07	0.31	0.08	0.28	0.08	0.29	0.08	0.27	0.08	0.31	0.07	0.04
63Y	MM	238	0.37	0.12	0.49	0.14	0.44	0.13	0.44	0.13	0.49	0.13	0.50	0.15	0.50	0.14	0.45	0.14	0.46	0.13	0.06
67N	MM	471	0.50	0.05	0.61	0.05	0.60	0.05	0.60	0.05	0.60	0.05	0.60	0.06	0.60	0.06	0.56	0.06	0.61	0.05	0.05
67T	MM	124	0.73	0.03	0.82	0.04	0.85	0.02	0.82	0.03	0.83	0.03	0.79	0.04	0.82	0.04	0.79	0.03	0.85	0.03	0.09
67U	MM	278	0.43	0.08	0.61	0.08	0.56	0.07	0.59	0.06	0.57	0.08	0.57	0.09	0.57	0.09	0.52	0.08	0.56	0.07	0.06
67V	MM	310	0.54	0.07	0.62	0.07	0.60	0.06	0.60	0.06	0.59	0.07	0.60	0.08	0.61	0.08	0.59	0.07	0.60	0.07	0.06
67Y	MM	241	0.52	0.09	0.63	0.09	0.63	0.07	0.56	0.08	0.67	0.07	0.68	0.09	0.67	0.09	0.62	0.09	0.63	0.08	0.06
68D	MM	121	0.53	0.09	0.74	0.07	0.66	0.07	0.71	0.06	0.67	0.08	0.69	0.08	0.69	0.08	0.62	0.09	0.66	0.07	0.09
68Q	MM	121	0.29	0.19	0.38	0.19	0.48	0.16	0.39	0.18	0.46	0.16	0.40	0.19	0.43	0.20	0.35	0.20	0.46	0.17	0.09
15D	OF	406	0.38	0.08	0.48	0.07	0.44	0.06	0.45	0.06	0.45	0.07	0.46	0.08	0.48	0.08	0.44	0.08	0.45	0.07	0.05
15E	OF	280	0.34	0.09	0.42	0.09	0.39	0.08	0.39	0.08	0.41	0.09	0.42	0.09	0.44	0.10	0.40	0.09	0.42	0.09	0.06
16B	OF	288	0.36	0.10	0.46	0.08	0.46	0.07	0.45	0.08	0.45	0.07	0.43	0.09	0.44	0.09	0.41	0.10	0.45	0.08	0.06
16C	OF	118	0.22	0.16	0.29	0.15	0.29	0.14	0.32	0.14	0.25	0.15	0.19	0.17	0.23	0.17	0.24	0.17	0.27	0.15	0.09
16D	OF	112	0.46	0.12	0.62	0.09	0.56	0.10	0.54	0.09	0.60	0.09	0.61	0.08	0.60	0.10	0.57	0.10	0.55	0.10	0.09
16H	OF	104	0.62	0.12	0.73	0.10	0.73	0.09	0.72	0.09	0.72	0.09	0.70	0.10	0.74	0.11	0.69	0.12	0.75	0.09	0.10
16J	OF	119	0.56	0.13	0.57	0.13	0.61	0.10	0.60	0.11	0.58	0.11	0.54	0.13	0.59	0.14	0.59	0.14	0.62	0.12	0.09
16R	OF	404	0.30	0.07	0.36	0.06	0.33	0.05	0.34	0.06	0.35	0.05	0.34	0.06	0.37	0.06	0.35	0.07	0.35	0.06	0.05
16S	OF	592	0.36	0.07	0.49	0.05	0.50	0.05	0.47	0.05	0.52	0.05	0.51	0.05	0.49	0.06	0.44	0.06	0.49	0.05	0.04
64C	OF	2959	0.37	0.02	0.48	0.02	0.47	0.02	0.44	0.02	0.47	0.02	0.47	0.02	0.47	0.02	0.43	0.02	0.47	0.02	0.02
94B	OF	3322	0.41	0.02	0.52	0.02	0.52	0.02	0.49	0.02	0.52	0.02	0.49	0.02	0.51	0.02	0.48	0.02	0.52	0.02	0.02
05B	SC	890	0.38	0.07	0.42	0.06	0.42	0.05	0.41	0.06	0.41	0.05	0.41	0.06	0.44	0.06	0.42	0.07	0.43	0.05	0.03
05C	SC	1971	0.38	0.04	0.48	0.03	0.49	0.03	0.47	0.03	0.48	0.03	0.47	0.03	0.47	0.04	0.43	0.04	0.48	0.03	0.02
05G	SC	119	0.64	0.13	0.72	0.09	0.61	0.11	0.68	0.10	0.62	0.10	0.67	0.10	0.71	0.11	0.70	0.12	0.67	0.12	0.09
17C	SC	187	0.35	0.14	0.38	0.11	0.35	0.10	0.36	0.11	0.34	0.10	0.34	0.11	0.38	0.12	0.38	0.13	0.37	0.11	0.07
72E	SC	562	0.40	0.07	0.56	0.05	0.56	0.04	0.52	0.05	0.59	0.04	0.58	0.05	0.56	0.06	0.49	0.07	0.56	0.05	0.04
05H	ST	110	0.75	0.07	0.72	0.06	0.70	0.07	0.75	0.04	0.66	0.07	0.66	0.07	0.73	0.08	0.77	0.07	0.73	0.07	0.10
13E	ST	678	0.39	0.04	0.46	0.04	0.47	0.04	0.48	0.04	0.45	0.04	0.43	0.04	0.45	0.04	0.43	0.04	0.47	0.05	0.04
54E	ST	270	0.46	0.08	0.49	0.07	0.50	0.08	0.52	0.07	0.47	0.07	0.47	0.07	0.49	0.08	0.49	0.08	0.51	0.08	0.06
74D	ST	98	0.65	0.10	0.69	0.09	0.71	0.10	0.70	0.09	0.67	0.10	0.62	0.10	0.69	0.11	0.69	0.10	0.72	0.11	0.10
82C	ST	516	0.47	0.06	0.56	0.05	0.56	0.05	0.58	0.05	0.53	0.05	0.54	0.05	0.54	0.06	0.52	0.06	0.53	0.06	0.04
91C	ST	233	0.55	0.06	0.58	0.05	0.57	0.05	0.57	0.05	0.55	0.05	0.55	0.05	0.57	0.05	0.58	0.05	0.56	0.06	0.07
91E	ST	301	0.49	0.07	0.53	0.06	0.57	0.06	0.56	0.06	0.54	0.06	0.50	0.06	0.52	0.06	0.51	0.06	0.57	0.06	0.06
91P	ST	159	0.44	0.12	0.40	0.12	0.39	0.12	0.43	0.12	0.36	0.12	0.41	0.11	0.41	0.13	0.43	0.12	0.38	0.14	0.08
91R	ST	145	0.61	0.09	0.59	0.09	0.56	0.11	0.59	0.09	0.55	0.10	0.54	0.09	0.59	0.10	0.64	0.09	0.58	0.12	0.08
92B	ST	364	0.33	0.08	0.41	0.07	0.45	0.07	0.43	0.07	0.42	0.07	0.37	0.08	0.38	0.08	0.36	0.08	0.44	0.07	0.05
93H	ST	114	0.54	0.13	0.58	0.11	0.60	0.11	0.60	0.12	0.57	0.11	0.56	0.12	0.59	0.13	0.56	0.13	0.62	0.12	0.09
95B	ST	3695	0.50	0.02	0.58	0.02	0.59	0.02	0.57	0.02	0.58	0.02	0.56	0.02	0.58	0.02	0.55	0.02	0.59	0.02	0.02
96B	ST	172	0.60	0.07	0.62	0.05	0.70	0.05	0.64	0.05	0.65	0.06	0.57	0.06	0.63	0.07	0.63	0.07	0.70	0.06	0.08
98C	ST	106	0.68	0.08	0.73	0.07	0.73	0.08	0.78	0.06	0.68	0.08	0.63	0.08	0.70	0.09	0.71	0.08	0.75	0.09	0.07

small MOS. MOS 05H (N=110) for example, consistently has bootstrap estimates .03 lower than the classical estimate.

Table 2 contains the adjusted validities and their standard errors. In general, these standard errors are larger than the standard errors of the corresponding unadjusted coefficients. Most of the standard errors are between one and two times the standard errors of the corresponding sample correlations. There are a few instances in which the inflation factor is as large as three. These cases are limited to small MOS, e.g., 71D (N=114), 76X (N=158), and 57E (N=126).

In very large MOS the standard errors are necessarily small, but there is some inflation. For example, MOS 13B (N=4778) consistently showed standard errors that are double the standard errors of uncorrected validities, but the absolute magnitude of the standard error was only .02.

It appears that the largest standard errors are associated with MOS in which the criterion variable has a highly skewed distribution. However, the converse is not true. The presence of a highly skewed distribution for an MOS with few observations was not always associated with very high standard errors. Table 3 presents some descriptive statistics on the four small MOS with highly skewed distributions.

Table 3

Descriptive Statistics for 4 Selected MOS

MOS	Criterion	N	Median	Interquartile Range	Inflation Factor
76X	Training	158	94	2.00	3.0
57E	Training	126	98	1.00	3.0
16C	Training	118	95	2.63	1.8
16D	Training	112	96	2.38	1.2

It appears that it is not possible to predict from sample size and shape of the criterion distribution which MOS will have highly inflated standard errors. It is more clear that standard errors may by substantially larger when the sample size is small. In this project, the average sample size was approximately 600, while the smallest sample size used in the research was 100. In investigations involving fewer observations, it is highly advisable to estimate standard errors using a repeated replication method. In particular, if the increase in the absolute magnitue of the validity coefficient due to the correction is also accompanied by an increase in the standard error of estimate, this finding should be reported.

Case 2: Conventional Cross-validation.

Method

A major goal of this validation effort was the identification of new "optimal" composites for the Army MOS. This identification requires that the MOS in our sample be clustered so that sets of relatively homogeneous MOS can be found. We would then find the linear combination of ASVAB subtests that was "optimal" for each of these MOS clusters.

This strategy depends crucially on a dependable clustering algorithm. If the clusters of MOS produced by a program are unstable, then the resulting search for "optimal" composites would be of no value. Therefore, a method of evaluating the stability of the MOS clustering results was essential.

We chose to evaluate the stability of the clustering in two stages. Frist, we used cross-validation to determine the dependability of the matrix of similarities that was used as input to the clustering program. Second, if those results appeared to be dependable, we would compare the assignments of MOS to clusters using cross-validation and determine the level of agreement.

Results

The two subsamples for the cross-validation were obtained by sorting the records by a scrambled ID and assigning each successive record into a different subsample. For each subsample, ridge regression coefficients were used to estimate each person's expected performance in each of the 98 MOS represented in our "combined criterion" file. This produced two order 98 correlation (similarity) matrices. These two similarity matrices were compared by correlating each row of one with the corresponding row of the other. Because a row of the similarity matrix represents a profile of the similarities of the corresponding MOS to all other MOS, the resulting correlations indicated the stability of the profiles of MOS similarity profiles. Figure 1 is a plot of the distribution of these correlations.

It is obvious that these correlations were disappointing. The distribution is centered around .15 to .20. This indicated that the similarity matrix from the full sample was too unstable to support t planned empirical identification of clusters.

Before abandoning the empirical clustering entirely, we hypoth zed that the unfavorable results might be due to outliers and/or ceili effects. We therefore transformed the data to normal scores and carried out the same cross-validation. Regrettably, the results were unchanged: the average correlation was .15. This did not indicate sufficient stabili y two support an emprical approach to clustering. We therefore decided to modify our research plan and identify the optimal composites for the existing MOS clusters.

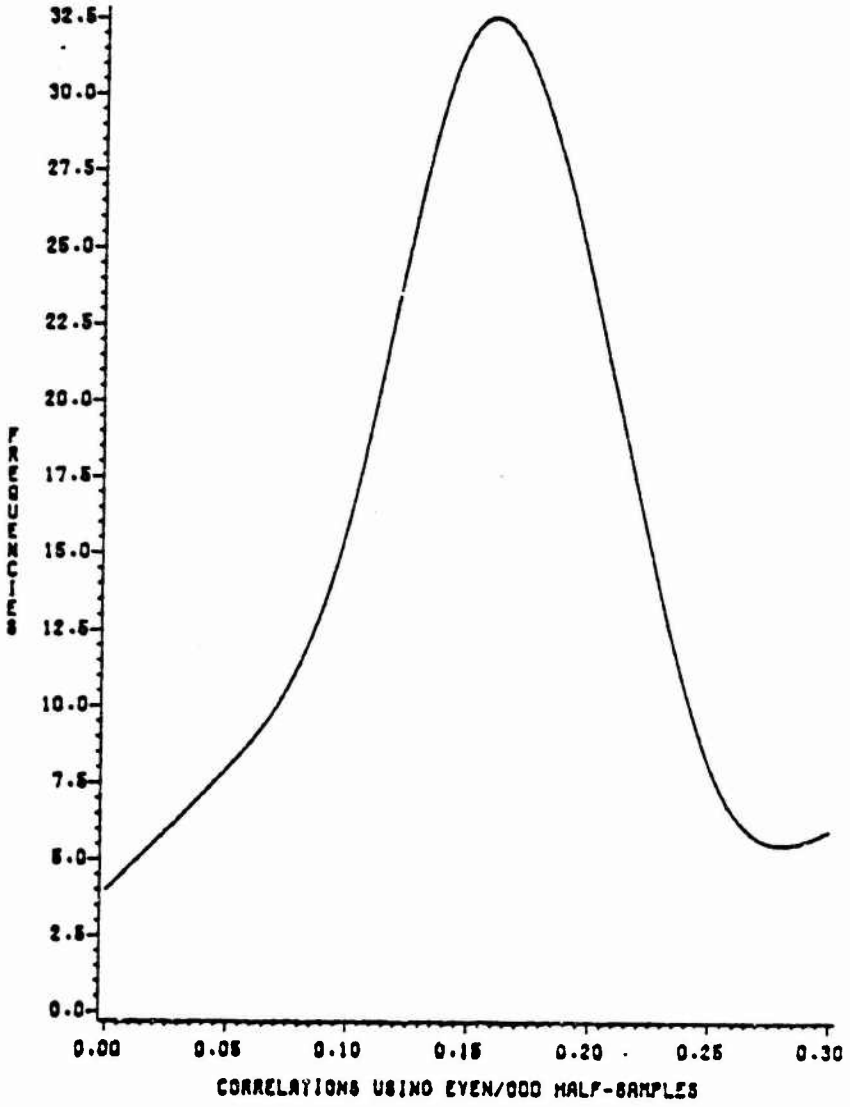


Figure 1. Frequency of Correlations of MOS Similarity Profiles.

The most likely reason for the instability of the similarity matrix was the highly skewed distribution of the similarity measures. The correlations among the expected performance scores were very high. Roughly three-forths were between .90 and 1.00. This reflected the high intercorrelations among the ASVAB subtests, the skewed criterion distributions, and the fact that the same subtests tended to be the best predictors for most MOS. Since similar regression equations were found in many MOS, little variability in expected performance scores across these MOS is possible. The cross-validation indicates that a major portion of that variability is not replicable across repeated samplings.

Summary

This paper describes two uses of repeated replication methods to assess the stability of sample statistics. In the investigation of the useability of the similarity matrix, we found that an elementary repeated replications method was sufficient to give a definitive answer. Sample statistics obtained from two orthogonal replications correlated so poorly that further work on empirical clustering was abandoned.

The evaluation of corrected validity coefficients was more complex. We needed a method of producing standard errors for these statistics. While the classical correction for restriction of range results in an increase in absolute validity, the accompanying increase in error of estimation is generally not reported. Some way of determining whether the increase in the level of validity is effectively offset by the decrease in precision is needed.

We found that the bootstrap method produced reasonable estimates of errors when compared to classical error estimates of sample correlations. The standard errors for corrected validities were generally between one and two times the standard errors of the corresponding sample correlations. Especially large increases in standard errors were found in relatively small MOS with skewed distributions of criterion scores. The standard errors of the very large MOS showed some inflation, but, since the absolute level of the standard error was so small, the increases were not important. Since the large MOS make the heaviest demands on computing power, we may in future work choose not to obtain bootstrap estimates for samples larger than a specificed size, say 1000.

In this validation, the smallest sample size included in the analysis was 100. In other work, where sample sizes are even smaller, the use of repeated replication techniques for variance estimation are likely to be even more important. The bootstrap or the jackknife can be implemented on the computer easily. The final method, BHS, while more difficult to program, is available from vendors. The availability of high speed and low cost computing power makes these methods practical alternatives.

References

Diaconis, P., & Efron, B. (1983) Computer-intensive methods in statistics. Scientific American, 248 (5), 116-130.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7 (1), 1-26.

Lawley, D. (1943) A note on Karl Pearson's selection formulae. Royal Society of Edinburgh, Proceedings, Section A., 62, 28-30.

Lord, F., & Novick, M. (1968) Statistical theory of mental test scores. Reading, Mass: Addison-Wesley.

McCarthy, P. J. (1976) The use of balanced half-sample replication in cross-validation studies. Journal of the American Statistical Association, 71, 596-604.

Mosteller, F., & Tukey, J. (1977) Data Analysis and Regression. Reading, Mass: Addison Wesley.

OSIRIS IV: Data Management and Statistical Software System. (1981) Ann Arbor, Mich: Institute for Social Research.

Plackett, R. L., & Burman, P. J. (1946) The design of optimum multifactorial experiments. Biometrika, 33, 305-325.

Wise, L. L. (1983) The PROC BRRVAR Procedure: Documentation. Palo Alto, CA: American Institutes for Research.

Subgroup Variation in the Validity of Army Aptitude Area Composites

Paul G. Rossmeissl
Army Research Institute

David A. Brandt
American Institutes for Research

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Paper presented at the Annual Convention of the American Psychological Association in Toronto, Canada.

An important scientific and policy issue is the question of predictive bias of the selection and classification procedures. The primary concern here is whether the use of an alternative set of ASVAB composites would lead to bias in the selection and classification of Army enlisted personnel. This question was addressed in three ways: First, the adjusted validities of the subgroups were calculated and compared. The subgroup validities were adjusted to the total applicant population rather than the separate subgroup populations. Second, the differences between the predicted scores for each subgroup were compared in the range of composite score values that contain the operational cutoff points. Third, the common and subgroup regression lines were plotted over this region. The sample regression lines were used as the basis for the latter two sets of comparisons. Unadjusted lines were used because the classical adjustment for restriction of range makes the assumption that the regression line in the selected group is the same as the regression line in the unselected population.

As noted in the earlier description of the data available for this research, we were limited in the analyses of subgroup differences to comparisons between race (blacks and whites) and between gender. We performed subgroup analyses only on those MOS that contained a sample of at least 100 soldiers of each subgroup. For race, this sample included 35 MOS and for gender it included 19 MOS. After the analyses had been obtained for each MOS, the results were aggregated to the cluster level. We will first discuss the analyses based upon comparisons between black and white soldiers and then turn to a discussion of analyses investigating differences as a function of gender.

Analyses of Differences by Race

The sample and adjusted validities of the current operational composites based upon the combined criterion as a function of race are presented in Table 1. Similar data based upon the proposed four alternative composites are given in Table 2.

Inspection of these two tables shows that, in general, both sets of composites predict performance in each of the subgroups well. The smallest adjusted validity in either table is a respectable .25, while the average adjusted validities are sizeable at .41 and .43 for the current and alternative composites respectively. While the validities in both tables are high, the validities obtained from the three alternative composites were consistently higher for both subgroups across all of the clusters.

Both tables show small differences between the validities obtained by whites in comparison to blacks. These differences are quite stable across the two different sets of composites. The average difference in adjusted validities between blacks and whites among the current composites was .08, while in the case of the alternative composites this value is slightly smaller at .07. The only sizeable changes in the black-white validity differences were found in GM and MM clusters, where the subgroup differences were .04 and .03 smaller for the alternative composites. The stability of these differences,

Table 1

Sample and Adjusted Validities for Blacks (B) and Whites (W):
Current Operational Composites
SQT and Training Criteria Combined

Cluster/ Composite	Sample Size W	Sample Size B	Sample Validities W	Sample Validities B	Adjusted Validities W	Adjusted Validities B	Difference (Adjusted)
CL	4780	6985	.30	.13	.51	.42	.09
CO	14523	3570	.30	.19	.44	.41	.03
EL	4527	3111	.26	.10	.43	.29	.14
FA	4936	3234	.36	.19	.56	.42	.14
GM	474	624	.20	.11	.41	.55	-.14
MM	2729	1039	.25	.12	.40	.34	.06
OF	6941	3316	.29	.14	.47	.39	.08
SC	3207	1708	.25	.11	.44	.30	.14
ST	6682	956	.27	.14	.41	.25	.16

Table 2

Sample and Adjusted Validities for Blacks (B) and Whites (W):
Four Alternative Composites
SQT and Training Criteria Combined

Cluster/ Composite	Sample Size W	Sample Size B	Sample Validities W	Sample Validities B	Adjusted Validities W	Adjusted Validities B	Difference (Adjusted)
CL/ACL	4780	6985	.41	.26	.57	.49	.08
CO/ACO	14523	3570	.31	.22	.45	.43	.02
EL/ACO	4527	3111	.27	.12	.44	.29	.15
FA/ACO	4936	3234	.37	.19	.57	.42	.15
GM/ACO	474	624	.29	.08	.46	.56	-.10
MM/AOP	2729	1939	.26	.18	.40	.37	.03
OF/AOP	6941	3316	.31	.22	.49	.42	.07
SC/AOP	3207	1708	.34	.22	.47	.33	.14
ST/AST	6682	956	.27	.18	.42	.26	.16

despite the radical changes in the makeup of the composites between the operational and the alternative sets, suggests that the small differences observed in ASVAB composite validities as a function of race are most likely attributable to the ASVAB subtests themselves or to the criterion measures, rather than to the way they are combined into composites.

Differences between subgroup validities such as those observed in Tables 1 and 2 above do not necessarily mean that either set of composites is culturally biased. Cronbach (1976) makes the distinction between equality of test validities and fairness in selection policies. The relationship of the subgroup regression lines to each other is the key issue in the analysis of predictive bias.

Clearly, predictive bias would not be an issue if both groups shared the same regression line. If this were true, each recruit would have the same predicted value on the criterion regardless of subgroup membership. Therefore, a natural way of investigating predictive bias is to identify values of the AA composite for which a significant difference in predicted criterion scores exists.

To compare the black and white regression lines, we calculated the predicted criterion scores for the two subgroups for composite scores ranging from 80 to 110 points. This range of values was selected because it contains all of the cutoff scores now in operational use by the Army. The two sets of predicted scores were then subtracted to obtain the difference score, and standard error of the difference was estimated using the formula for the variance of the difference given in Rogosa (1980). The differences between the two regression lines are given in Table 3 for the current operational composites and Table 4 for the proposed four alternative composites.

Inspection of these two tables shows that, in general, for both sets of composites the two subgroup regression lines tend to be close over this range of composite scores. The average differences between the two lines for the current composites are: 3.80 for the CL cluster, 2.76 for the CO cluster, 2.38 for the EL cluster, 4.88 for the FA cluster, 4.23 for the GM cluster, 3.40 for the MM cluster, .88 for the OF cluster, 5.93 for the SC cluster, and 2.56 for the ST cluster. The average differences for the proposed alternative composites were 1.57 (CL), 2.10 (CO), .89 (EL), 2.77 (FA), 4.70 (GM), 1.10 (MM), -.90 (OF), .90 (SC) and .87 (ST). While some of these differences and those given in the tables are statistically significant, they tend to be relatively small in comparison to the standard deviation of the combined SQT and training criterion, which had been standardized to a value of 20 for accessions into each MOS. Only for fairly high values of the composite scores (around 110) did the differences in predicted scores for the two subgroups become large. These findings are typical of the comparisons of black and white regression lines found in other educational, employment, and military research (i.e., Hanser & Grafton, 1983).

Tables 3 and 4 show that the relationships between the black and white regression lines are similar for both sets of composites. In both tables the differences most often have positive values, indicating that the white regression lines lie above the black regression lines. In other words, the black criterion scores are overpredicted by the regression line based upon the white subgroup. This relationship of average overprediction of the black

Table 3

Predicted Criterion Scores for Blacks (B) and Whites (W): Current Operational Composites

Composite Score	Predicted (Combined	Criterion Score B	W)	Subgroup Difference (W/-/B)	Standard Error of the Difference
CL Cluster					
80	88.44	90.03	90.14	.10	2.54
85	91.07	91.69	93.03	1.34	2.17
90	93.70	93.34	95.92	2.58	1.84
95	96.33	95.00	98.81	3.80*	1.57
100	98.95	96.66	101.70	5.04*	1.39
105	101.58	98.32	104.59	6.27*	1.35
110	104.21	99.97	107.48	7.50*	1.44
CO Cluster					
80	90.66	89.37	91.59	2.22	1.19
85	93.14	91.54	93.94	2.40*	1.14
90	95.61	93.71	96.29	2.58*	1.11
95	98.09	95.88	98.64	2.76*	1.11
100	100.57	98.05	100.99	2.94*	1.14
105	103.04	100.22	103.34	3.12*	1.19
110	105.52	102.39	105.69	3,30*	1.26
EL Cluster					
80	90.18	91.50	90.86	-.63	2.23
85	93.24	93.60	93.97	.37	1.88
90	96.31	95.70	97.07	1.37	1.60
95	99.37	97.80	100.18	2.38	1.43
100	102.44	99.90	103.28	3.38*	1.42
105	105.50	102.00	106.39	4.39*	1.56
110	108.57	104.10	109.49	5.39*	1.81
FA Cluster					
80	90.07	90.33	92.54	2.21	1.14
85	93.12	92.30	95.38	3.08*	1.01
90	96.18	94.26	98.22	3.95*	.91
95	99.23	96.23	101.06	4.82*	.83
100	102.29	98.20	103.90	5.69*	.80
105	105.34	100.17	106.74	6.56*	.81
110	108.40	102.14	109.58	7.44*	.87

(cont'd)

Predicted Criterion Scores for Blacks (B) and Whites (W):
Current Operational Composites (Continued)

Composite Score	Predicted Criterion Score (Combined	B	W)	Subgroup Difference (W/-/B)	Standard Error of the Difference
GM Cluster					
80	95.09	97.24	95.84	-1.40	2.50
85	98.53	98.78	99.26	.48	2.27
90	101.97	100.32	102.68	2.36	2.35
95	105.41	101.86	106.09	4.23	2.69
100	108.85	103.40	109.51	6.12	3.22
105	112.29	104.93	112.93	8.00*	3.86
110	115.73	106.47	116.35	9.88*	4.56
MM Cluster					
80	91.50	89.84	93.74	3.90*	1.60
85	94.12	92.29	96.02	3.73*	1.48
90	96.74	94.75	98.31	3.56*	1.42
95	99.36	97.20	100.60	3.40*	1.41
100	101.98	99.65	102.88	3.23*	1.45
105	104.60	102.10	105.17	3.06.	1.56
110	107.22	104.56	107.46	2.90	1.70
OF Cluster					
80	93.14	90.73	94.87	4.14*	1.17
85	96.03	94.38	97.43	3.05*	1.11
90	98.91	98.03	100.00	1.97	1.10
95	101.80	101.68	102.56	.88	1.45
100	104.68	105.32	105.12	-.20	1.25
105	107.57	108.97	107.69	-1.29	1.39
110	110.45	112.62	110.25	-2.37	1.56
SC Cluster					
80	89.64	86.83	93.82	6.99*	2.09
85	92.25	89.32	95.96	6.64*	1.82
90	94.87	91.81	98.09	6.28*	1.59
95	97.48	94.30	100.22	5.92*	1.41
100	100.09	96.79	102.36	5.57*	1.29
105	102.71	99.28	104.49	5.22*	1.27
110	105.32	101.76	106.62	4.86*	1.34
ST Cluster					
80	85.52	85.78	86.02	.24	1.30
85	88.54	87.98	89.00	1.02	1.26
90	91.56	90.19	91.98	1.79	1.24
95	94.58	92.40	94.96	2.56*	1.25
100	97.60	94.61	97.95	3.34*	1.29
105	100.62	96.82	100.93	4.12*	1.35
110	103.64	92.02	103.91	4.89*	1.44

*p .05

Table 4

Predicted Criterion Scores for Blacks (B) and Whites (W): Four Alternative Composites

Composite Score	Predicted Criterion Score (Combined	B	W)	Subgroup Difference (W/-/B)	Standard Error of the Difference
CL Cluster					
80	93.10	92.90	92.36	-.54	2.43
85	96.08	95.47	95.63	.16	2.08
90	99.06	98.04	98.91	.87	1.76
95	102.04	100.61	102.18	1.57	1.50
100	105.02	103.18	105.45	2.27	1.33
105	107.99	105.75	108.72	2.98*	1.29
110	110.97	108.32	112.00	3.68*	1.37
CO Cluster					
80	92.17	91.22	92.35	1.13	1.18
85	94.46	93.17	94.62	1.45	1.13
90	96.74	95.11	96.89	1.78	1.11
95	99.03	97.06	99.16	2.10	1.11
100	101.32	99.00	101.43	2.42*	1.13
105	103.60	100.95	103.69	2.74*	1.18
110	105.89	102.90	105.96	3.07*	1.25
EL Cluster					
80	95.35	95.58	94.87	-.71	2.23
85	97.71	97.52	97.34	-.17	1.88
90	100.06	99.46	99.82	.36	1.60
95	102.42	101.39	102.29	.90	1.43
100	104.77	103.33	104.76	1.43	1.42
105	107.12	105.27	107.23	1.96	1.56
110	109.48	107.21	109.70	2.50	1.81
FA Cluster					
80	95.02	94.92	95.08	.16	1.14
85	97.42	96.50	97.53	1.03	1.01
90	99.81	98.08	99.99	1.90*	.90
95	102.20	99.67	102.44	2.77*	.83
100	104.59	101.25	104.89	3.64*	.80
105	106.99	102.83	107.35	4.52*	.81
110	109.38	104.41	109.80	5.39*	.87

(cont'd)

Predicted Criterion Scores for Blacks (B) and Whites (W): Four Alternative Composite Solution (Continued)

Composite Score	Predicted Criterion Score (Combined	B	W)	Subgroup Difference (W/-/B)	Standard Error of the Difference
GM Cluster					
80	97.92	98.69	97.73	-.95	2.49
85	100.62	99.78	100.71	.93	2.27
90	103.32	100.86	103.68	2.82	2.34
95	106.02	101.96	106.66	4.70	2.68
100	108.72	103.05	109.64	6.59*	3.21
105	111.41	104.14	112.61	8.48*	3.85
110	114.11	105.23	115.59	10.36*	4.55
MM Cluster					
80	93.55	91.71	94.80	3.09	1.59
85	95.88	94.49	96.92	2.43	1.48
90	98.22	97.28	99.04	1.76	1.41
95	100.55	100.06	101.16	1.10	1.40
100	102.88	102.84	103.27	.44	1.45
105	105.21	105.62	105.39	-.23	1.55
110	107.54	108.40	107.51	-.89	1.70
OF Cluster					
80	92.84	91.40	93.82	2.42*	1.16
85	95.76	95.19	96.50	1.31	1.09
90	98.68	98.98	99.19	.21	1.08
95	101.60	102.78	101.88	-.90	1.13
100	104.52	106.57	104.56	-2.00	1.23
105	107.44	110.36	107.25	-3.11*	1.37
110	110.35	114.15	109.94	-4.21*	1.54
SC Cluster					
80	92.59	91.30	93.45	2.16	2.00
85	95.17	94.13	95.87	1.74	1.75
90	97.74	96.96	98.28	1.32	1.52
95	100.31	99.79	100.69	.90	1.35
100	102.89	102.62	103.10	.49	1.24
105	105.46	105.44	105.51	.07	1.22
110	108.03	108.28	107.93	-.35	1.28
ST Cluster					
80	85.20	88.38	85.19	-3.19*	1.29
85	88.37	90.22	88.38	-1.84	1.25
90	91.54	92.06	91.58	-.48	1.23
95	94.70	93.90	94.77	.87	1.24
100	97.87	95.74	97.96	2.22	1.28
105	101.04	97.58	101.16	3.58*	1.34
110	104.21	99.42	104.35	4.93*	1.43

* p .05

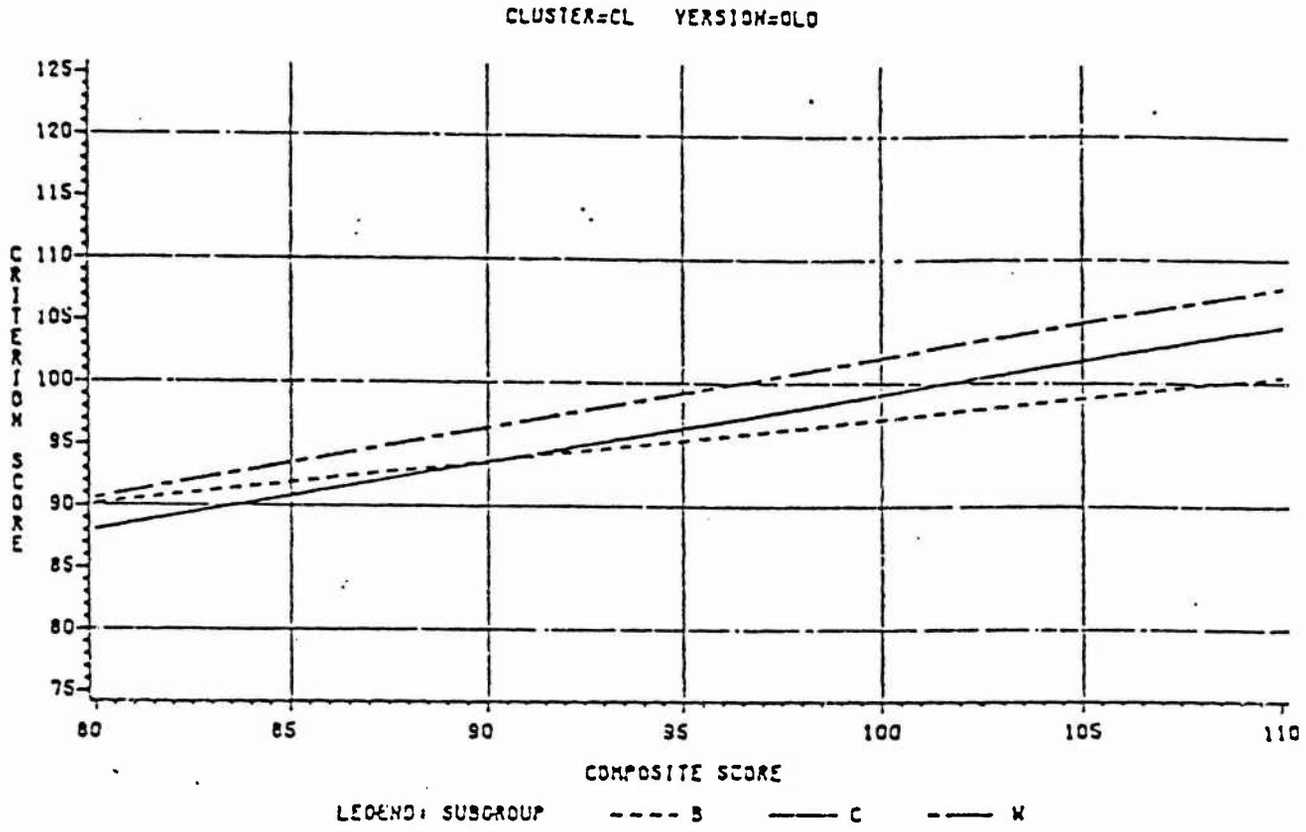
regression line by the white regression line across this range of composite scores was true for all of the current composites and all but one (OF) of the proposed alternative composites.

The alternative composites differ from the current operational set in two ways. Overall, the differences in predicted criterion scores observed in the alternative composites are smaller than the differences found with the operational composites. The average of the absolute values of differences from the current composites is 3.42, while the proposed alternative composites show an average absolute value of the differences of 1.76. Again, both of these values are fairly small when compared to a criterion standard deviation of 20. The other noticeable aspect in which the two sets of composites differ has already been noted above. When the alternative OF (AOP) composite is used to predict performance for the OF cluster of MOS, the white regression line tends to slightly underpredict rather than overpredict the black regression line. Tables 3 and 4 show that the basic pattern of general overprediction of the black regression line by the white regression with some underprediction for low composite scores is the case for both the operational and the alternative composites.

Given that the Army does not use separate black and white regression lines line to the common regression line becomes important when significant differences between the subgroup lines exist. If the criterion scores for a subgroup are substantially underpredicted by the common regression line (e.g., the subgroup line falls above the common line), use of the common line to select and classify potential personnel would be unfair to that subgroup since its "true" predicted criterion would be higher than the value predicted by the common selection/classification instrument.

Underprediction of any subgroup is a serious problem only when the underprediction is for values of the composite near the cutoff point for that MOS. This is true because an individual is able to enlist in his or her MOS of choice as long as his or her composite score is above the appropriate cutoff. Composite scores well above the cutoff do not have any real meaning to the system. For example, if two individuals with composite scores of 95 and 105, respectively, wished to enter an MOS with a cutoff score of 90, both would be allowed to enlist in the MOS. The ten-point difference in their composite scores would not affect either person's selection or classification.

To investigate the relationships between the subgroup regressions and the common regression lines, we plotted the black, white, and common lines in the region that contains the cutoff scores for the Army MOS. These plots are presented in Figures 1 through 9. These plots show that the predicted values of all three lines tend to have higher slopes for the alternative composites than for the current composites. This finding is in agreement with the earlier validity data which showed somewhat higher criterion predictability with the use of the alternative composites. In each of the figures, the plots based upon the alternative composites tend to show the three lines being closer together than they are in the plots obtained from the current composites. This is consistent with Table 4, which showed that the alternative composites have the smaller differences among the predicted criterion scores from the two subgroups.



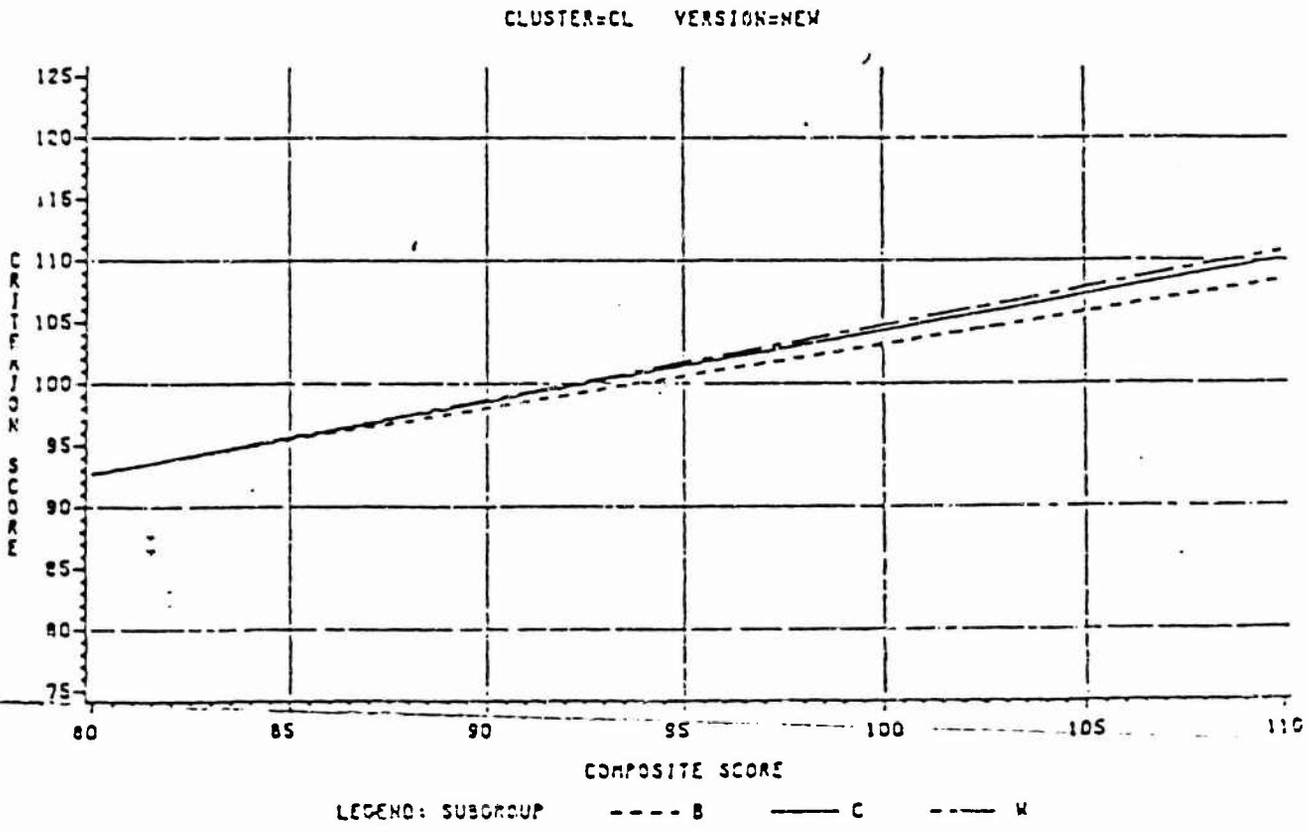
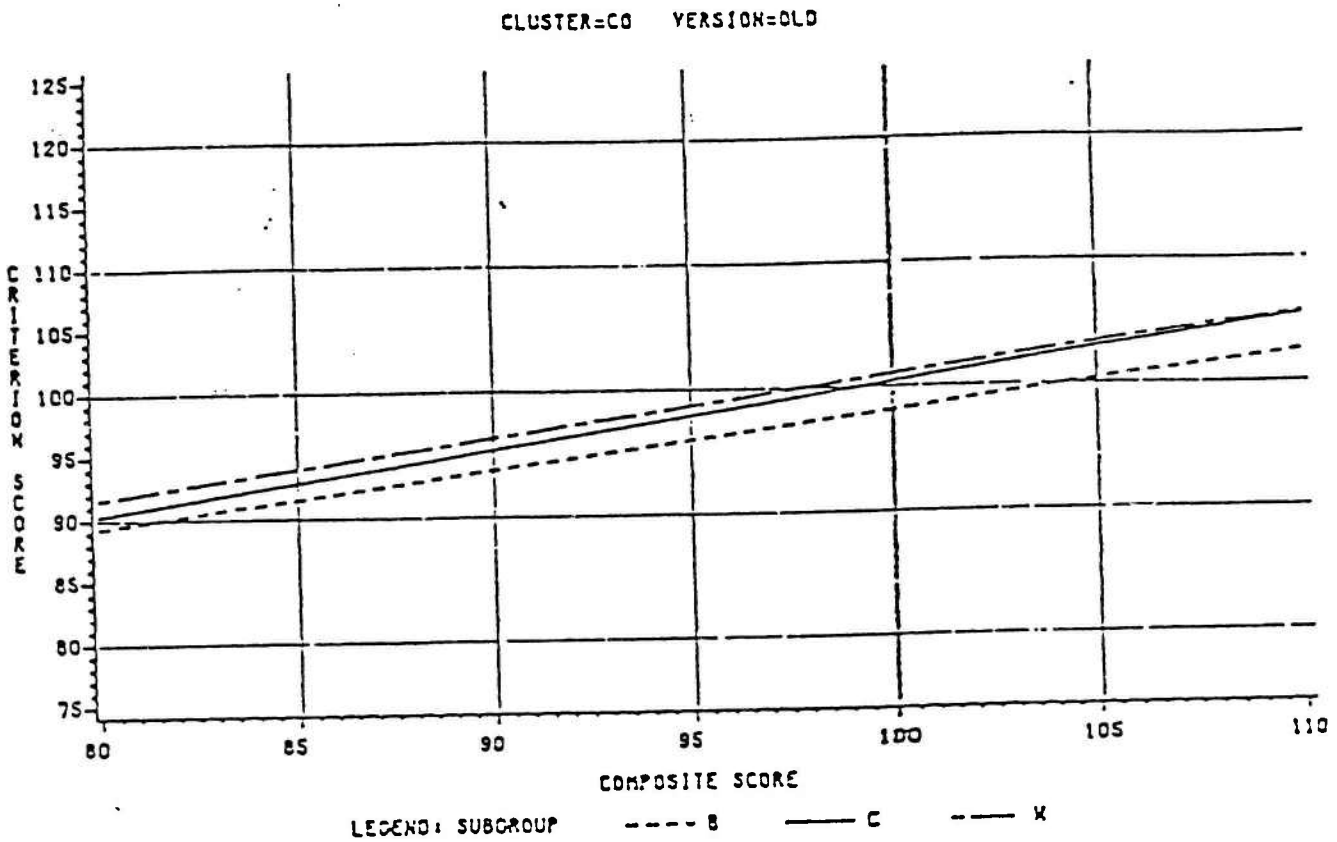


Figure 1. Regression lines for Current and Alternative (old and new) Composites for CL MOS, by Race



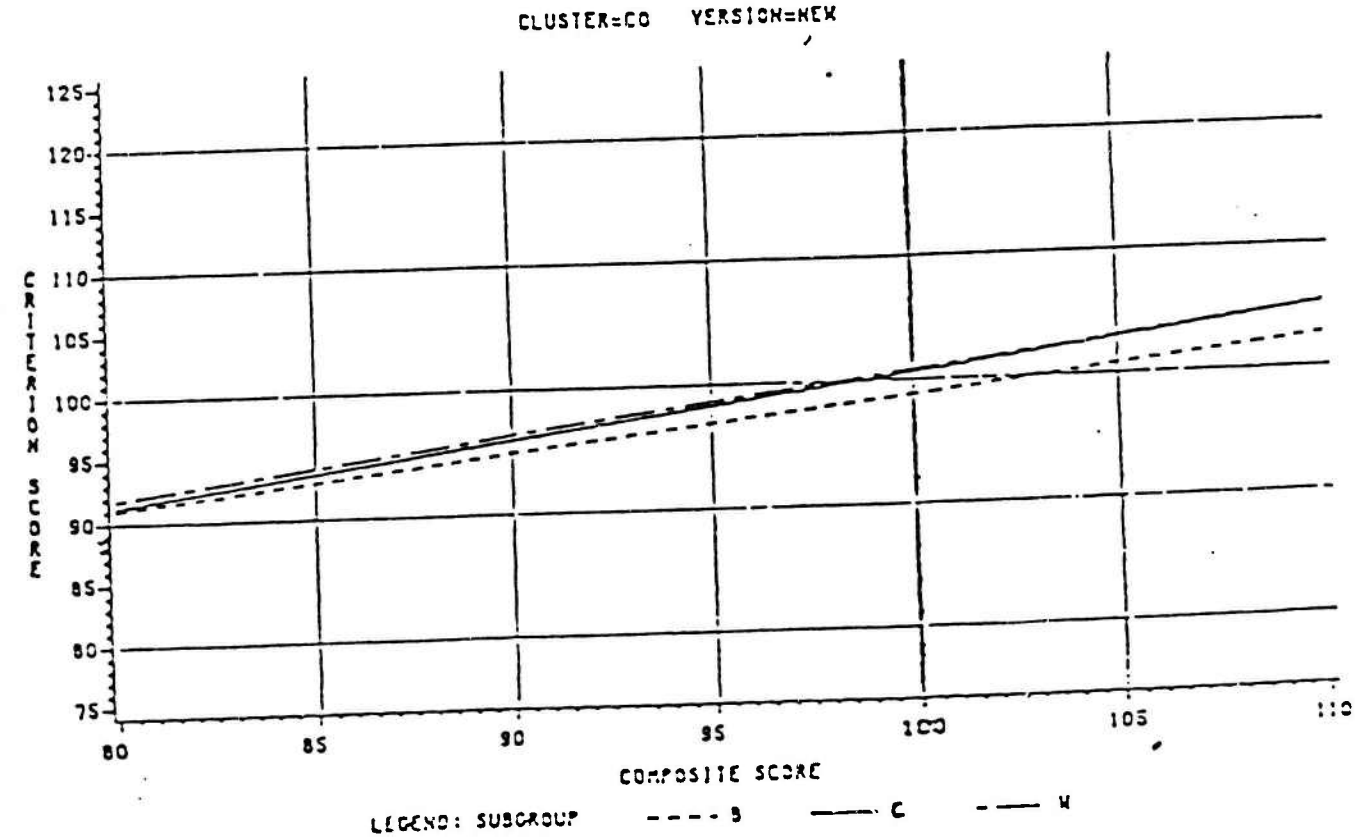
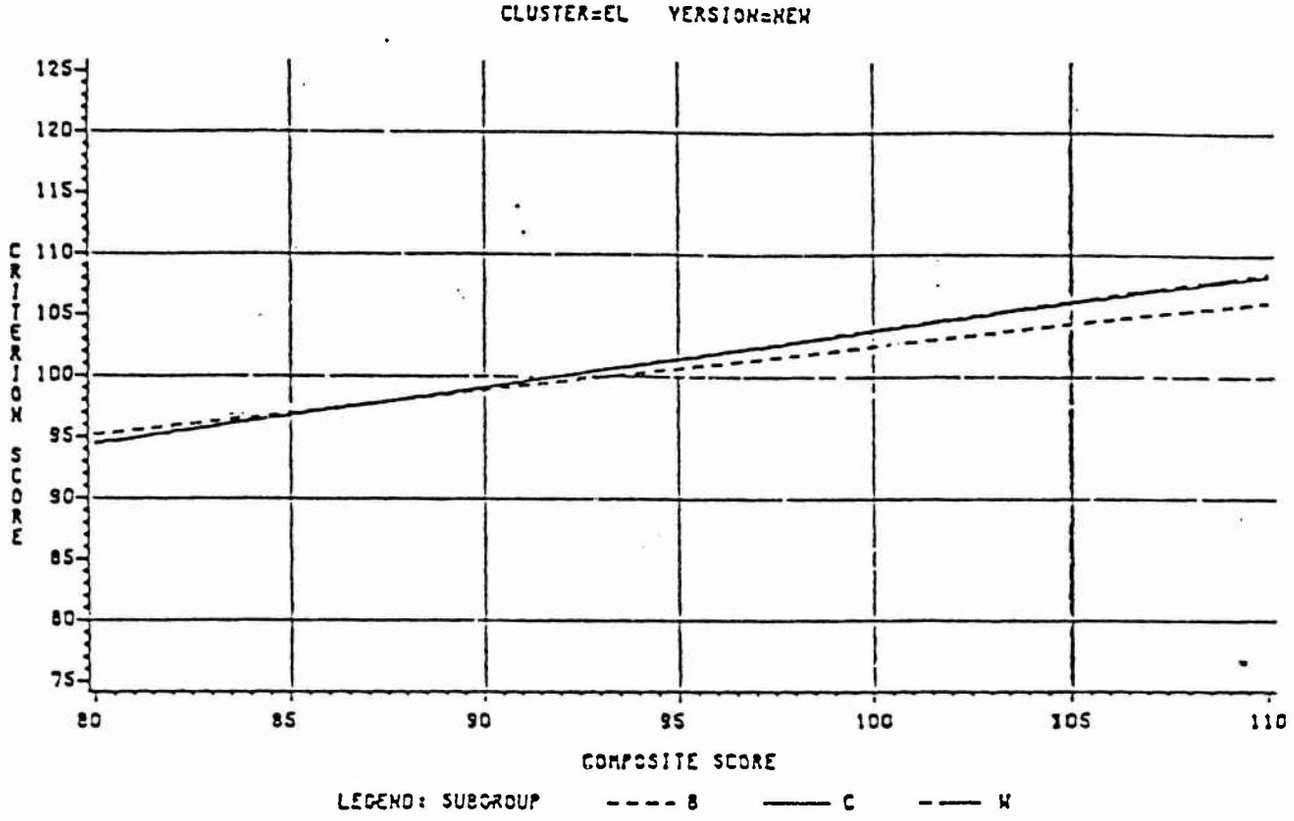


Figure 2. Regression lines for Current and Alternative (old and new) Composites for CO MOS, by Race



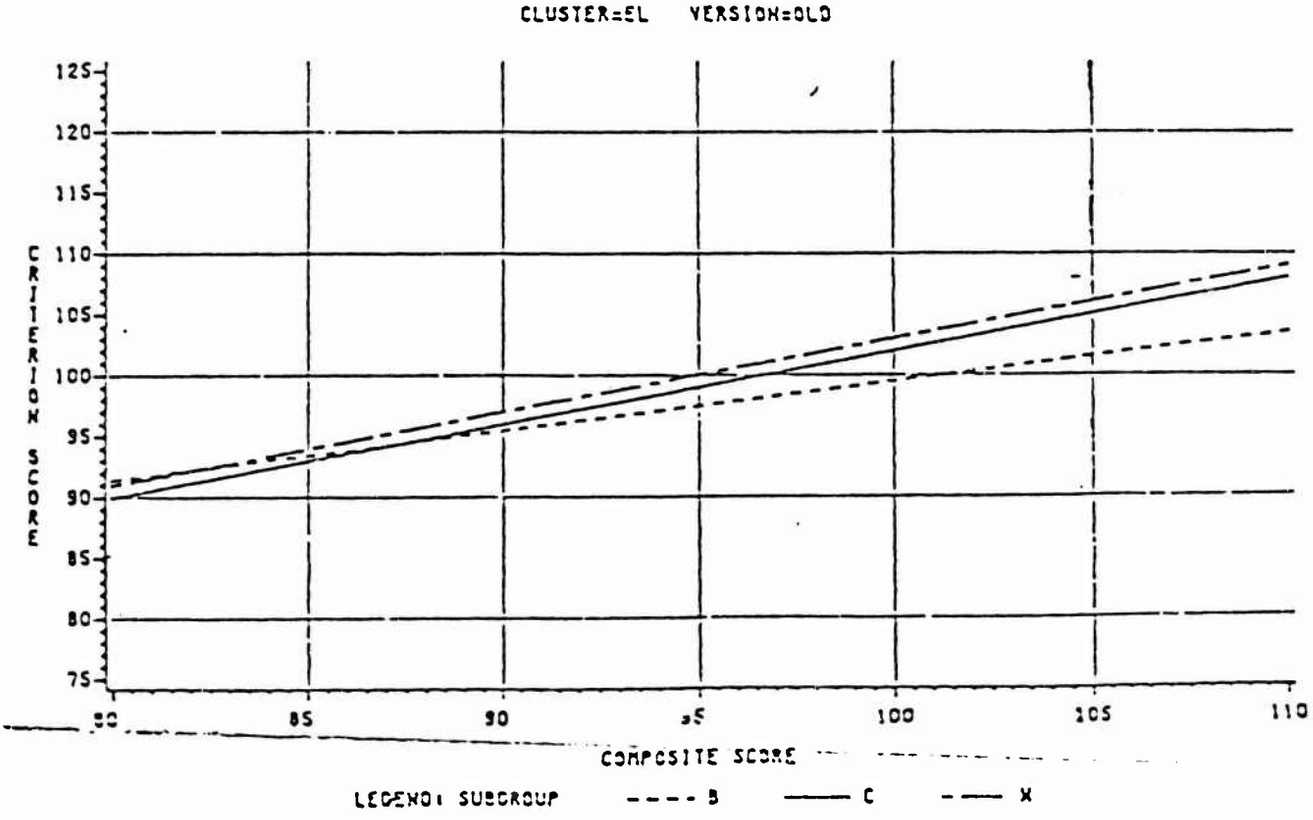
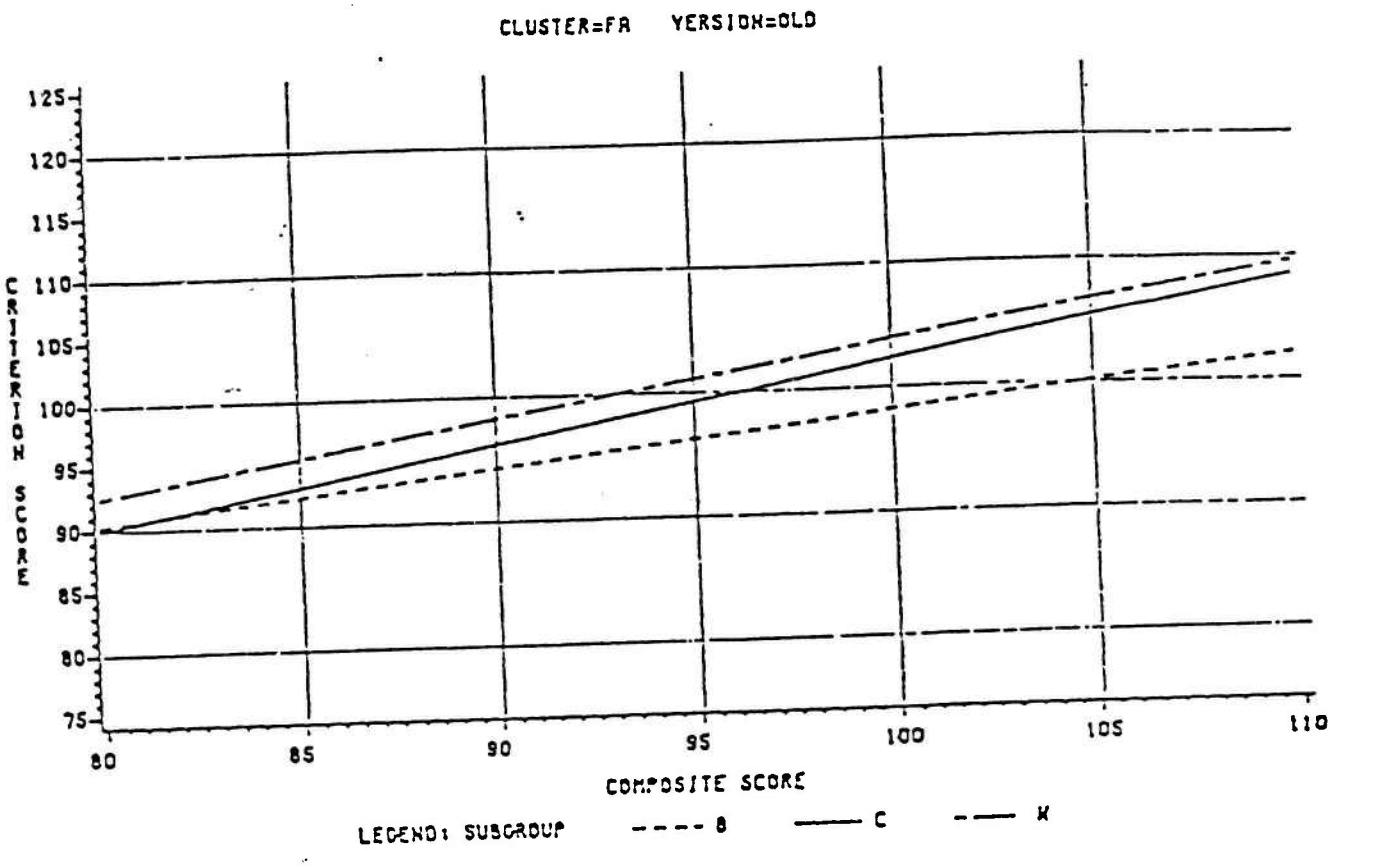


Figure 3. Regression lines for Current and Alternative (old and new) Composites for EL MOS, by Race



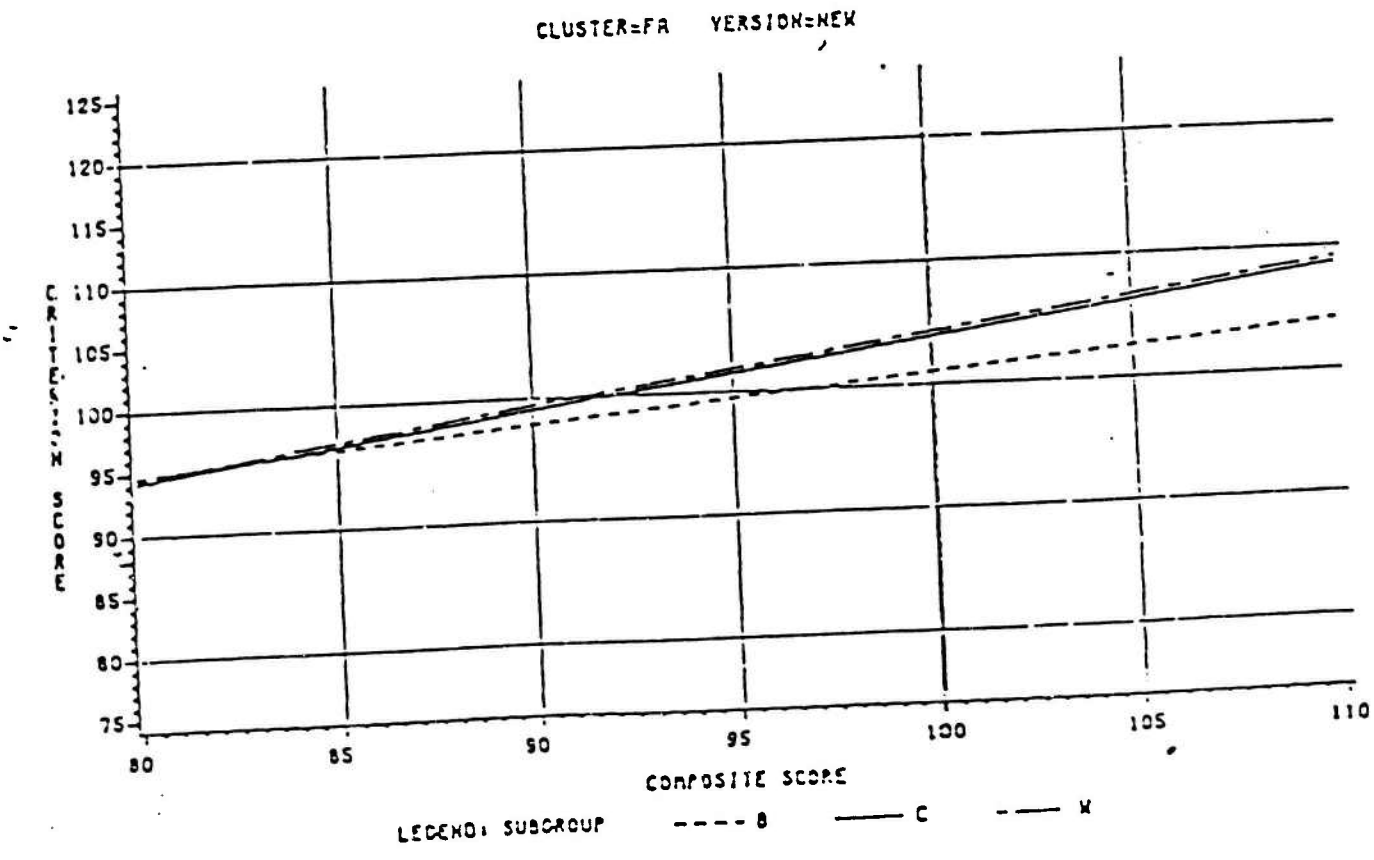
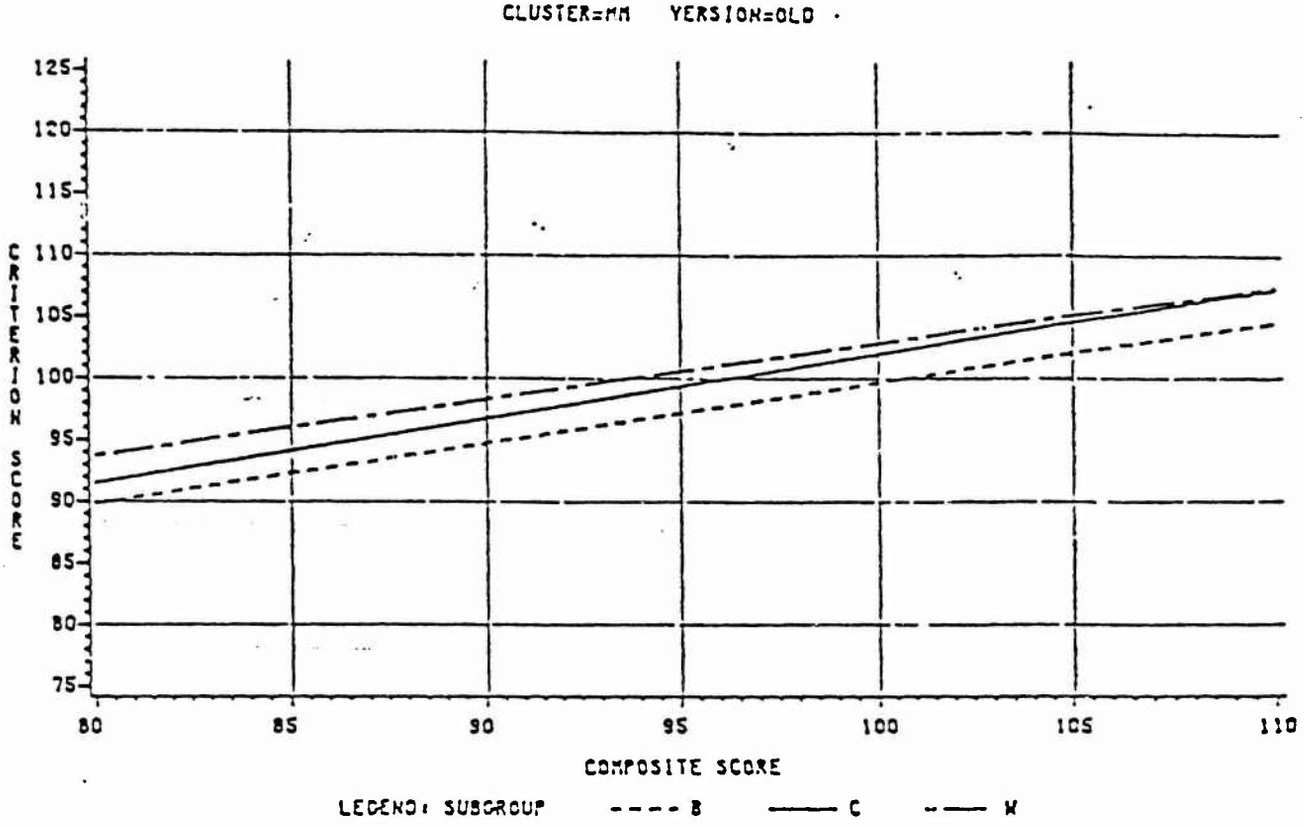


Figure 4. Regression lines for Current and Alternative (old and new) Composites for FA MOS, by Race



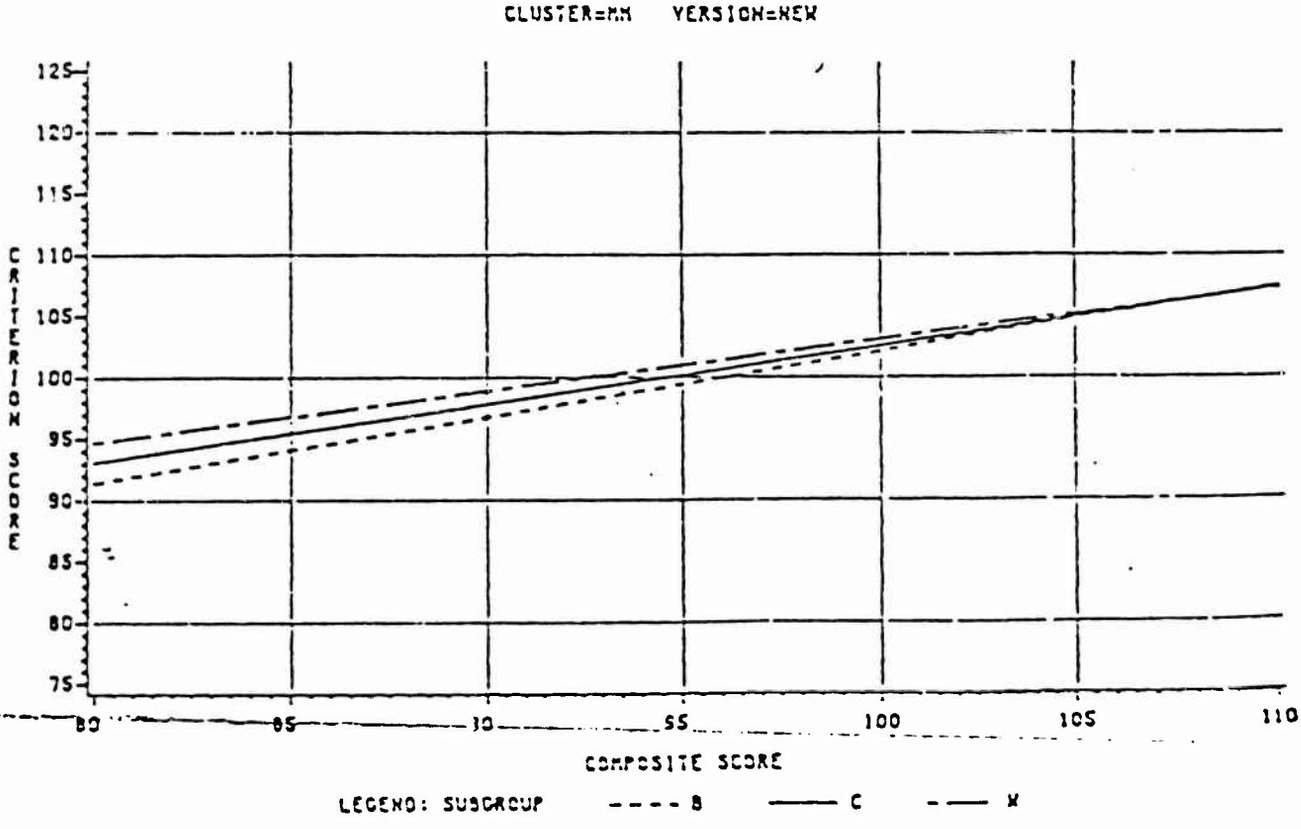
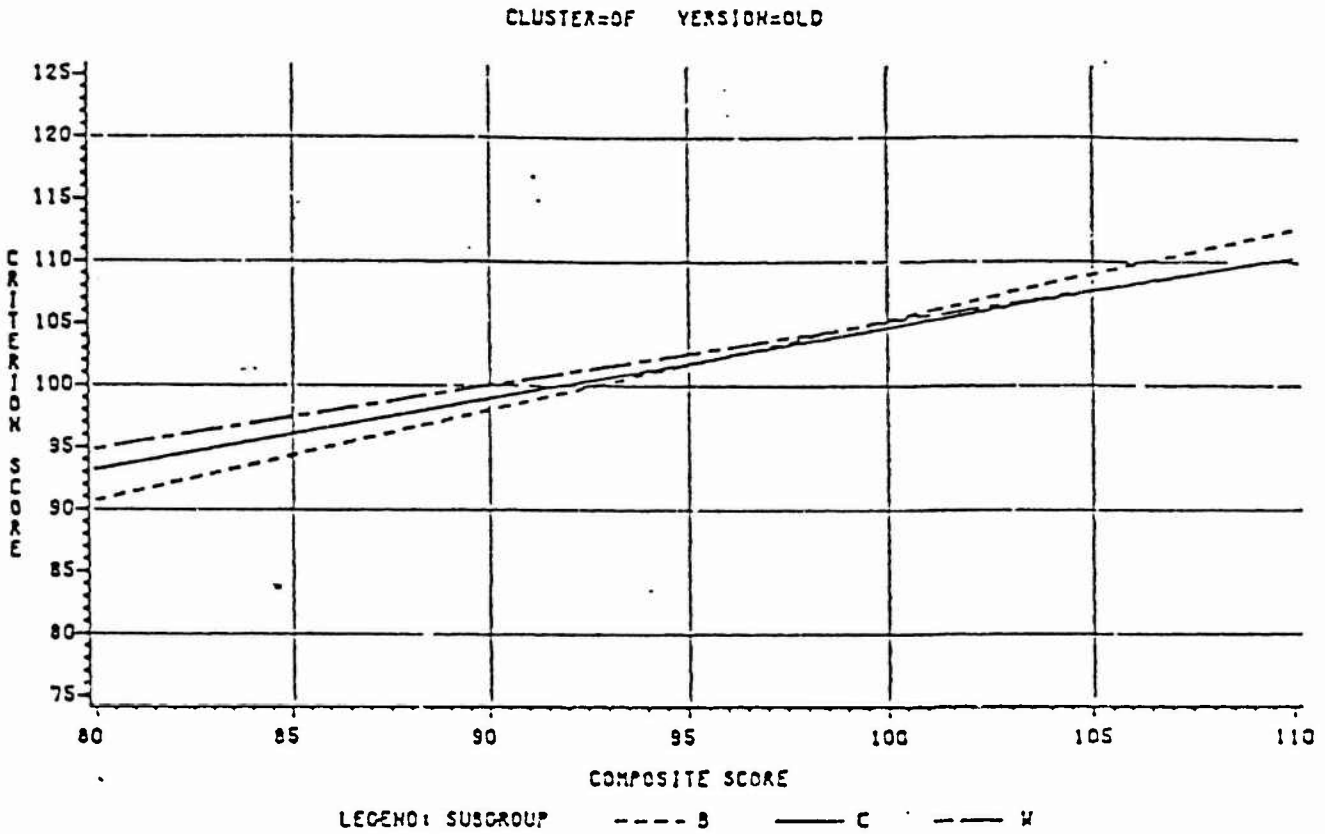


Figure 5. Regression lines for Current and Alternative (old and new) Composites for MM MOS, by Race



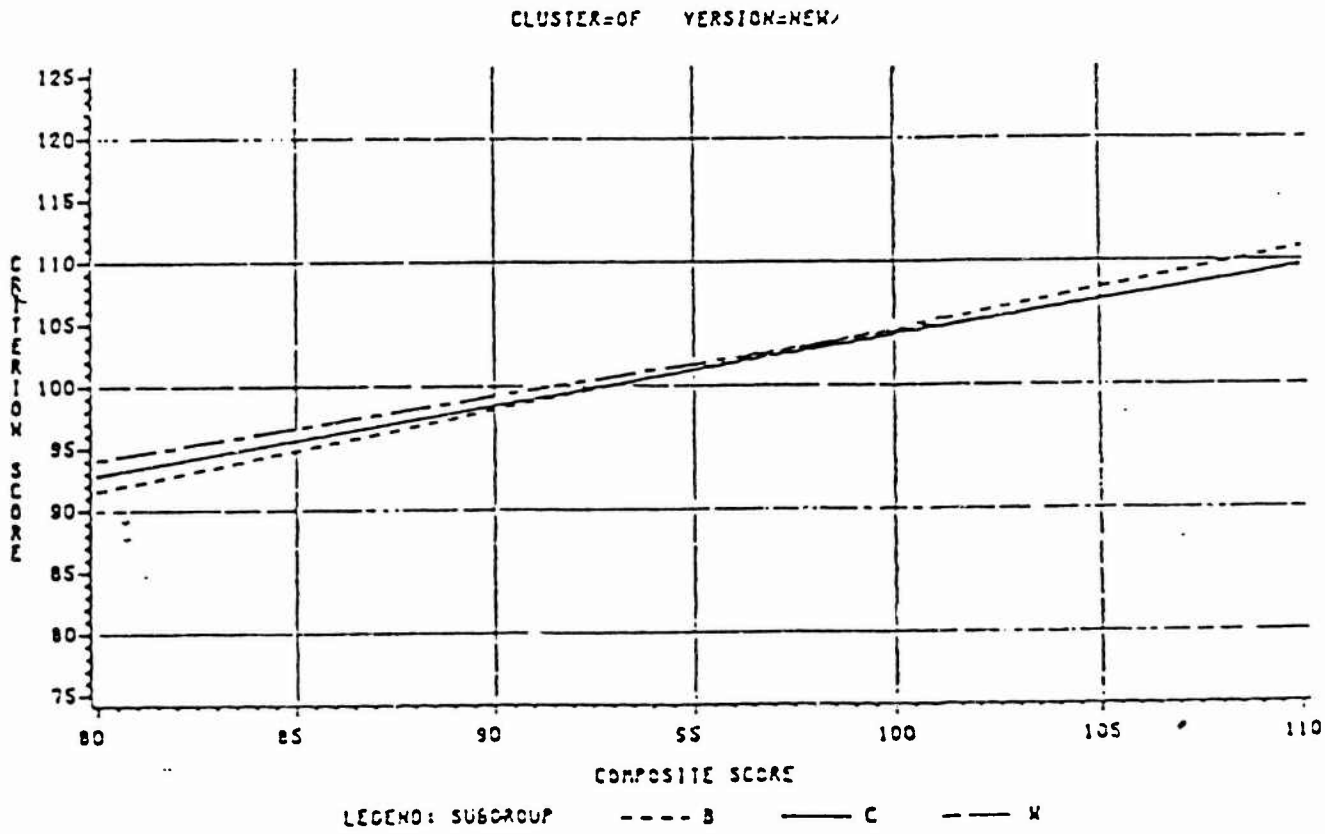
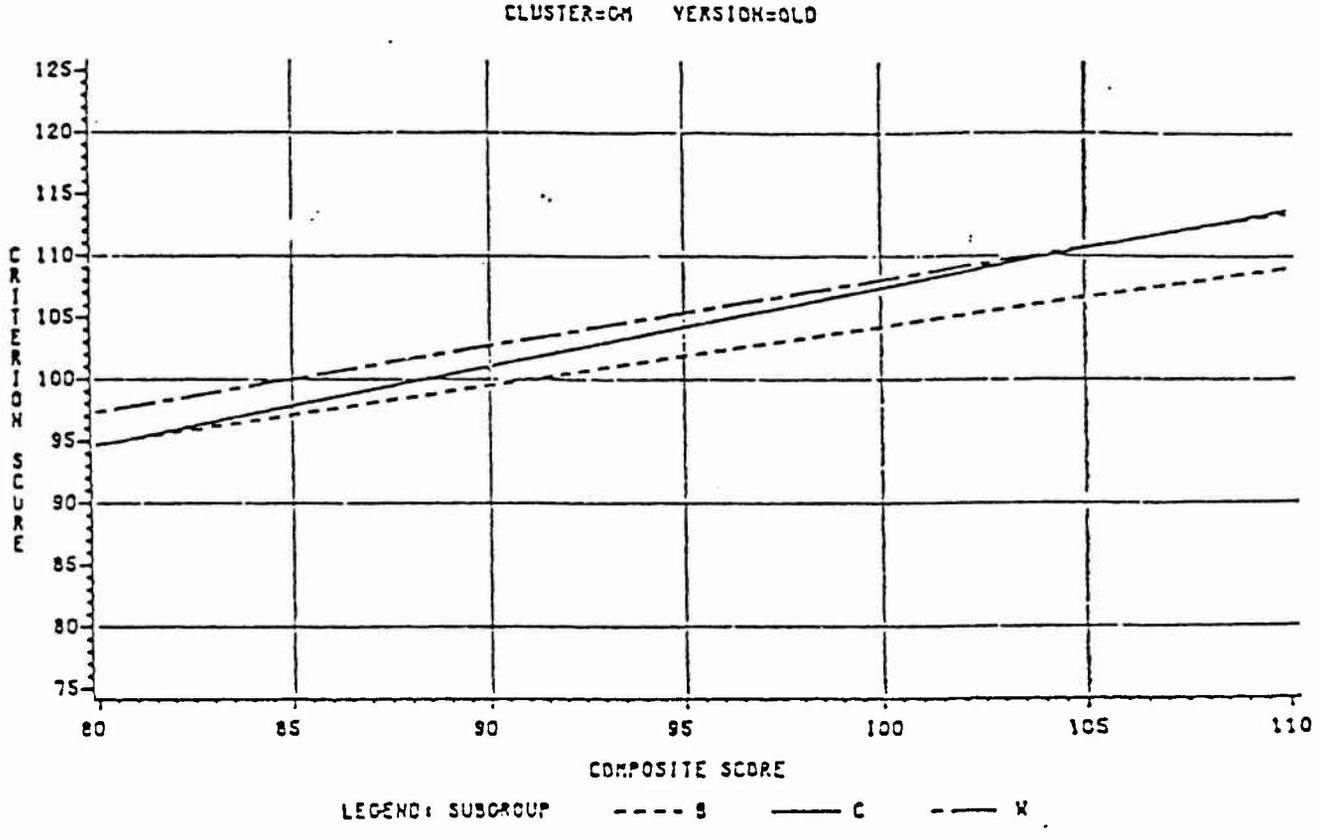


Figure 6. Regression lines for Current and Alternative (old and new) Composites for OF MOS, by Race



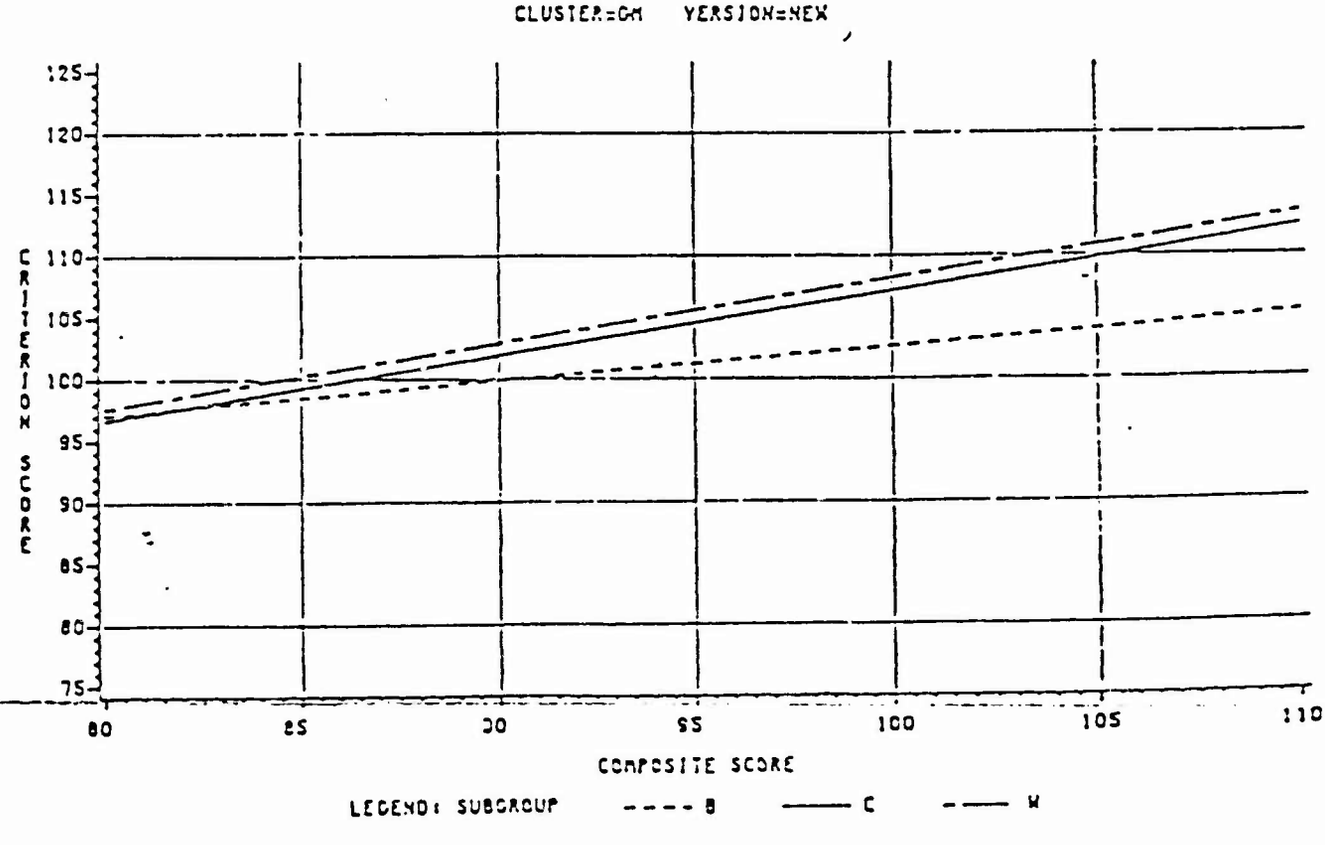
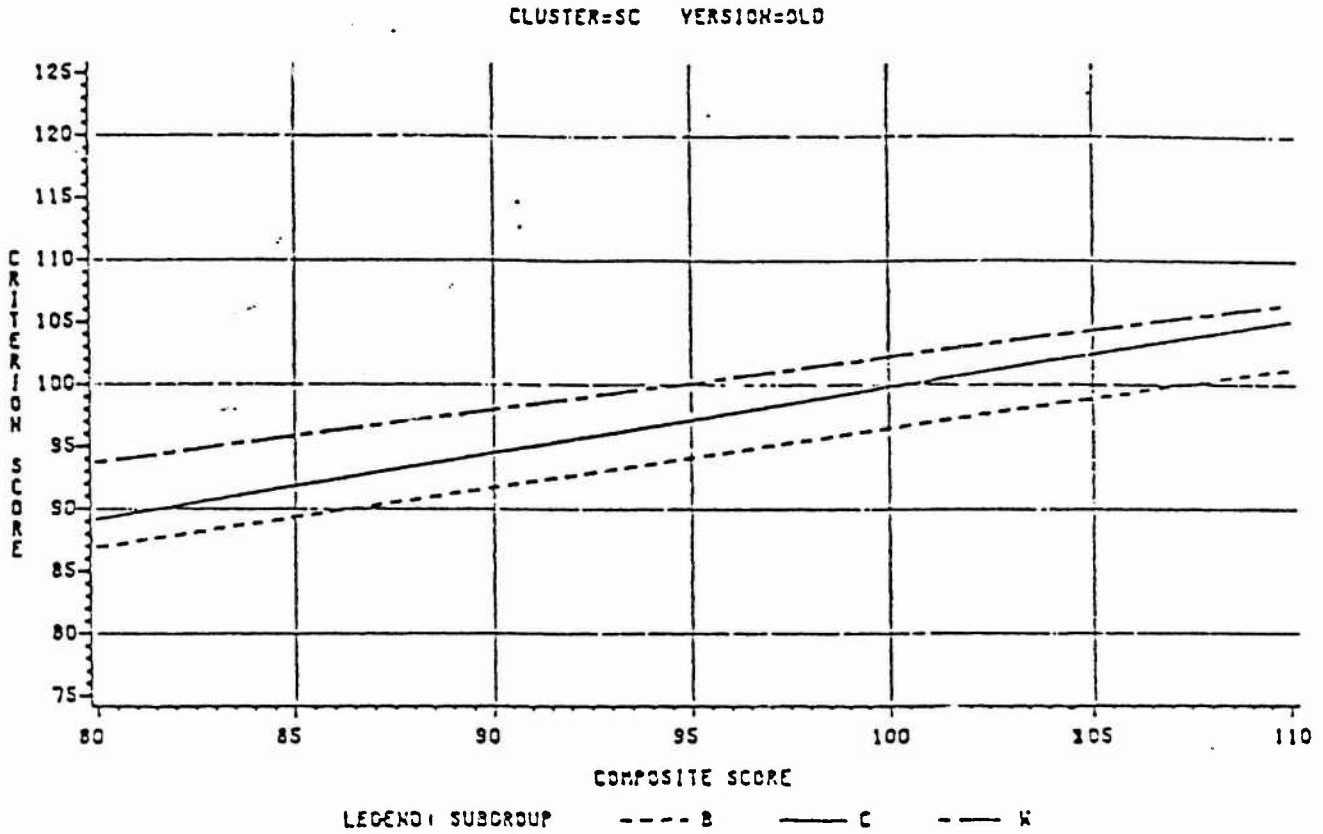


Figure 7. Regression lines for Current and Alternative (old and new) Composites for GM MOS, by Race



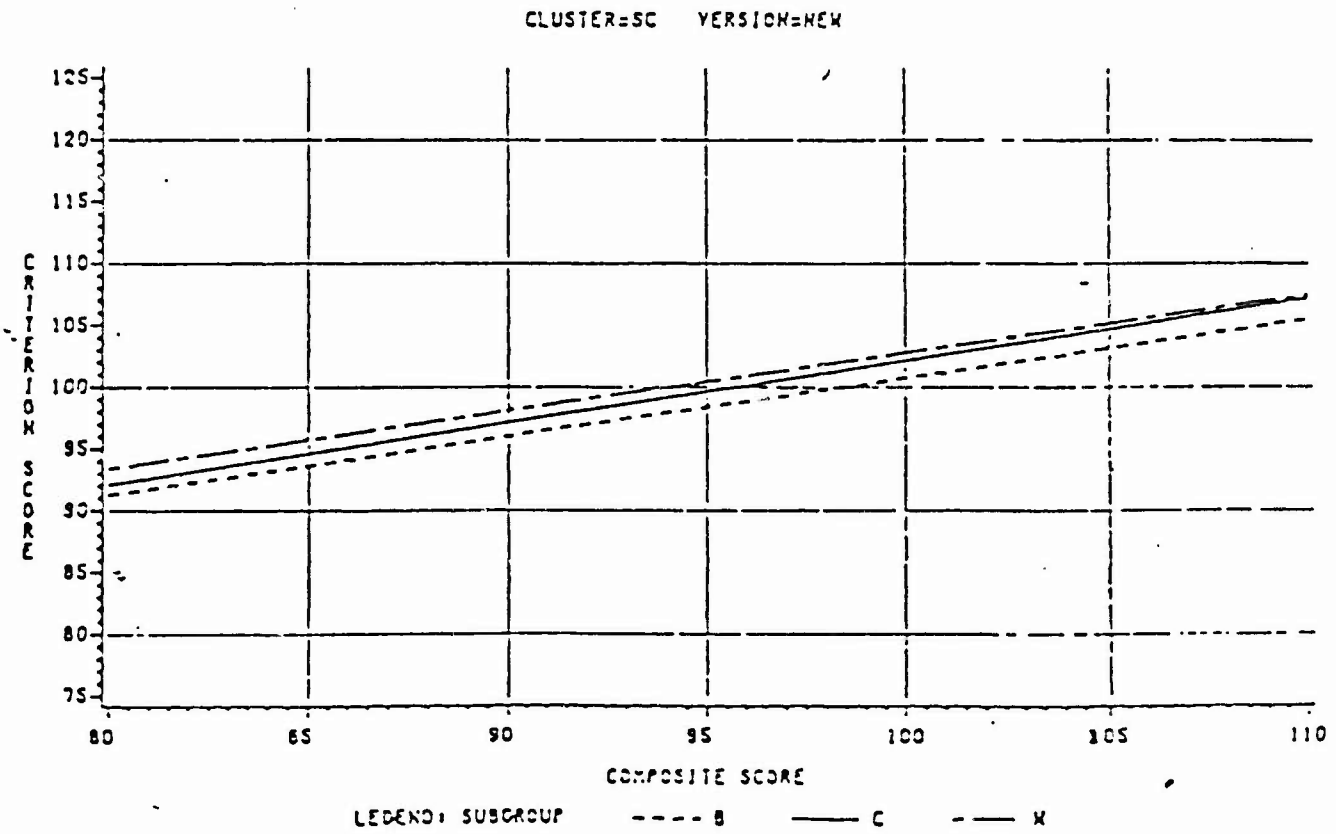
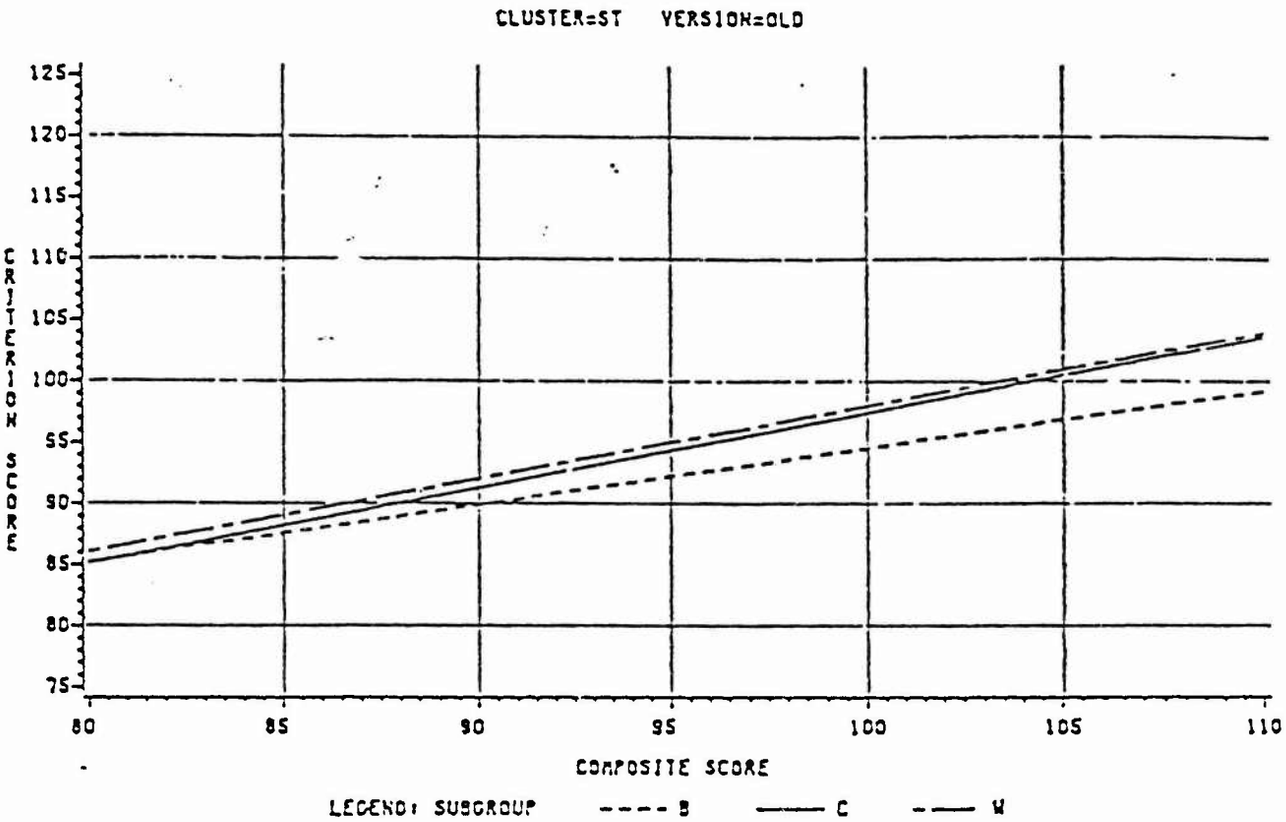


Figure 8. Regression lines for Current and Alternative (old and new) Composites for SC MOS, by Race



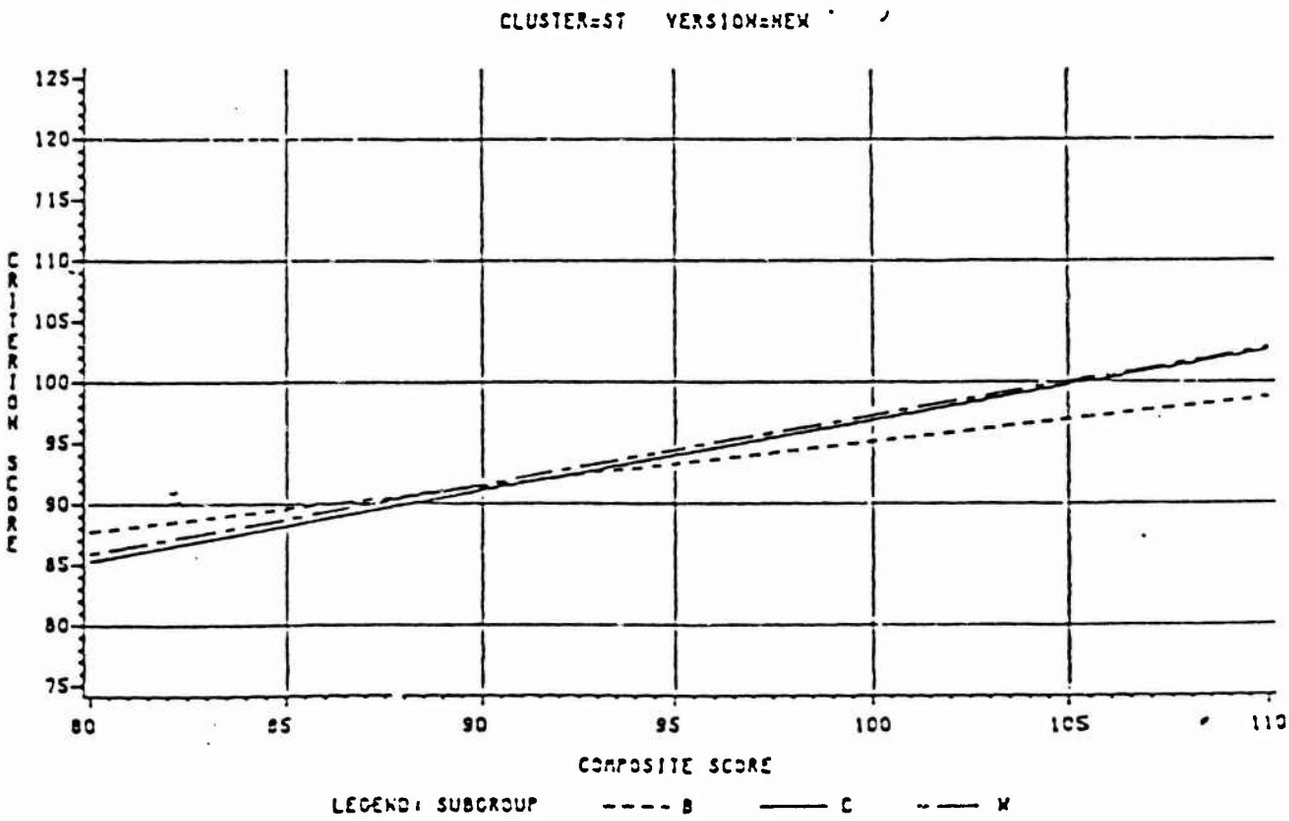


Figure 9. Regression lines for Current and Alternative (old and new) Composites for ST MOS, by Race

Figures 1 through 9 also show that the pattern of the relationships between the three regression lines near the possible cutoff values is quite similar for the two sets of composites for most of the nine clusters. Eight of the nine composites show overprediction of performance by blacks rather than the more serious underprediction of performance. The one exception was the OF cluster, which revealed overprediction for the lower composite scores and underprediction for the higher values. This pattern among the regression lines of the OF cluster was observed for both the operational and the proposed alternative composites.

To summarize the findings of the investigation of black versus white predictive bias, it appears that there are small differences in the predictive validities and the regression lines for the two groups for all composites in both their operational and alternative versions. The subgroup regression lines are also not perfectly approximated by a single common regression line, although for both sets this difference results in overprediction rather than underprediction of blacks in the region of the lines where selection and classification takes place. However, since the validity and regression line differences were not very large and the use of the common regression line does not, in general, result in underprediction of performance by blacks, either set of composites could be used without adversely impacting the enlistment of black soldiers.

Analyses of Differences by Gender

The sample and adjusted validities for gender subgroups of the current AA operational composites based upon the combined SQT and training criterion are presented in Table 5. Similar data but based upon the four proposed alternative composites are found in Table 6. The CO, FA, and GM clusters are not included in either of these tables because no MOS in these clusters met the criterion of at least 100 female soldiers (CO and FA do not contain MOS that are currently open to enlistment for women).

As was the case in the analysis of racial subgroups, both sets of composites tend to be accurate predictors of performance in each subgroup. Here the mean overall validities were .42 and .45 respectively for the current and alternative composites. The tables also show that the adjusted validities for the alternative composites tended to be higher than the values obtained by the current composites for both subgroups and across all clusters. The one exception to this rule was the validity of the ACO composite when used to predict the performance of men in the EL cluster. In this case the adjusted validities were equal for the current and alternative composite.

Another similarity between the data presented in Tables 5 and 6 and the validity differences discussed earlier for the black and white subgroups is that there was little change in the adjusted validity differences between the groups as a function of the two composite sets. The mean difference between male and female adjusted validities was .06 for the operational composites and .05 for the alternative composites. The change in validity differences between the two tables is only .01 and for two of the clusters (CL and SC) the difference in subgroup validities was consistent across composite sets. This finding further suggests that differences in the predictive

Table 5

Sample and Adjusted Validities for Males (M) and Females (F)
Current Operational Composites
SQT and Training Criteria Combined

Cluster/ Composite	Sample Size M	Sample Size F	Sample Validities M	Sample Validities F	Adjusted Validities M	Adjusted Validities F	Difference (Adjusted)
CL	9035	4352	.30	.19	.48	.45	.03
EL	3110	852	.25	.10	.41	.16	.25
MM	2238	195	.30	.33	.43	.51	-.08
OF	8142	1536	.31	.23	.47	.43	.03
SC	4113	1097	.29	.13	.47	.28	.19
ST	5912	1195	.27	.31	.46	.50	-.04

Table 6

Sample and Adjusted Validities for Males (M) and Females (F)
Four Alternative Composites
SQT and Training Criteria Combined

Cluster/ Composite	Sample Size M	Sample Size F	Sample Validities M	Sample Validities F	Adjusted Validities M	Adjusted Validities F	Difference (Adjusted)
CL / ACL	9035	4352	.42	.32	.56	.53	.03
EL / ACO	3110	852	.26	.14	.41	.19	.22
MM / AOP	2238	195	.31	.34	.43	.52	-.09
OF / AOP	8142	1536	.35	.27	.50	.46	.04
SC / AOP	4113	1097	.37	.25	.51	.32	.19
ST / AST	5912	1195	.27	.35	.46	.52	-.06

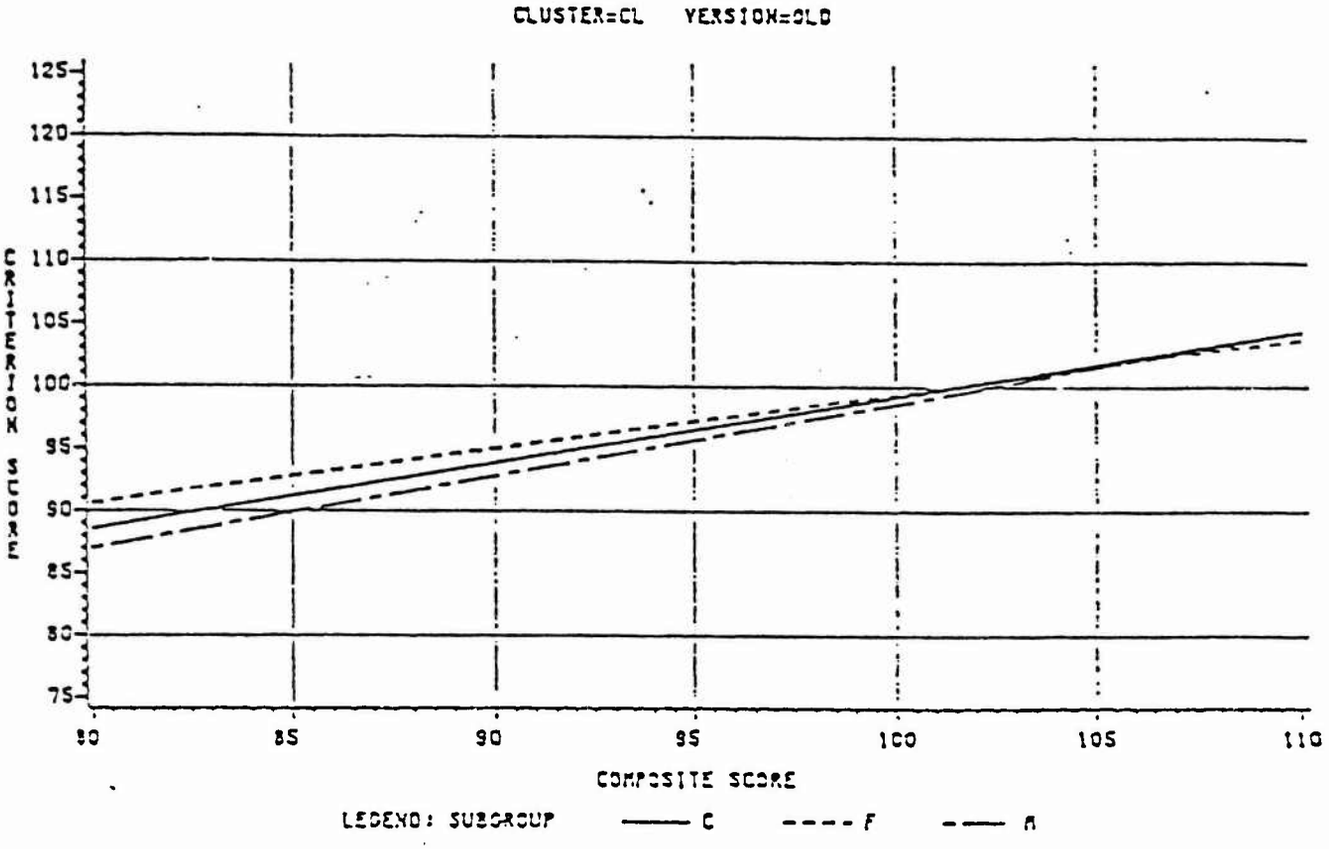
validity of ASVAB composites between cultural or racial subgroups is primarily a function of the the ASVAB subtests and not the manner those subtests are combined into composites.

Both Tables 5 and 6 show that in two clusters (EL and SC), there were fairly large differences in predictive validity between males and females. For the EL cluster this difference was .25 for the current composite and .22 for the alternative composite. The difference for the SC cluster was consistent at .19 for both composites. Whether these validity differences impact upon the selection and classification of women into these MOS clusters will be further discussed in the analysis of the differences between the regression lines and the discussion of the plots of the common and subgroup regression lines.

Table 7 presents the comparisons between the female and male regression lines for the current operational composites, while similar data are given for the four alternative composites in Table 8. The data in these tables were obtained in the manner that has been previously described in the analyses of racial subgroups.

Tables 7 and 8 show that, despite the higher predictive validity of the alternative AA composites for both subgroups, the subgroup regression lines based upon the operational composites tend to be closer together than the female/-/male regression lines based upon the alternative composites. The mean absolute value of the differences in predicted criterion scores between the two groups was 1.69 for the operational composites in comparison to 2.79 for the alternative composites. Four clusters (CL, MM, OF, and SC) showed sizeable increases in the absolute value of the differences, but of these the change for MM should not present an issue for the assignment of personnel to MOS. It represents an increase in overprediction of the female regression line by the male regression rather than underprediction. A more serious concern is the apparent underprediction of female performance by the proposed alternative AA composites in the CL, OF, and SC clusters. It should be noted, however, that an observed average of about two and a half units of underprediction for these clusters is fairly small in comparison to the combined criterion standard deviation of 20. The seriousness of these differences in regression lines also depends on where along the common regression line they are found, and this issue can be best addressed by examining the plots of the three regression lines for each cluster.

Figures 10 through 15 present the plots of the female, male, and common regression lines across for the range of composite scores that contain the cutoff values, for both the current operational and the proposed alternative composites. A comparison of the figures for these two sets of composites shows that for one cluster (ST) the pattern among the plotted regression lines is quite similar for the two sets of AA composites. For two other clusters (EL and MM) the alternative composites show more overprediction of female soldier performance than do the current operational composites. Since in both of these cases the female line is overpredicted by the common line, a switch to the alternative should not hinder the enlistment of women into the MOS that comprise the MM and ST clusters. The plots for the remaining three clusters (CL, OF, and SC) all showed an increase in underprediction of female performance with the alternative composites. For the CL and SC clusters, the current composites also showed underprediction of the female criterion scores, and the new composites produced a small increase in that underprediction, particularly for high composite scores. In the case of the



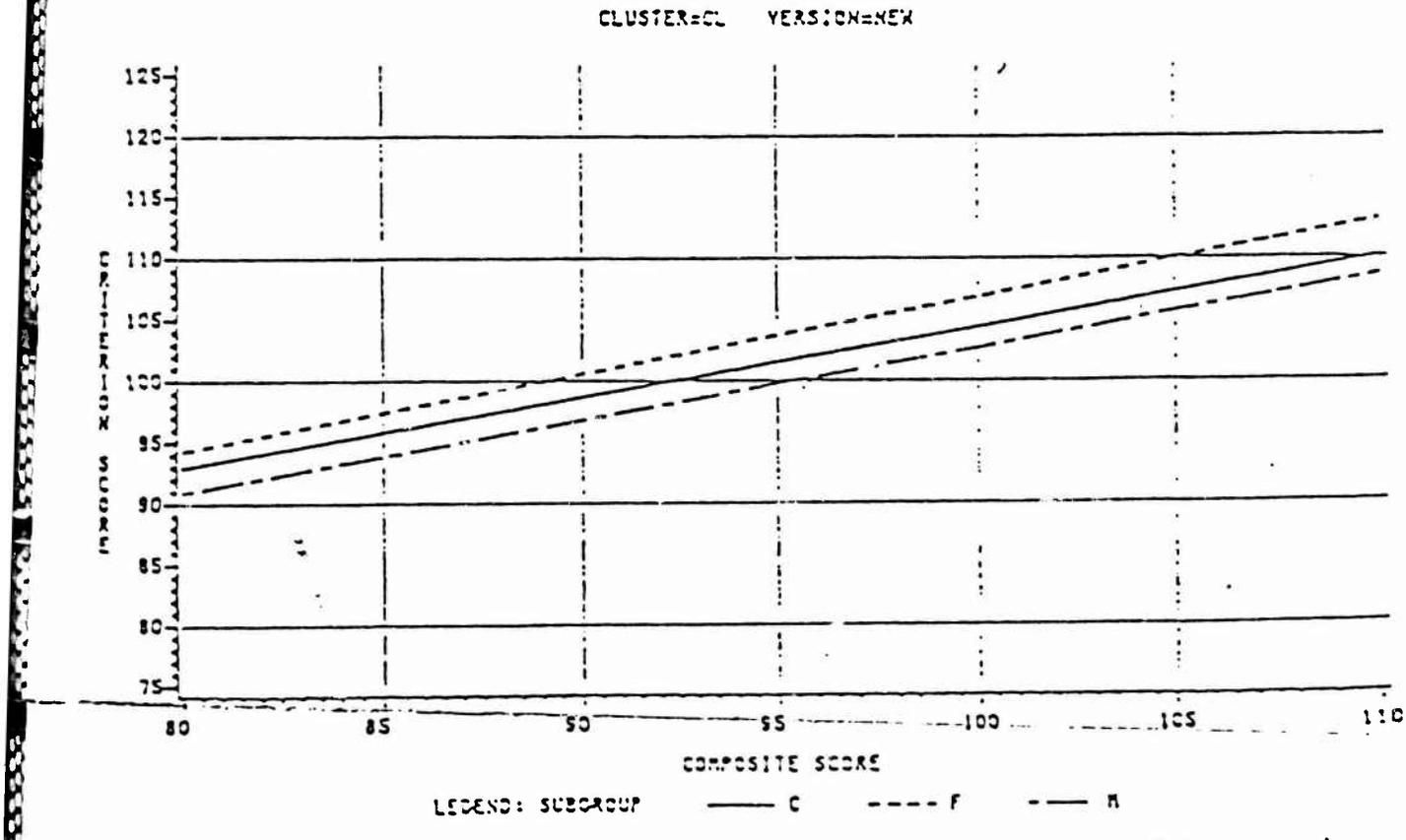
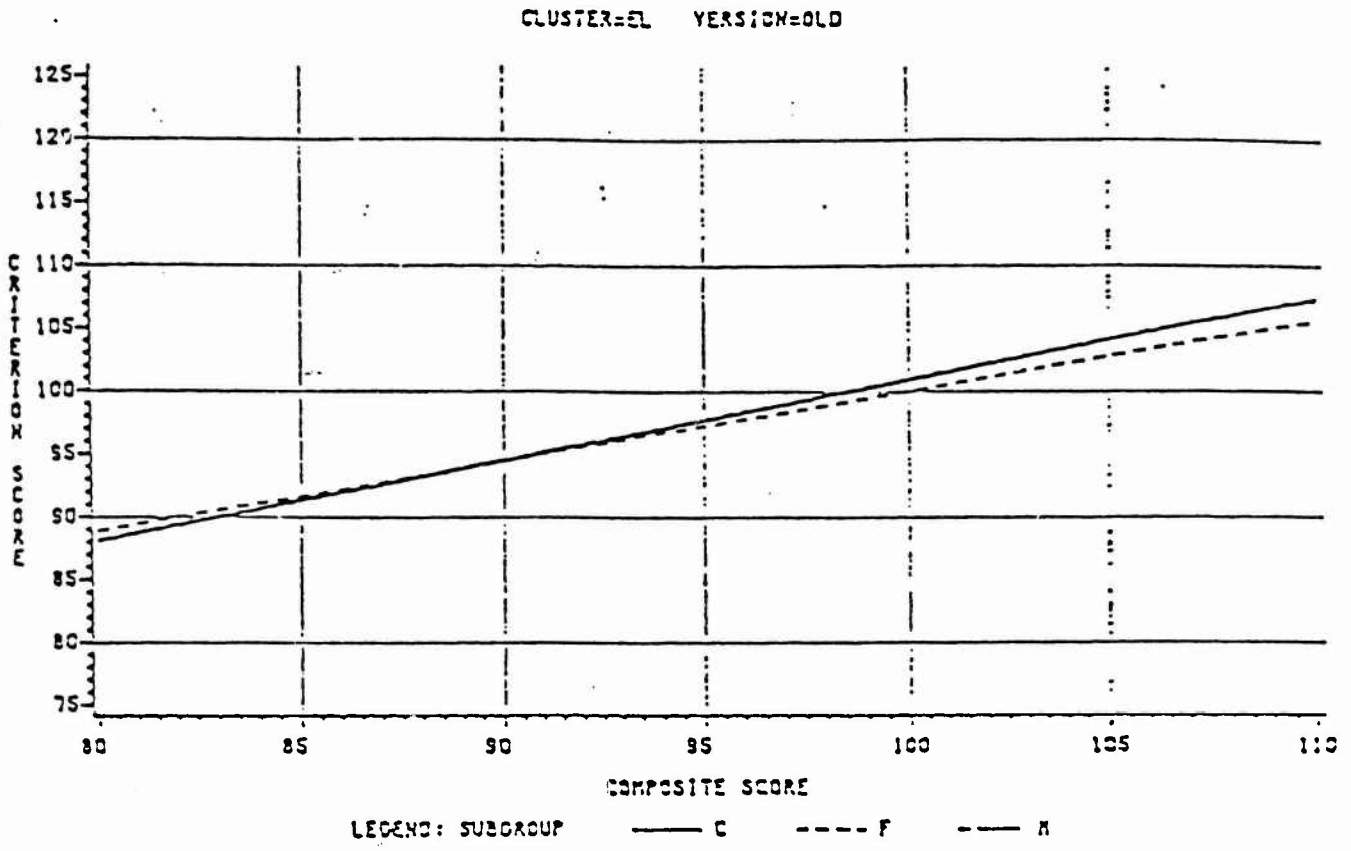


Figure 10. Regression lines for Current and Alternative (old and new) Composites for CL MOS, by Gender.



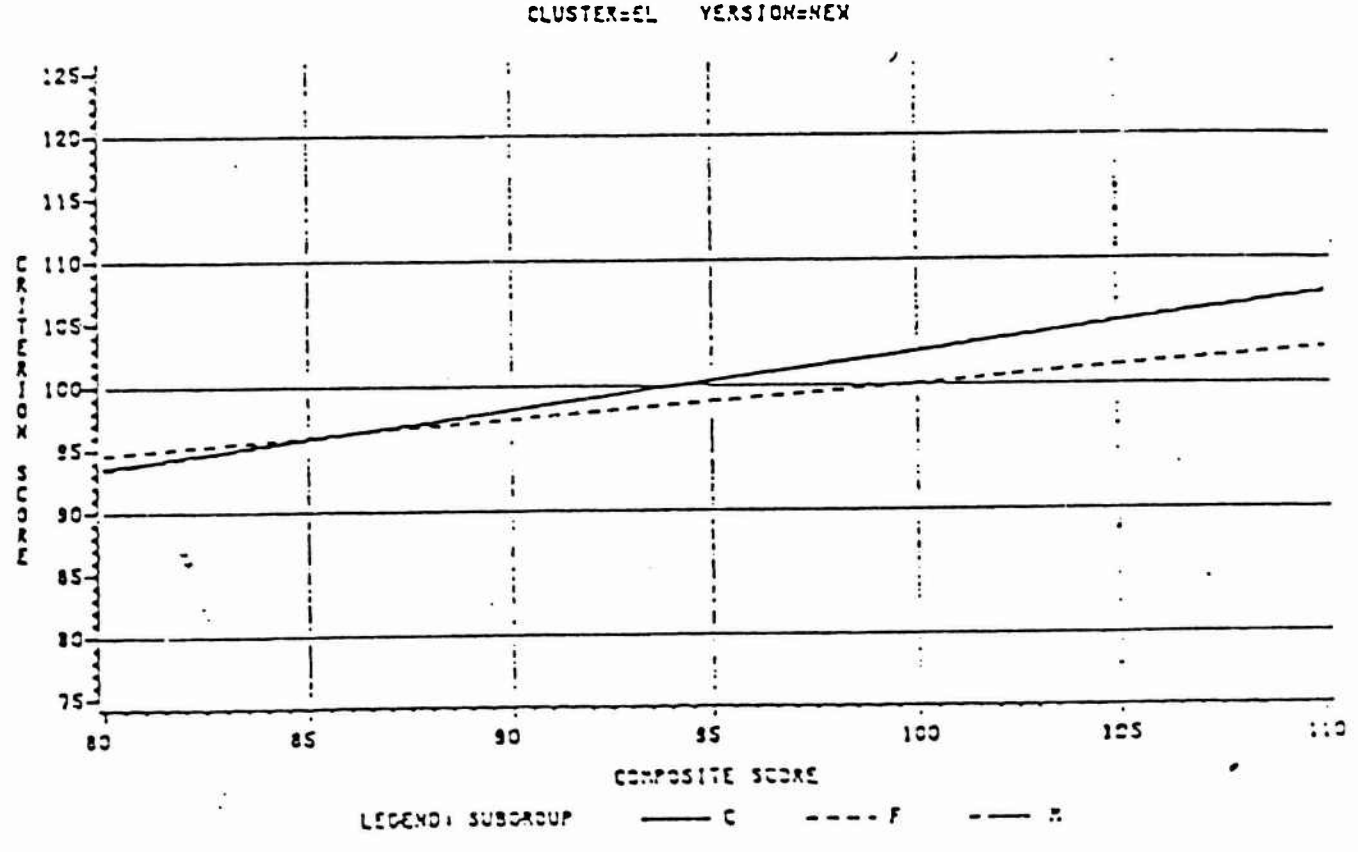
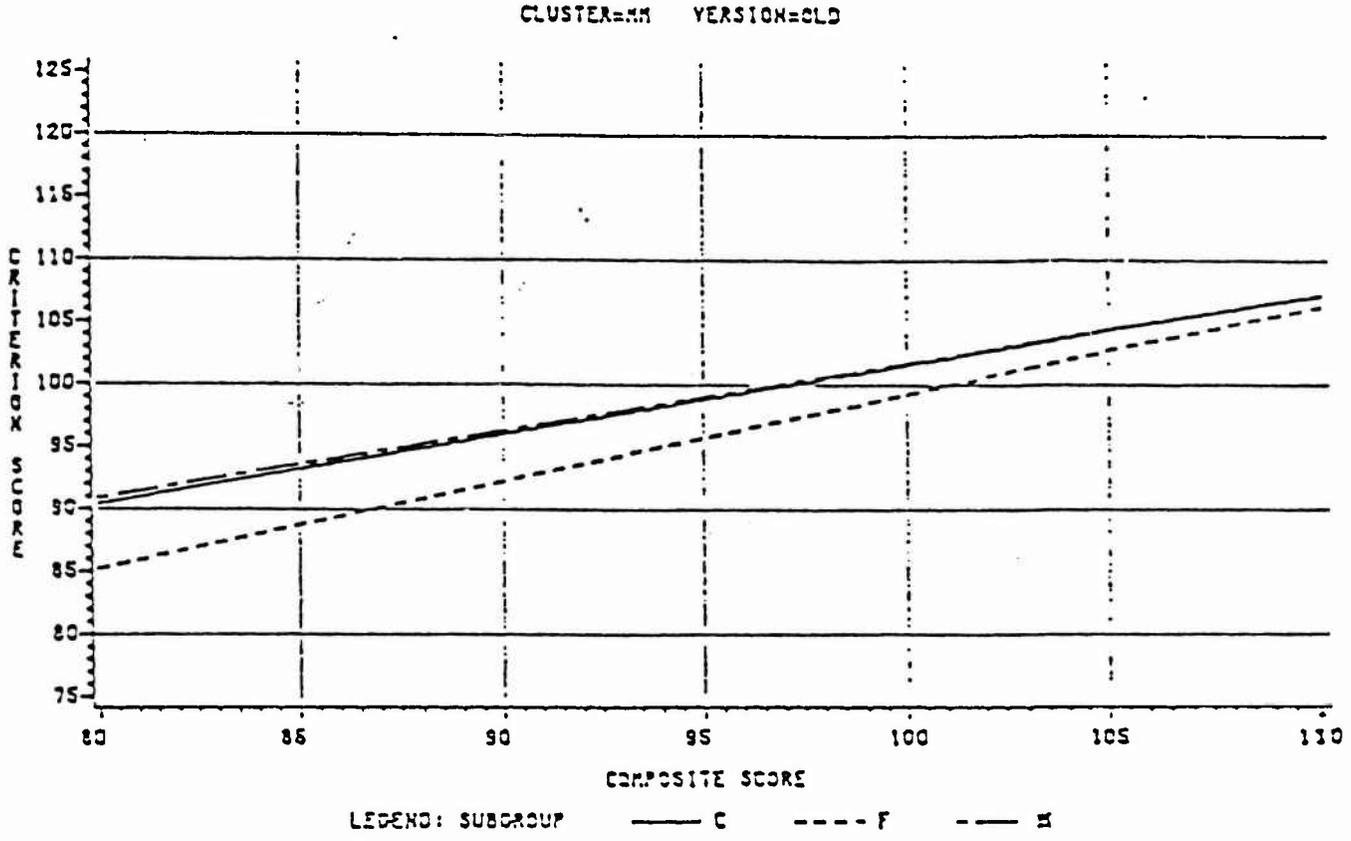


Figure 11. Regression lines for Current and Alternative (old and new) Composites for EL MOS, by Gender.



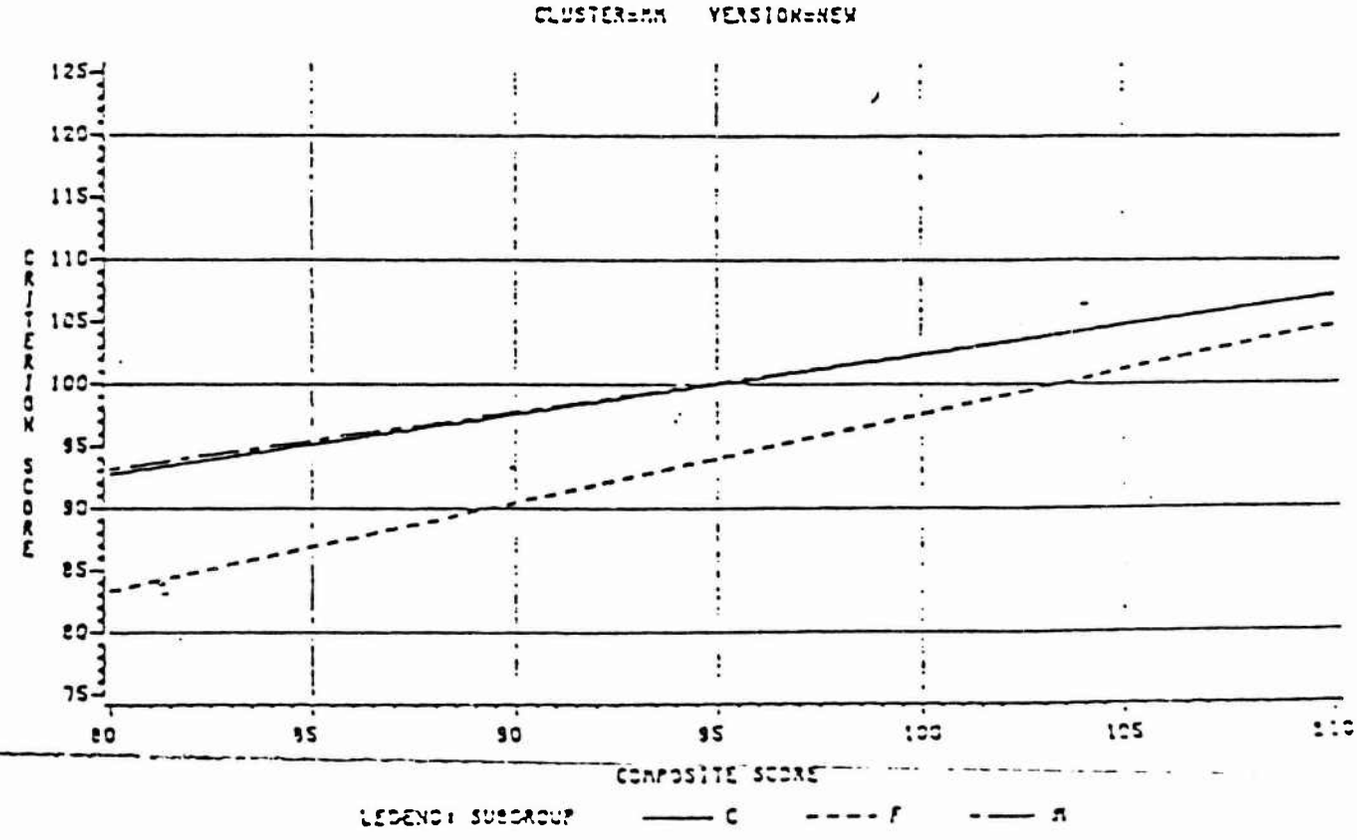
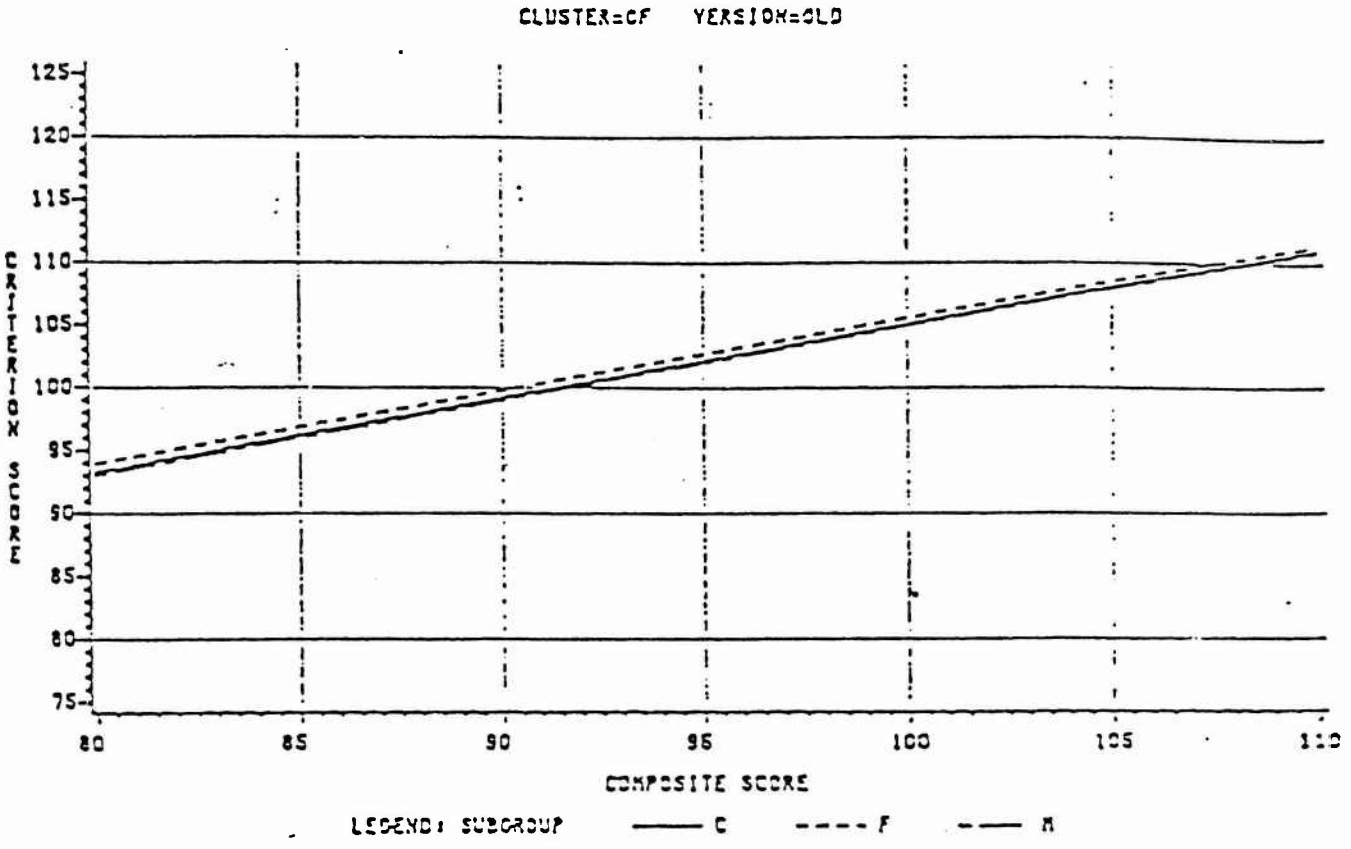


Figure 12 Regression lines for Current and Alternative (old and new) Composites for MM MOS, by Gender.



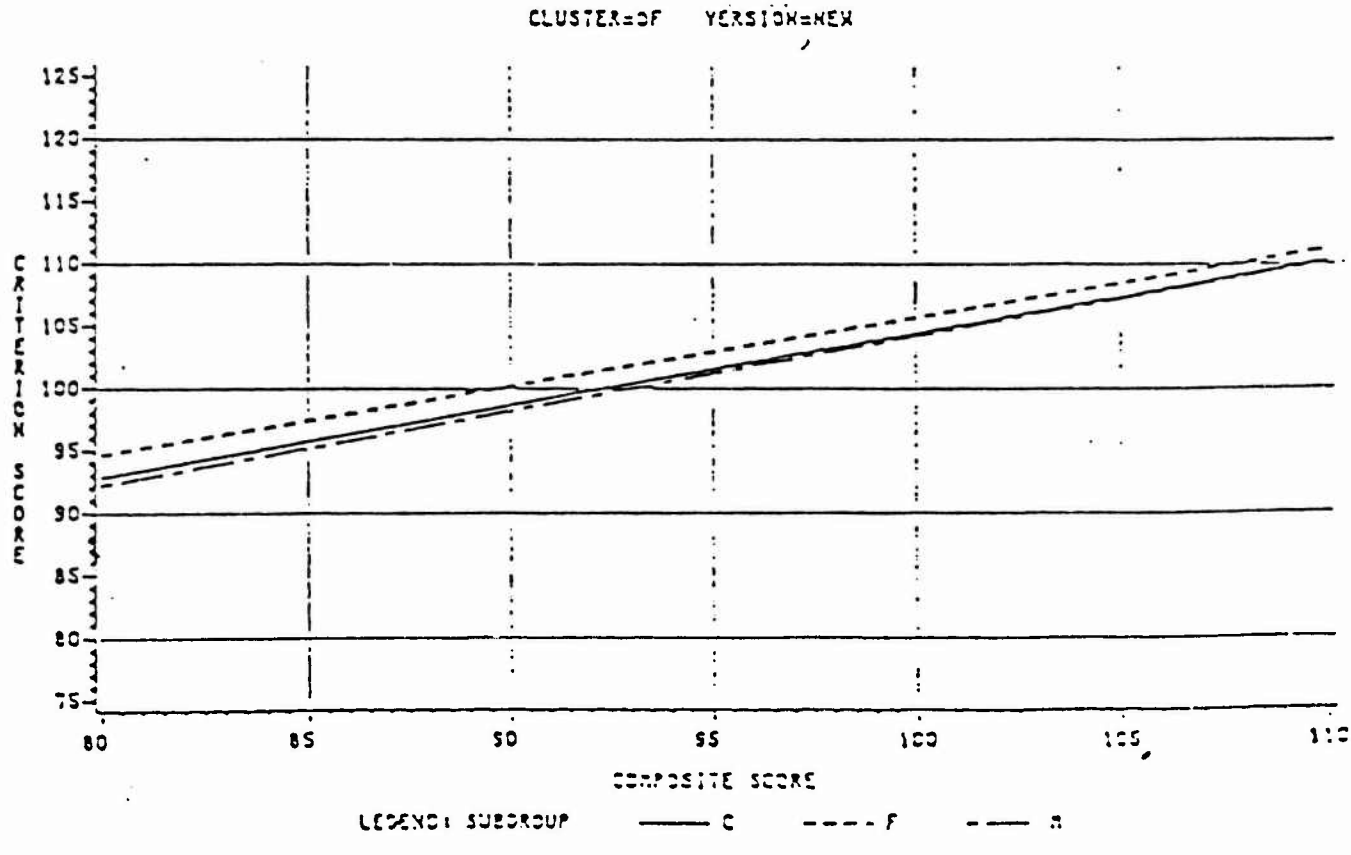
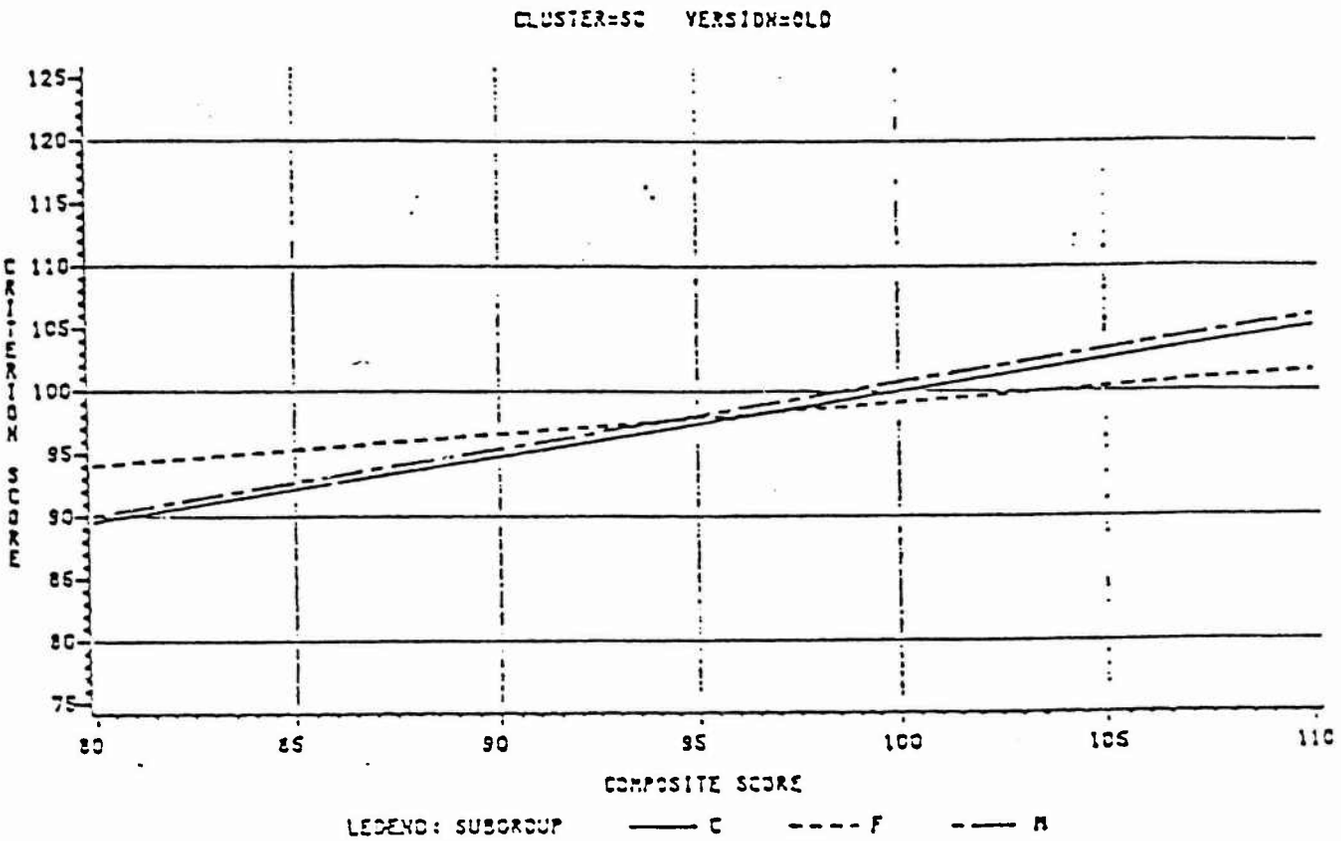


Figure 13 Regression lines for Current and Alternative (old and new) Composites for OF MOS. by Gender.



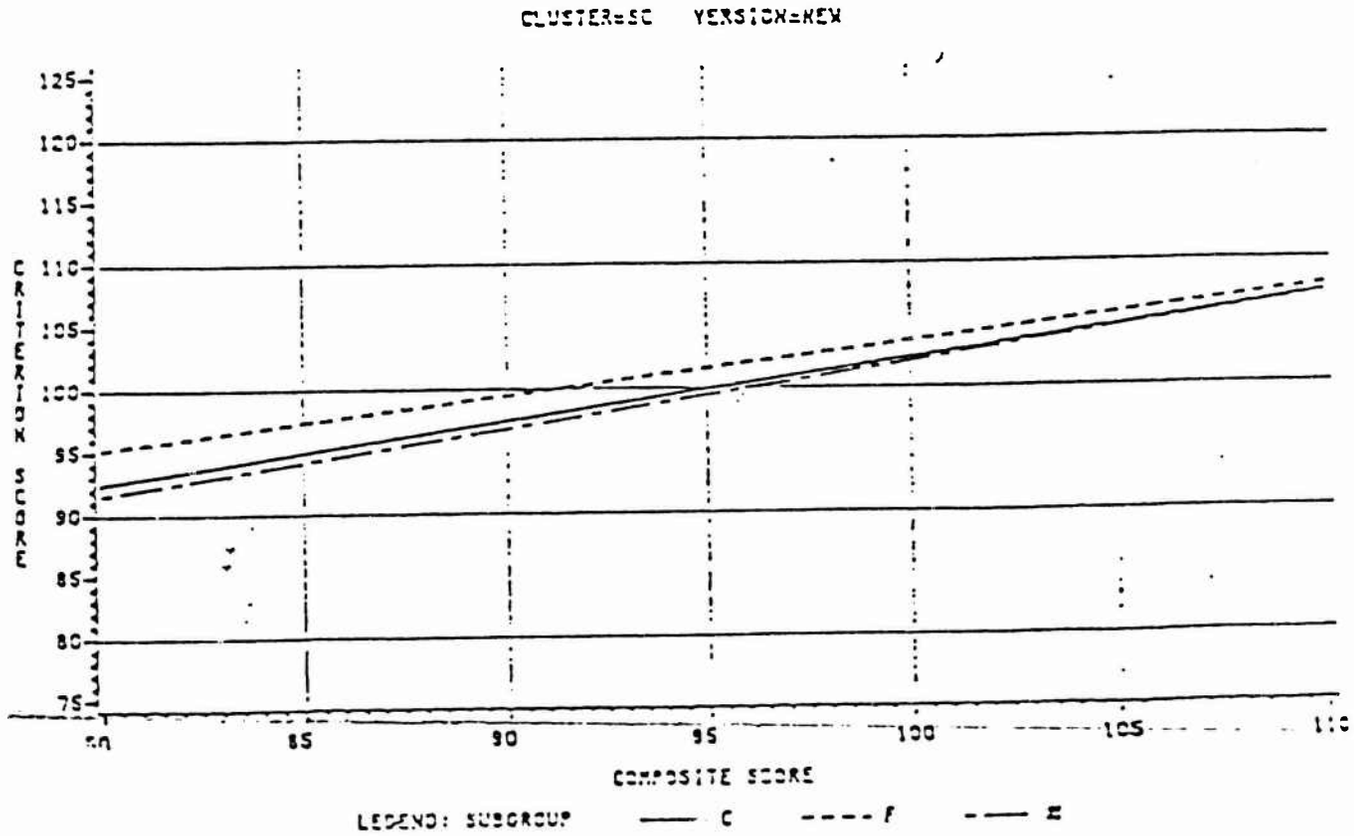
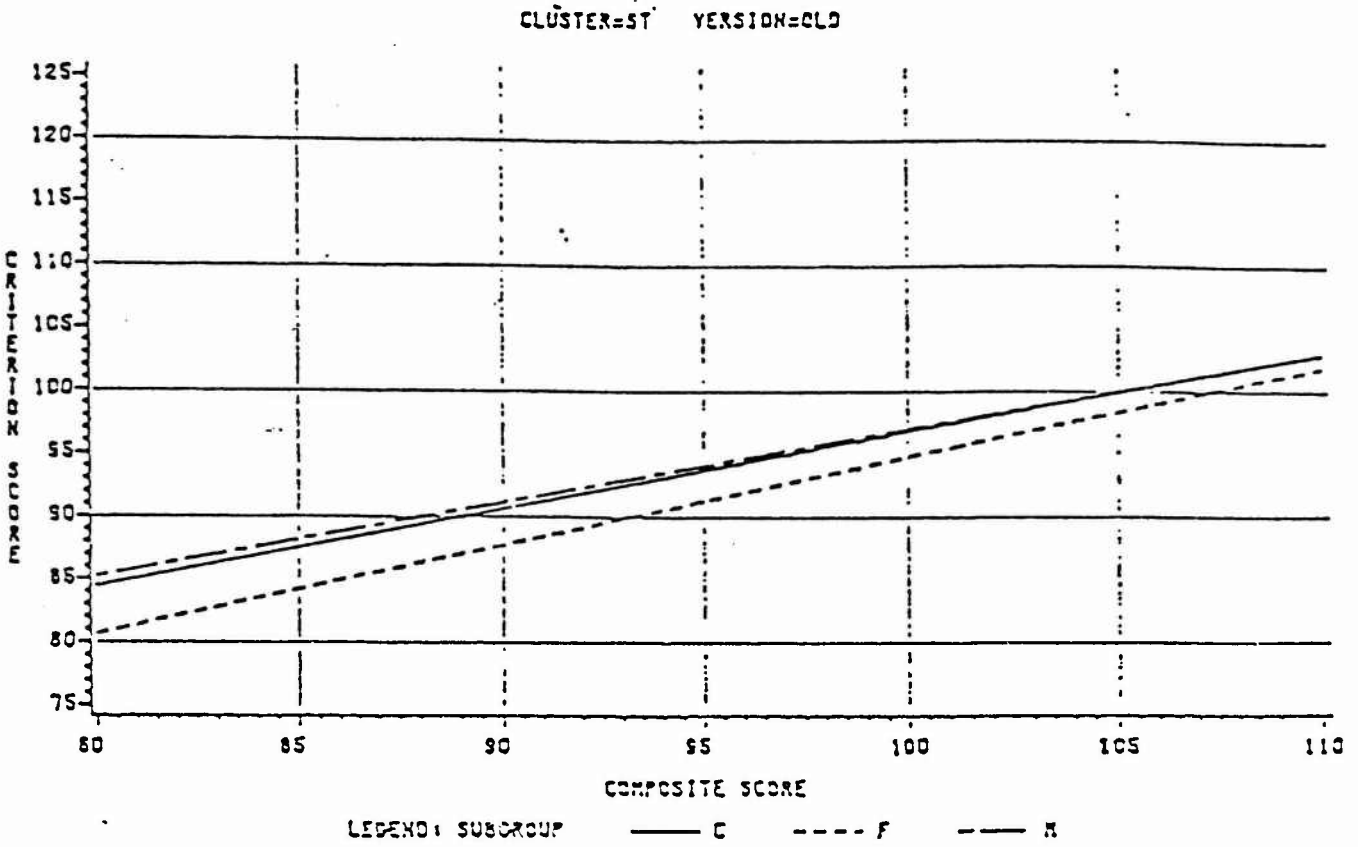


Figure 14 Regression lines for Current and Alternative (old and new) Composites for SC MOS, by Gender.



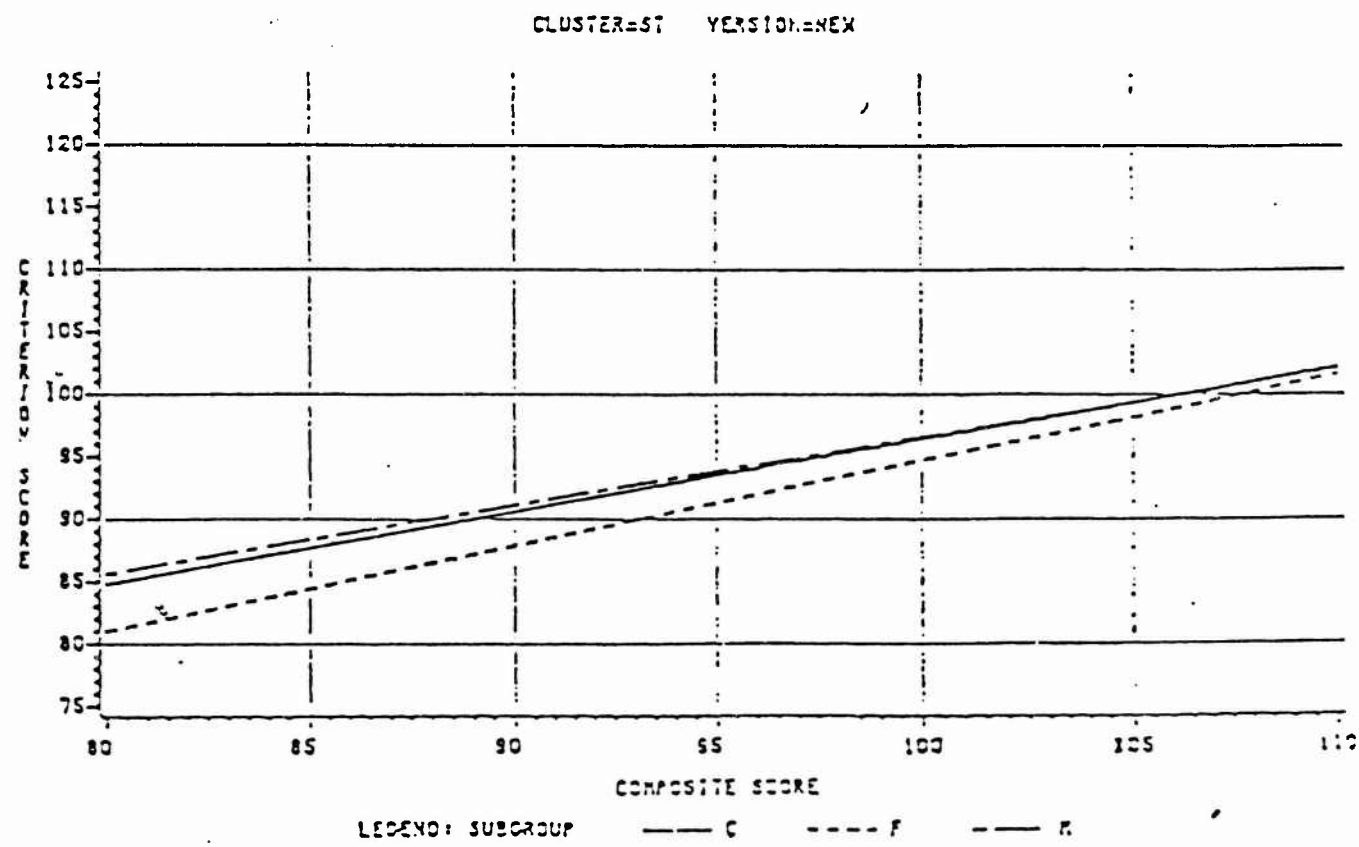


Figure 15 Regression lines for Current and Alternative (old and new) Composites for ST MOS, by Gender.

Table 7

Predicted Criterion Scores for Males (M) and Females (F): Current Composites

Composite Score	Predicted Criterion Score (Combined	F	M)	Subgroup Difference (M/-/F)	Standard Error of the Difference
CL Cluster					
80	88.39	91.26	86.77	-4.49	3.02
85	91.03	93.36	89.68	-3.68	2.50
90	93.66	95.46	92.59	-2.87	2.03
95	96.30	97.56	95.50	-2.06	1.65
100	98.93	99.66	98.41	-1.25	1.42
105	101.56	101.77	101.33	-.44	1.38
110	104.20	103.87	104.24	.37	1.58
EL Cluster					
80	84.23	86.20	83.98	-2.22	2.97
85	88.28	89.74	88.09	-1.65	2.39
90	92.32	93.28	92.20	-1.08	1.83
95	96.37	96.83	96.31	-.51	1.35
100	100.42	100.37	100.42	.05	1.03
105	104.46	103.91	104.54	.62	1.04
110	108.51	107.46	108.65	1.19	1.39
MM Cluster					
80	90.37	85.24	90.88	5.63	2.98
85	93.16	88.74	93.59	4.85*	2.38
90	95.95	92.25	96.31	4.06*	1.92
95	98.74	95.75	99.02	3.27	1.71
100	101.53	99.25	101.74	2.49	1.86
105	104.32	102.75	104.46	1.70	2.29
110	107.11	106.25	107.17	.92	2.87
OF Cluster					
80	93.26	93.93	93.00	-.93	1.65
85	96.22	96.84	95.99	-.85	1.27
90	99.18	99.75	98.97	-.78	1.01
95	102.13	102.66	101.95	-.70	.97
100	105.09	105.57	104.94	-.63	1.17
105	108.05	108.47	107.92	-.55	1.52
110	111.00	111.38	110.90	-.48	1.95

(cont'd)

Predicted Criterion Scores for Males (M) and Females (F):
Current Composites (Continued)

Composite Score	Predicted Criterion Score (Combined	F	M)	Subgroup Difference (M/-/F)	Standard Error of the Difference
SC Cluster					
80	89.64	94.26	90.44	-3.82	4.05
85	92.25	95.45	93.08	-2.37	3.29
90	94.87	96.63	95.72	-.92	2.57
95	97.48	97.82	98.36	.54	1.94
100	100.09	99.01	101.00	1.99	1.50
105	102.71	100.20	103.64	3.44*	1.46
110	105.32	101.38	106.28	4.90*	1.84
ST Cluster					
80	85.21	81.52	86.03	4.51	3.19
85	88.22	84.89	88.92	4.03	2.67
90	91.22	88.26	91.82	3.55	2.18
95	94.23	91.64	94.71	3.07	1.72
100	97.23	95.01	97.60	2.59	1.33
105	100.24	98.39	100.50	2.11	1.10
110	103.24	101.76	103.39	1.63	1.13

* p .05

OF cluster, the three regression lines of the current composite are essentially equal, while a switch to the alternative composite would result in some underprediction along the entire regression line. In general, the degree of underprediction of female scores shown in these three clusters is relatively small. The CL cluster is perhaps the most extreme case and here the common regression line falls only about two points below the female line.

Considering all of the data discussed above, it appears that the alternative AA composites could replace the composites now being used operationally without increasing predictive bias on the basis of gender. The differences in predictive validity of the two sets of composites are quite similar, and the degree of underprediction of female performance by a common regression line is much the same for both composite sets.

Other Analyses of Subgroup Differences

One possible explanation of the lower predictive validities for blacks in Tables 1 and 2 and for females in some clusters of Tables 5 and 6 is that these subgroups showed less variability in their criterion scores than the other two subgroups. The data relevant to this hypothesis can be found

Table 8

Predicted Criterion Scores for Females (F) and Males (M): Four Alternative Composites

Composite Score	Predicted (Combined	Criterion Score F	M)	Subgroup Difference (M/-/F)	Standard Error of the Difference
CL Cluster					
80	93.10	94.36	90.98	-3.38	2.86
85	96.08	97.63	94.06	-3.57	2.37
90	99.06	100.91	97.14	-3.77*	1.93
95	102.03	104.18	100.21	-3.97*	1.56
100	105.01	107.45	103.29	-4.16*	1.34
105	107.99	110.73	106.36	-4.36*	1.31
110	110.97	114.00	109.44	-4.56*	1.49
EL Cluster					
80	91.37	92.47	91.02	-1.45	2.95
85	94.49	95.38	94.20	-1.18	2.37
90	97.61	98.30	97.38	-.92	1.82
95	100.74	101.21	100.55	-.66	1.34
100	103.86	104.12	103.73	-.40	1.02
105	106.98	107.04	106.90	-.14	1.04
110	110.11	109.95	110.08	.13	1.38
MM Cluster					
80	92.70	84.42	93.28	8.86	2.97
85	95.13	87.82	95.66	7.83	2.36
90	97.57	91.23	98.03	6.80	1.90
95	100.00	94.63	100.40	5.77	1.70
100	102.43	98.04	102.77	4.74	1.85
105	104.86	101.44	105.14	3.70	2.27
110	107.29	104.84	107.52	2.67	2.86
OF Cluster					
80	92.75	94.06	92.21	-1.85	1.62
85	95.81	97.24	95.32	-1.91	1.25
90	98.88	100.42	98.44	-1.98*	.99
95	101.94	103.59	101.55	-2.04*	.95
100	105.01	106.77	104.67	-2.10	1.15
105	108.07	109.95	107.78	-2.17	1.49
110	111.14	113.13	110.90	-2.23	1.91

(cont'd)

Predicted Criterion Scores for Females (F) and Males (M):
Four Alternative Composites Solution (Continued)

Composite Score	Predicted Criterion Score (Combined	F	M)	Subgroup Difference (M/-/F)	Standard Error of the Difference
SC Cluster					
80	92.59	95.16	92.26	-2.90	3.88
85	95.17	97.41	94.89	-2.53	3.16
90	97.74	99.66	97.51	-2.15	2.47
95	100.31	101.91	100.14	-1.78	1.86
100	102.89	104.16	102.76	-1.40	1.44
105	105.46	106.41	105.39	-1.02	1.40
110	108.03	108.66	108.02	-.65	1.76
ST Cluster					
80	85.38	81.69	86.14	4.46	3.18
85	88.46	85.27	89.10	3.83	2.67
90	91.54	88.86	92.07	3.21	2.17
95	94.63	92.44	95.03	2.59	1.71
100	97.71	96.03	97.99	1.96	1.33
105	100.79	99.61	100.95	1.34	1.10
110	103.87	103.20	103.91	.71	1.13

p .05

in Table 4 for the comparison of racial subgroups and Table 5 for comparisons based upon gender. It should be noted that all of the standard deviations in these tables are similar, because the criterion measures had been standardized to have a standard deviation of twenty in each MOS.

Examination of Table 9 shows that the small differences observed between black and white composite validities are not due to any major restriction in the variability of the criterion for black soldiers, relative to white soldiers. For seven of the nine clusters the black subgroup showed greater criterion variability than did the white subgroup. The differences in predictive validity between these groups, therefore, cannot be attributed to differences in criterion variability.

The data in Table 9 do suggest an explanation for the observed overprediction of black soldier performance by the use of a common regression line. For all nine clusters in this table, the mean criterion score for blacks is slightly smaller than the value for whites. Such a relationship normally leads to common line overprediction of the subgroup with the lower mean criterion score.

Table 9

Means and Standard Deviations Of the Combined Criterion Scores for Black (B) and White (W) Subgroups

Cluster	Means		Standard Deviations	
	B	W	B	W
CL	97.63	105.68	19.18	19.63
CO	94.46	103.71	20.05	19.08
EL	99.60	105.75	19.72	18.74
FA	97.23	107.55	19.24	18.25
GM	99.74	104.72	19.28	20.77
MM	95.13	103.99	20.63	18.97
OF	97.51	104.34	20.43	19.00
SC	97.00	105.78	19.93	18.62
ST	96.34	104.27	19.85	18.68

Table 10 shows that lower criterion variances cannot explain the differences in validities between females and males in Tables 5 and 6. For the clusters that had shown somewhat lower validities for females than males (CL, EL, OF, and ST), only in the CL cluster did the criterion scores from female soldiers have less observed variance than the male criterion scores.

Table 10

Means and Standard Deviations Of the Combined Criterion Scores for Female (F) and Male (M) Subgroups

Cluster	Means		Standard Deviations	
	F	M	F	M
CL	103.63	101.00	18.12	19.89
EL	102.09	105.05	19.16	18.92
MM	96.09	101.36	20.76	19.92
OF	99.34	101.78	20.23	19.77
SC	98.40	104.23	20.10	19.03
ST	99.86	103.76	20.37	18.74

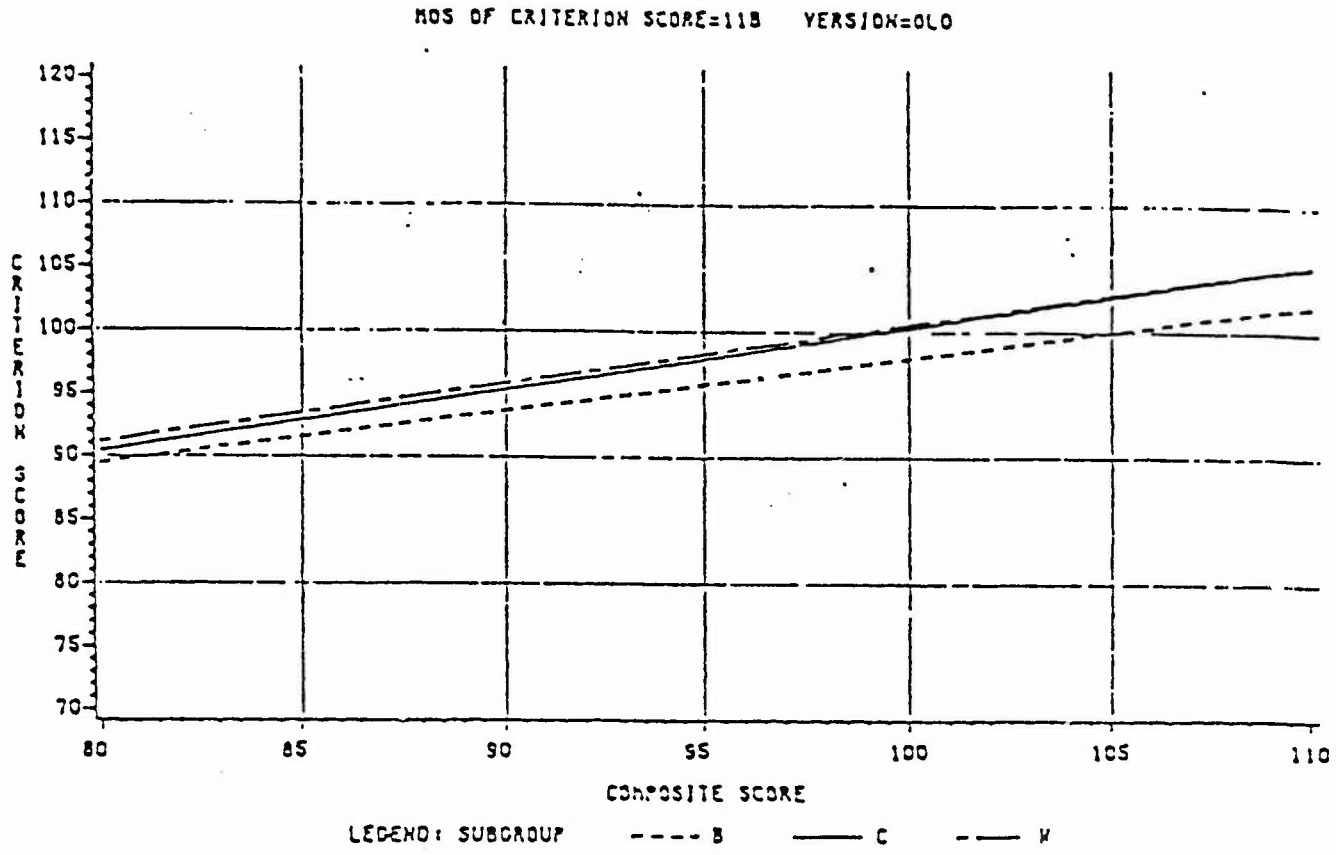
For all of the comparisons among subgroup predictive validities and regression lines discussed above, the reporting of analyses has been at the cluster rather than the MOS level. In order to aggregate the information to this level, the statistics were first calculated for each MOS. The resulting data were then pooled (weighted by sample size) across the appropriate MOS to obtain the analyses for each cluster or composite. While this approach is the most reasonable way to aggregate MOS-level data to the cluster level, it does not inform about MOS-level relationships. This question is particularly relevant for MOS with different proportions of subgroup populations.

We addressed this question by comparing regression lines for sets of two MOS within each cluster. The particular MOS for these analyses were selected according to the following criteria: First, there had to be at least two MOS within a cluster for which we had data for at least 100 soldiers in each subgroup. Second the two MOS within each cluster were selected by taking the two that showed the greatest difference in the ratio of subgroup sample sizes. For example, in the the analyses of racial differences within the CL cluster, the two MOS examined were 71L and 75D. In the case of 71L the ratio of whites to blacks was 1.07, while in 75D the same ratio was .43. This procedure was followed in order to maximize the probability of uncovering differences in the regression lines as a function of the distribution of subgroups within the MOS. The procedure had the side effect of allowing for the reporting of analyses of MOS with relatively small sample sizes in comparison to the other analyses of this report, but the minimum sample of at least 100 soldiers per subgroup was still large in comparison to past research.

For the MOS meeting these criteria the differences between subgroup regression lines for both racial and gender comparisons are given in Tables 11 and 12 for the current composites and in Tables 13 and 14 for the alternative composites. The comparison of the subgroup regressions to the common regression line are presented in Figures 16 through 29.

Three important findings emerge from these tables and figures. First, these data indicate that in general a switch to the alternative composites would not result in an increase in predictive bias for either blacks or women. Most (nine out of fourteen) of the MOS show quite similar patterns among the subgroup regression lines drawn from the current and alternative composites. For the comparisons based upon race, only MOS 11H and 13F showed substantial change with the new composites. For MOS 11H, the switch to the new composites would tend to result in overprediction of black soldier performance while the current system produces some underprediction. For MOS 13F, the new composites produce a subgroup regression line that is closer and no longer nearly parallel to the common regression line. Neither of these changes would negatively impact the enlistment of blacks into these MOS.

Likewise, a change to the alternative composites does not appear to present serious problems for the enlistment of female soldiers even when the pattern among the regression lines appears to change with the composites. In the case of MOS 05C this change results only in the regression lines being closer together, and therefore showing less underprediction of female performance by the common regression line. For MOS 75C the relative degree of underprediction versus overprediction is fairly constant for the two sets of composites, but where each occurs along the common regression line



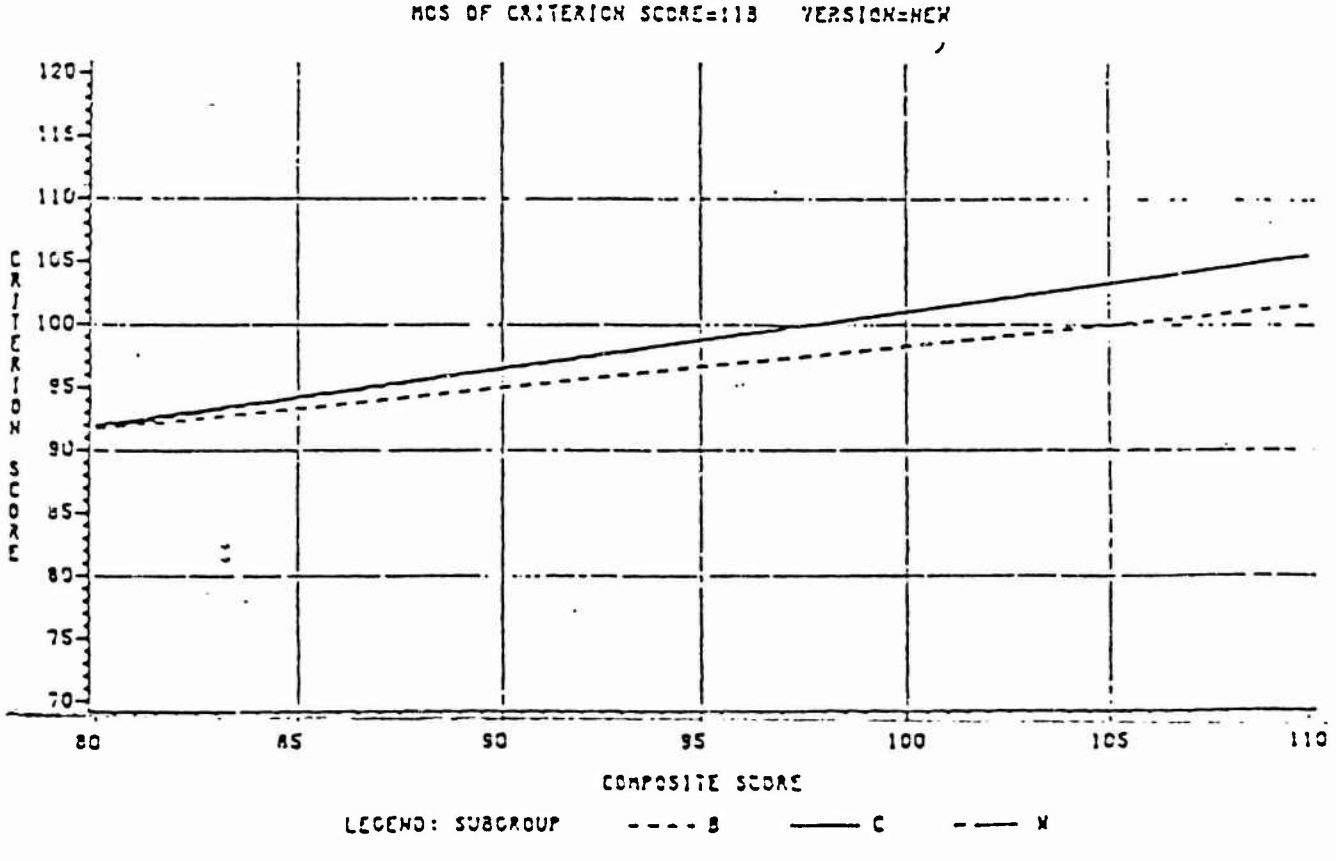
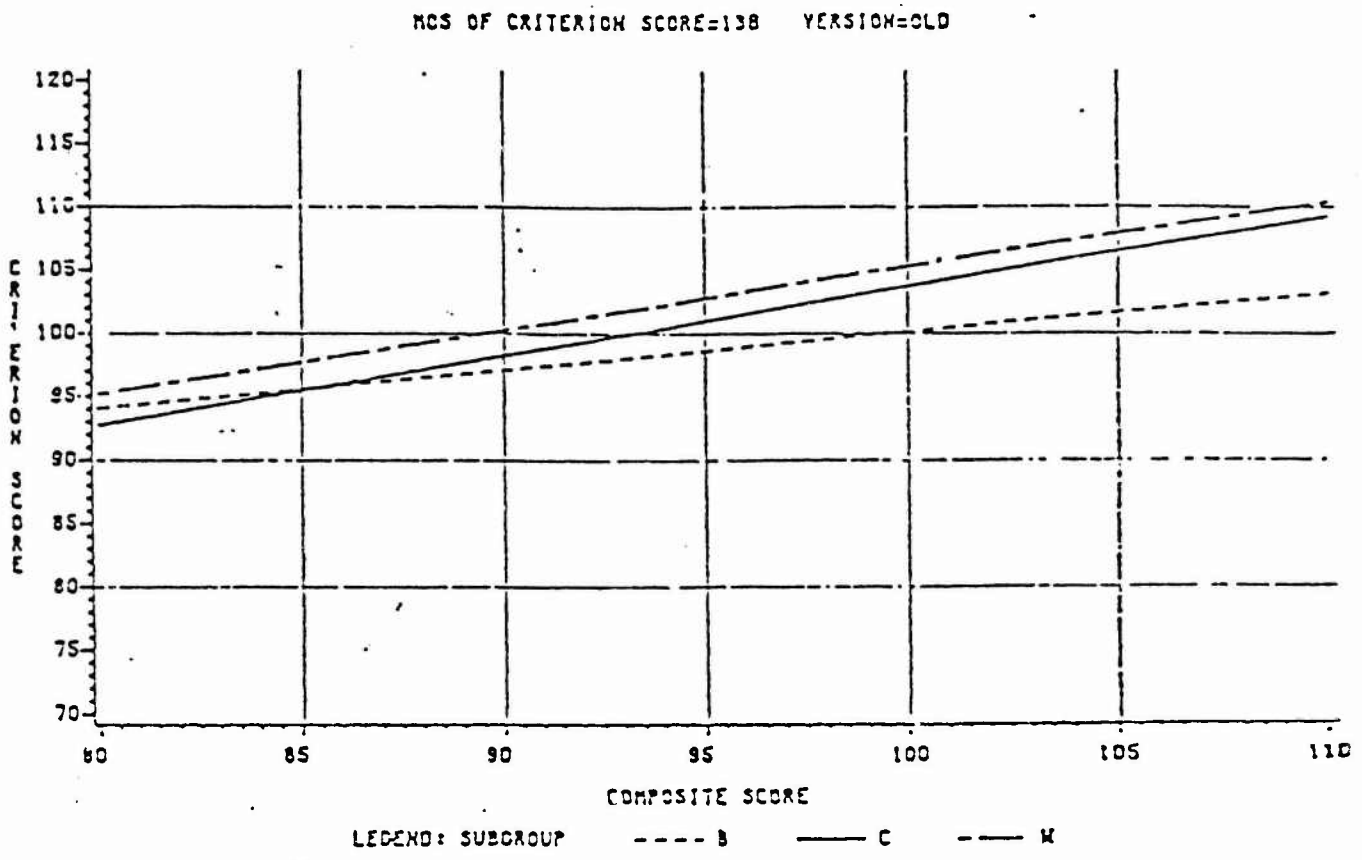


Figure 16 Regression lines for Current and Alternative (old and new) Composites for MOS 11B, by Race



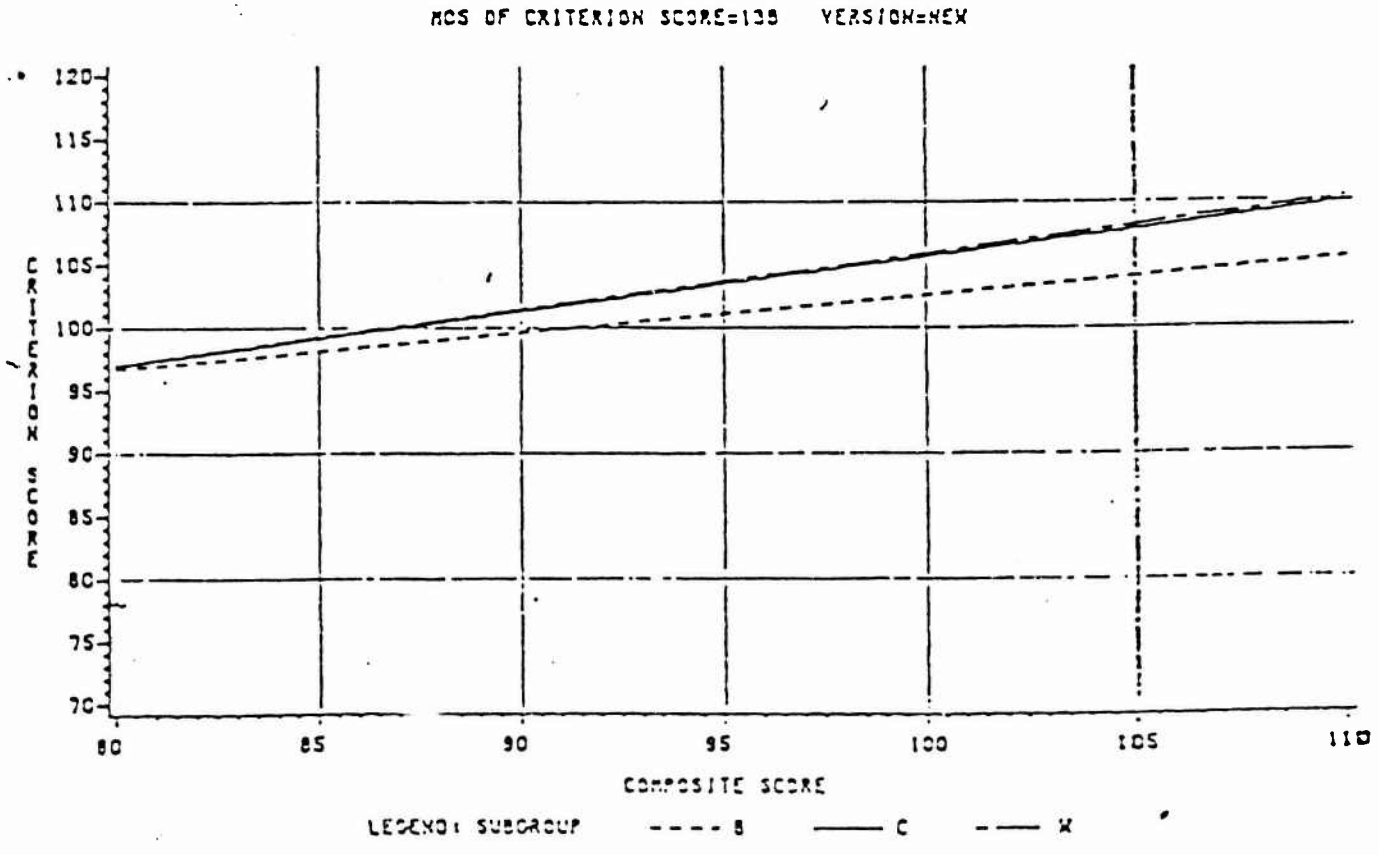
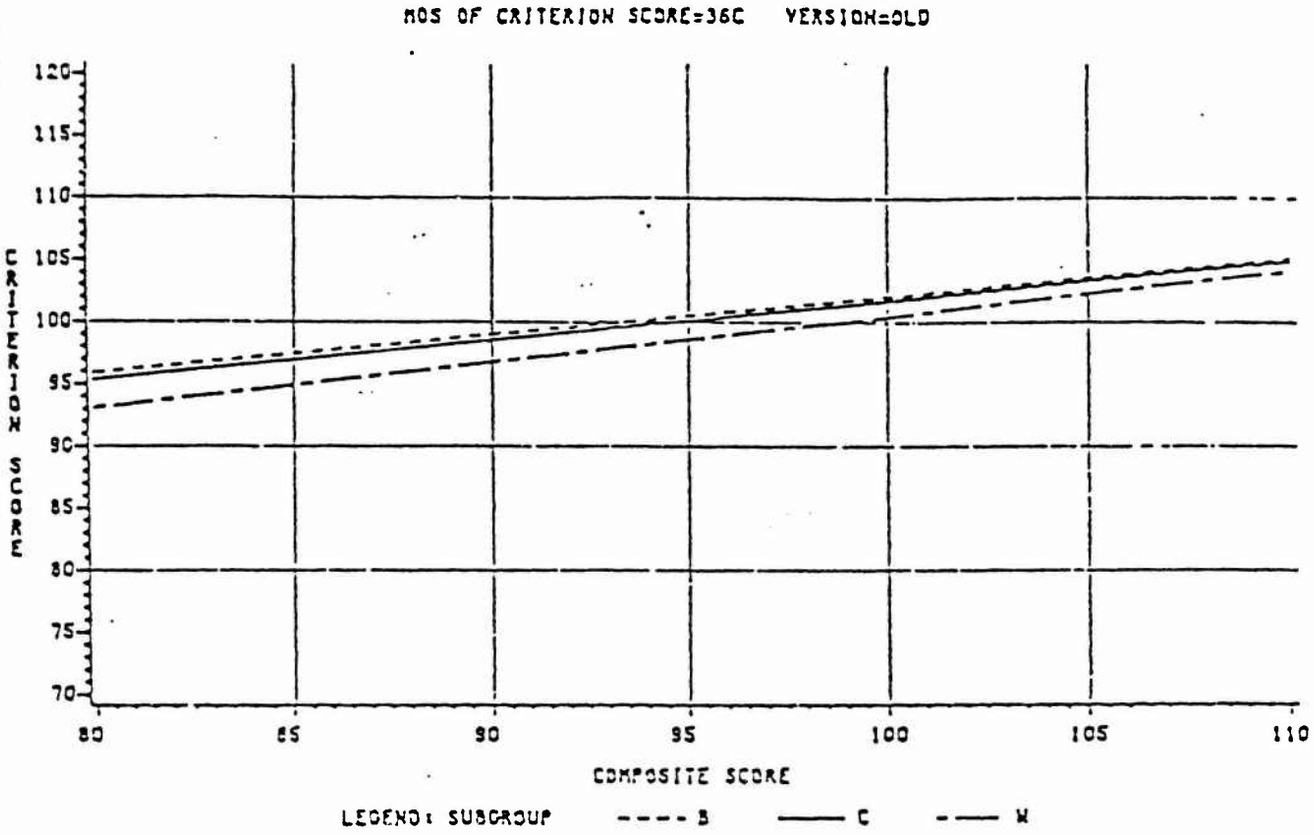


Figure 17 Regression lines for Current and Alternative (old and new) Composites for MOS 138, by Race



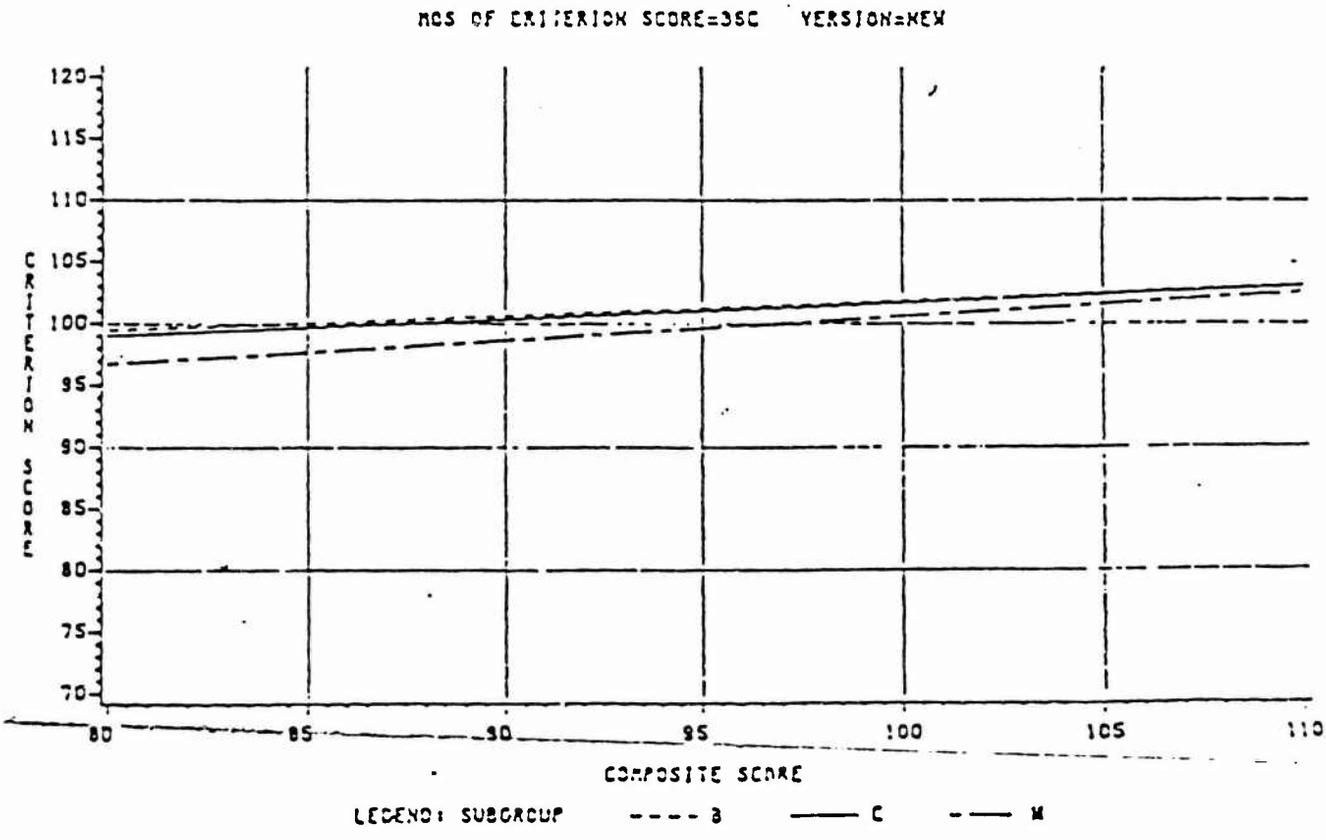
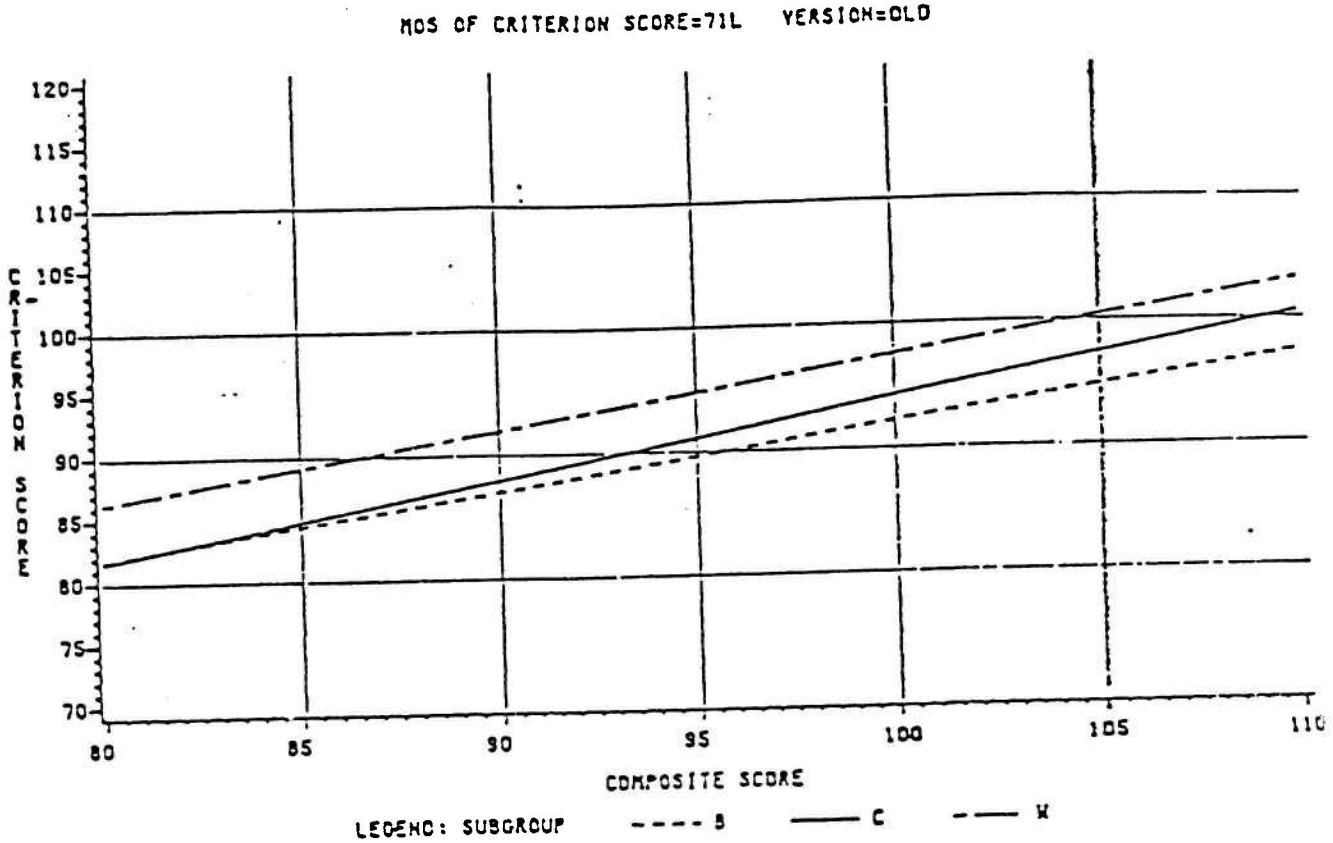


Figure 18 Regression lines for Current and Alternative (old and new) Composites for MOS 36C, by Race



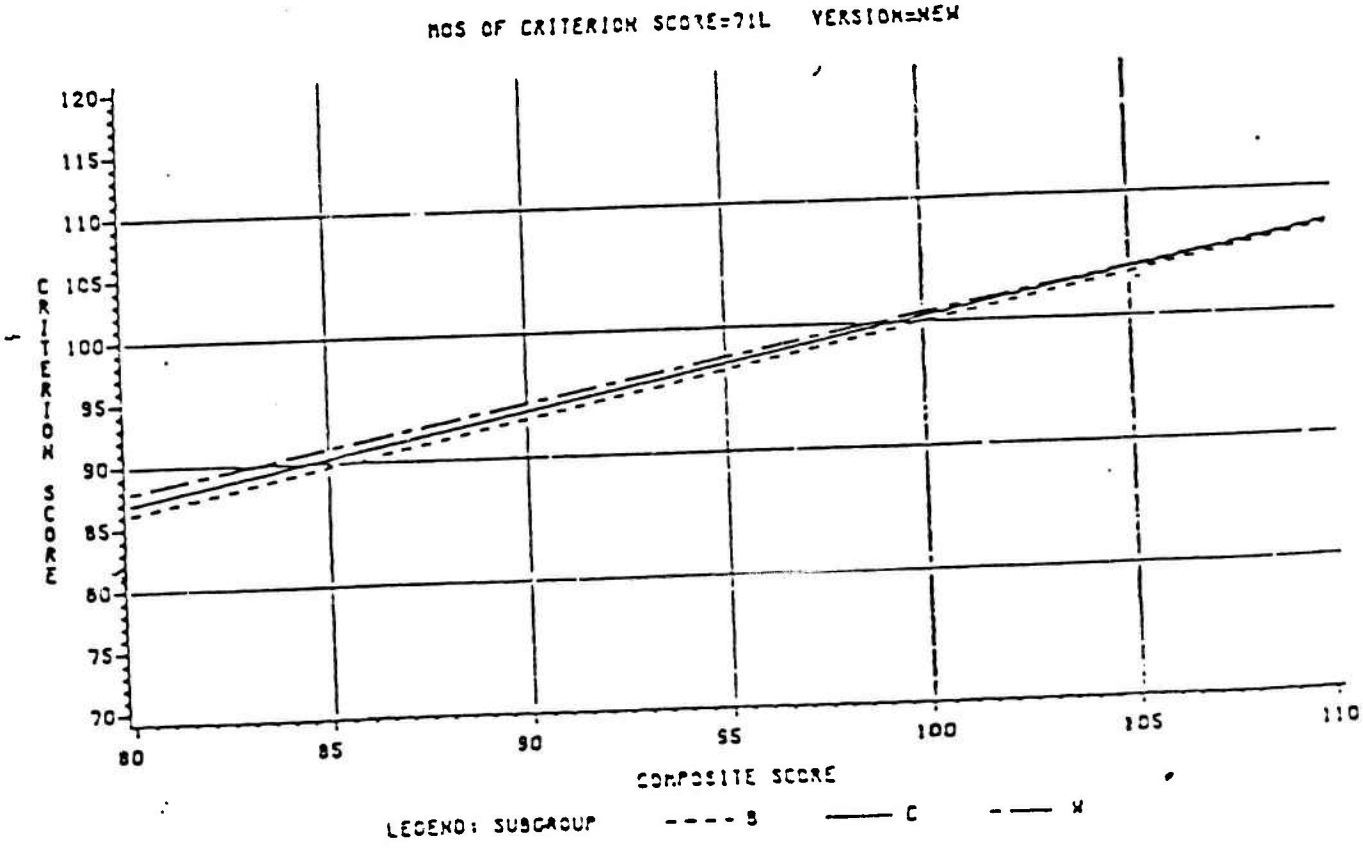
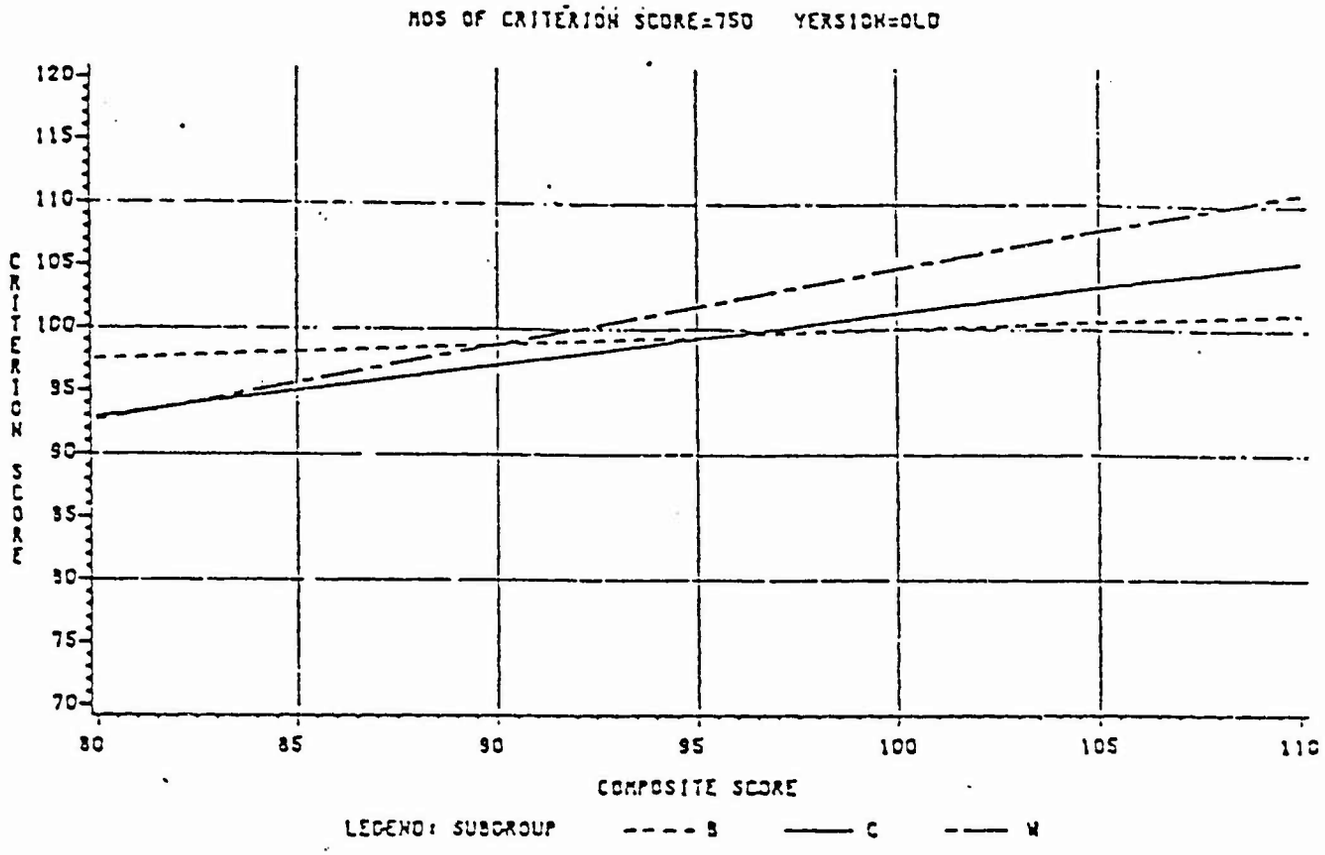


Figure 19 Regression lines for Current and Alternative (old and new) Composites for MOS 71L, by Race



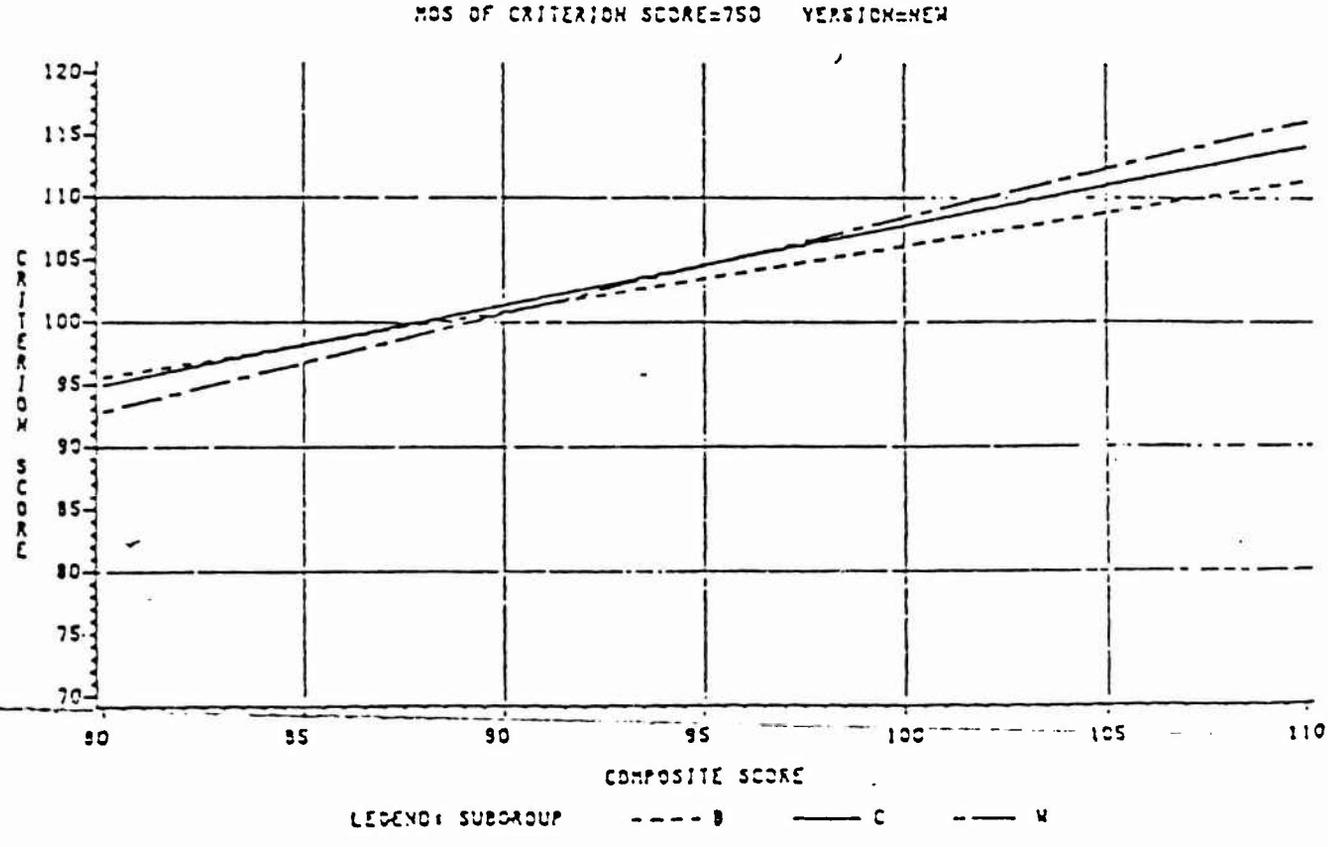
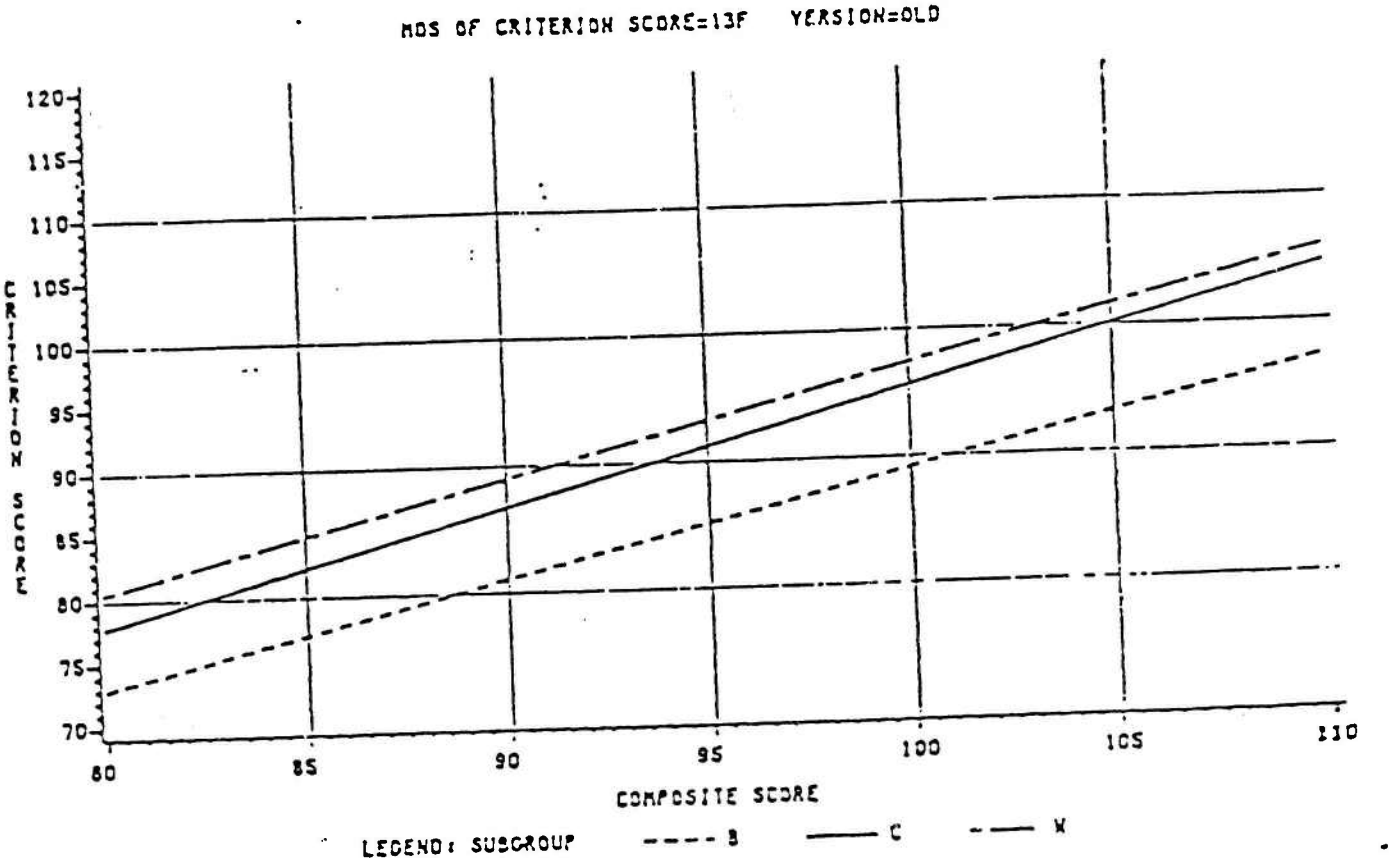


Figure 20 Regression lines for Current and Alternative (old and new) Composites for MOS 75D, by Race



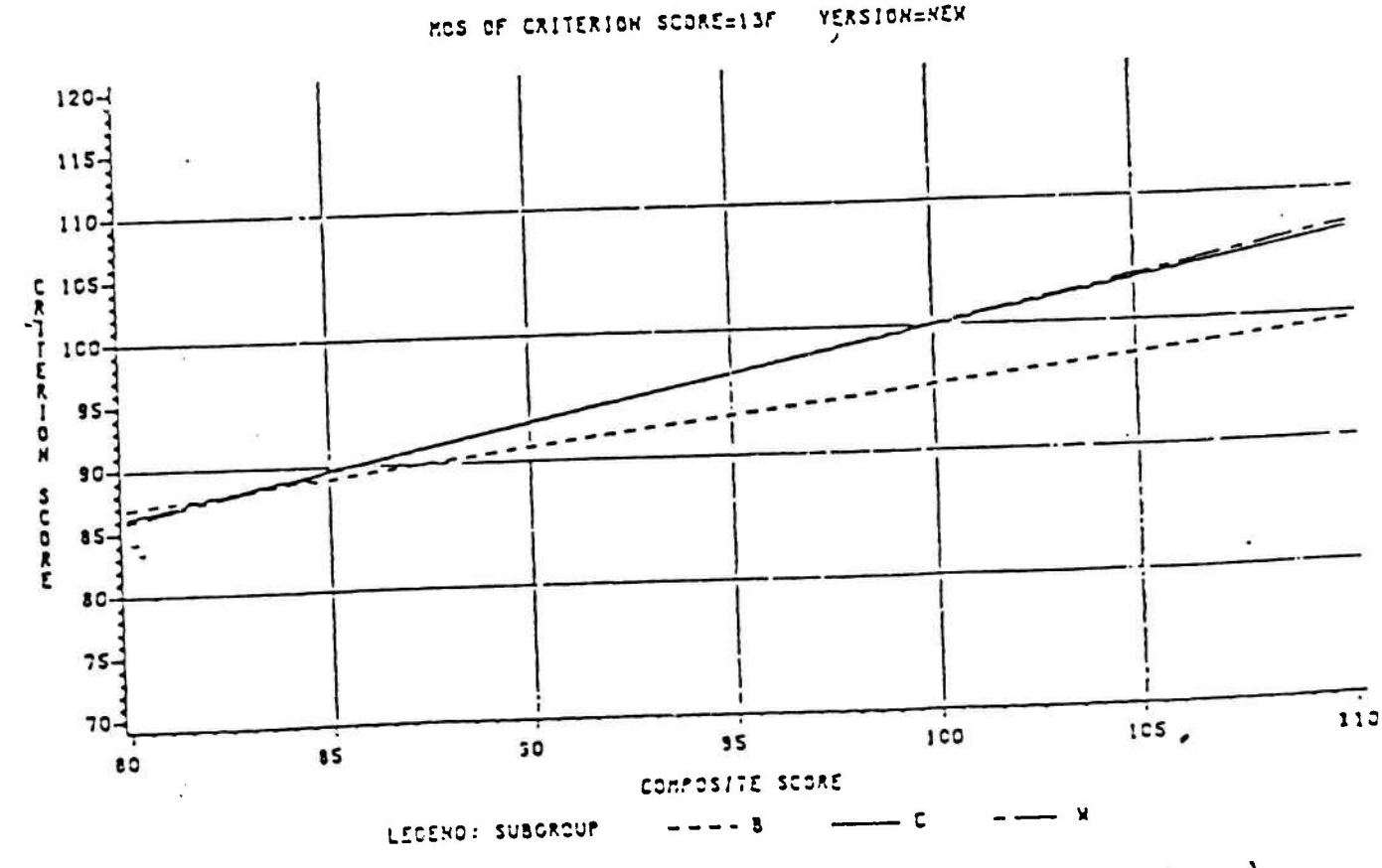
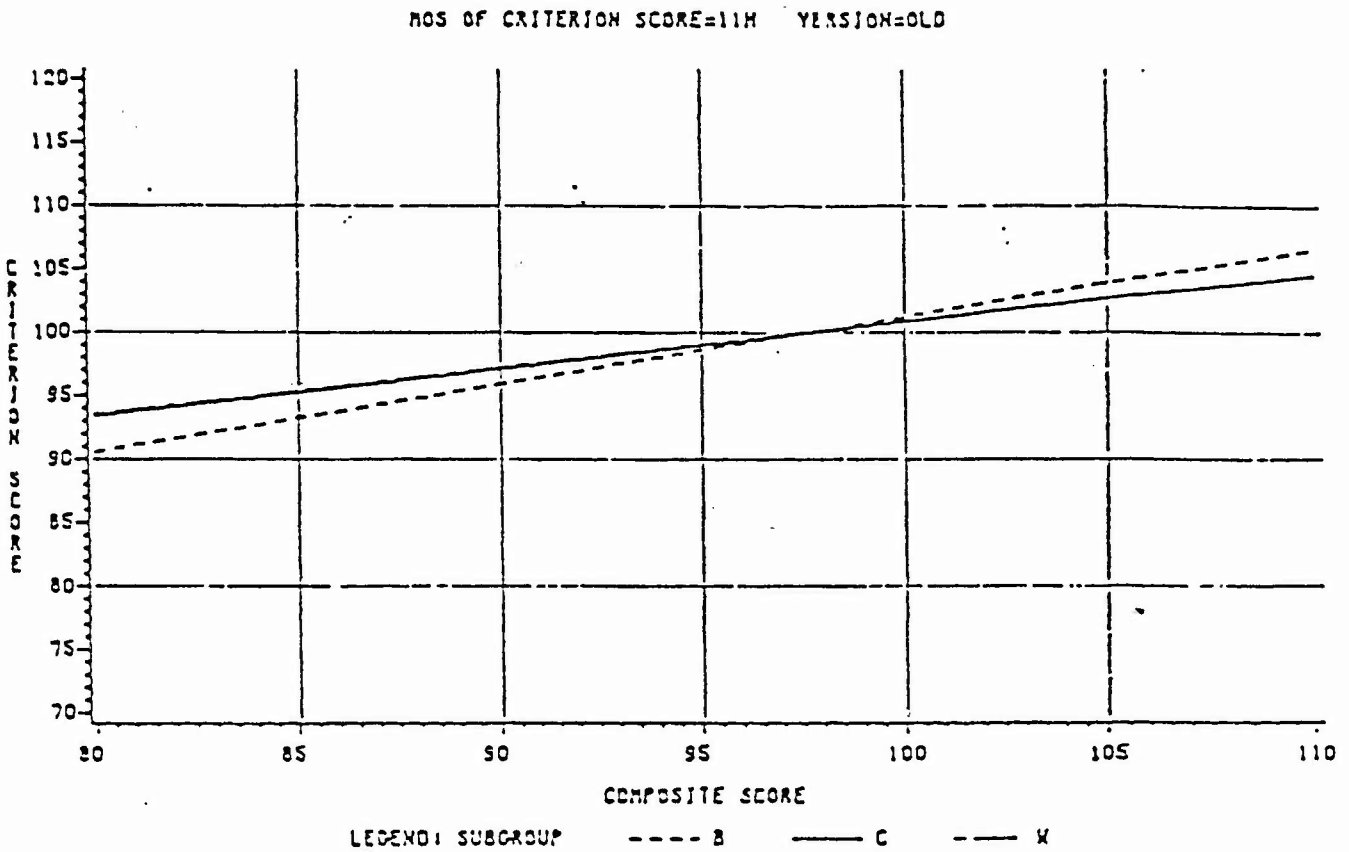


Figure 21 Regression lines for Current and Alternative (old and new) Composites for MOS 13F, by Race



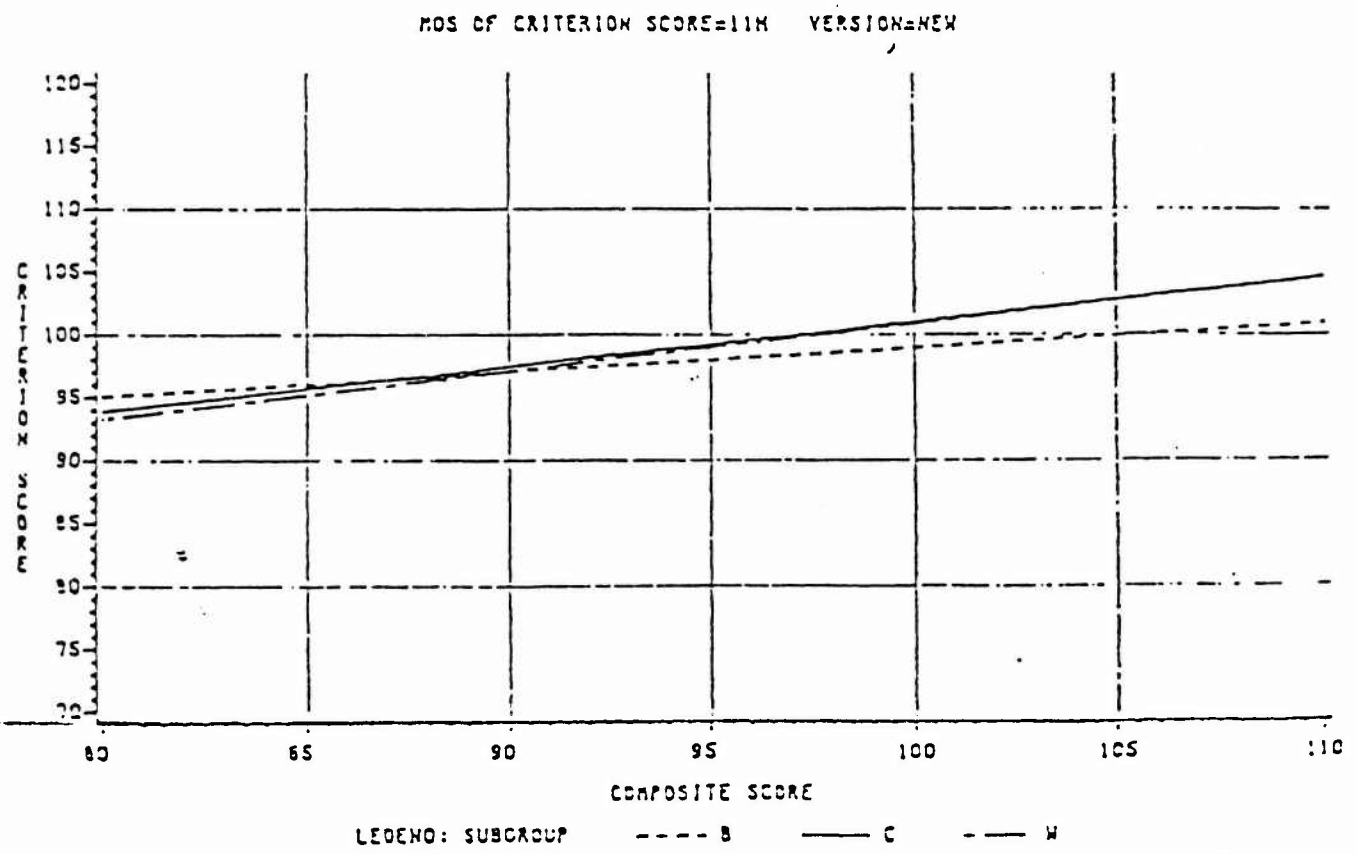
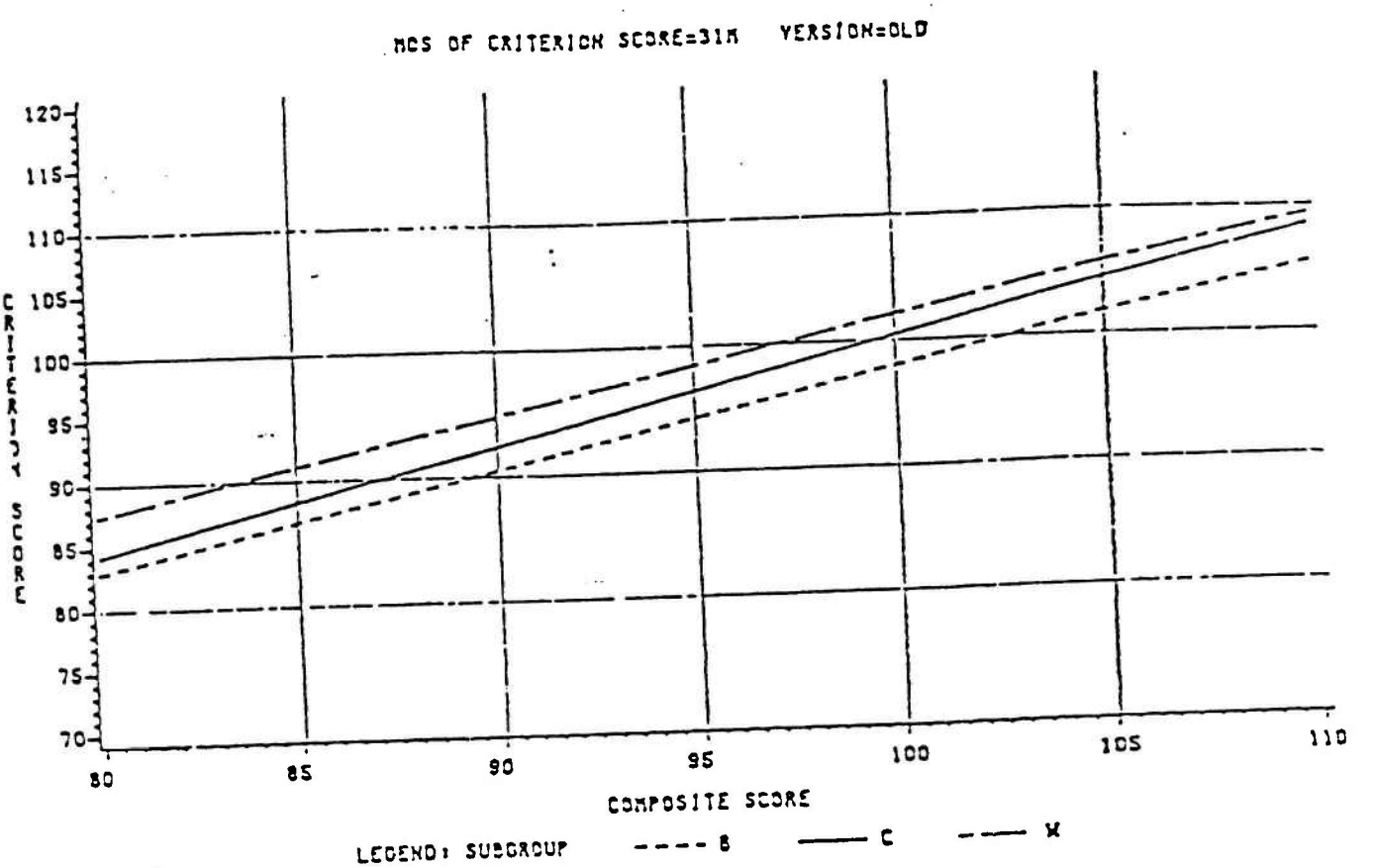


Figure 22 Regression lines for Current and Alternative (old and new) Composites for MOS 11H, by Race



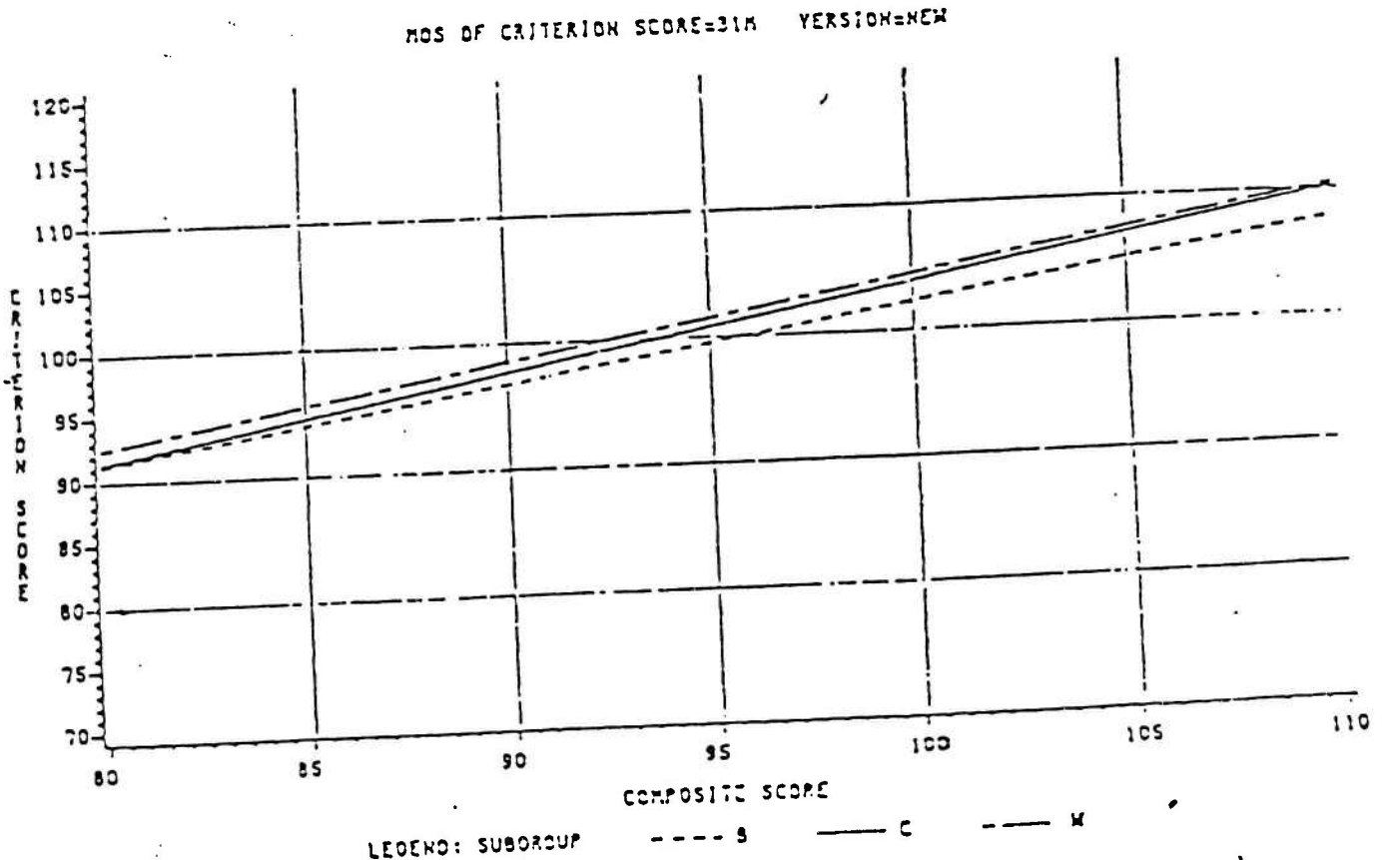


Figure 23 Regression lines for Current and Alternative (old and new) Composites for MOS 31M, by Race

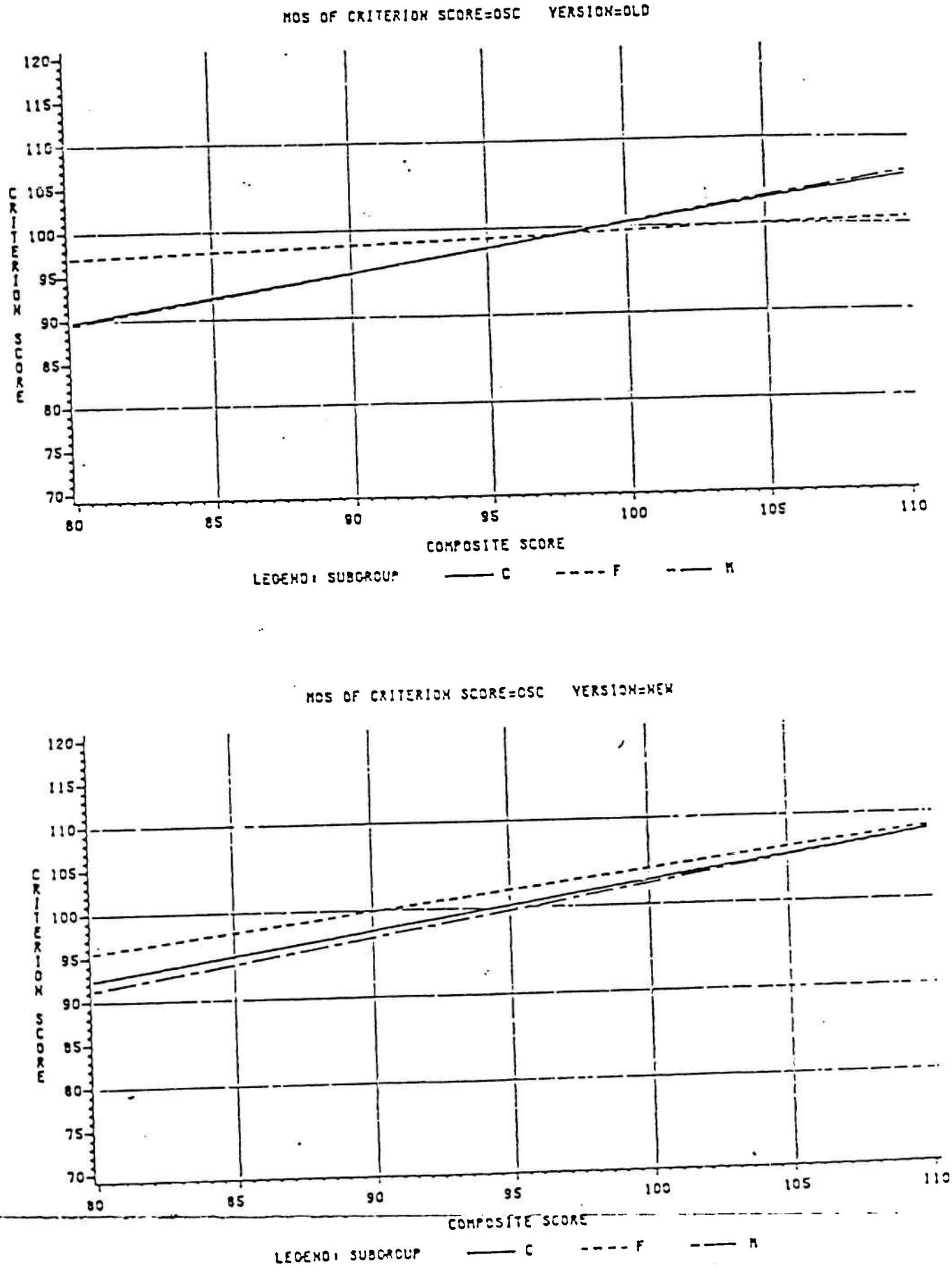
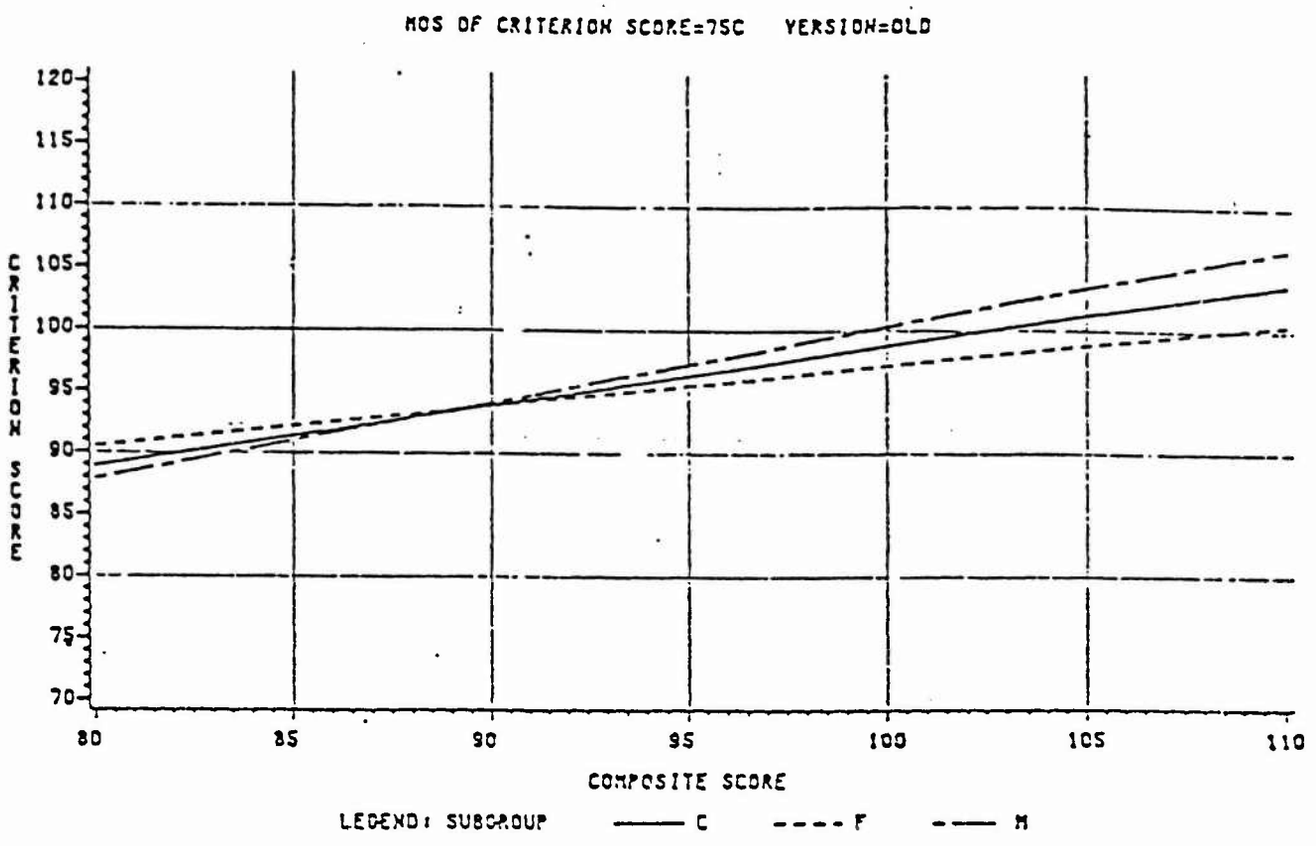


Figure 24 Regression lines for Current and Alternative (old and new) Composites for MOS 05C, by Gender



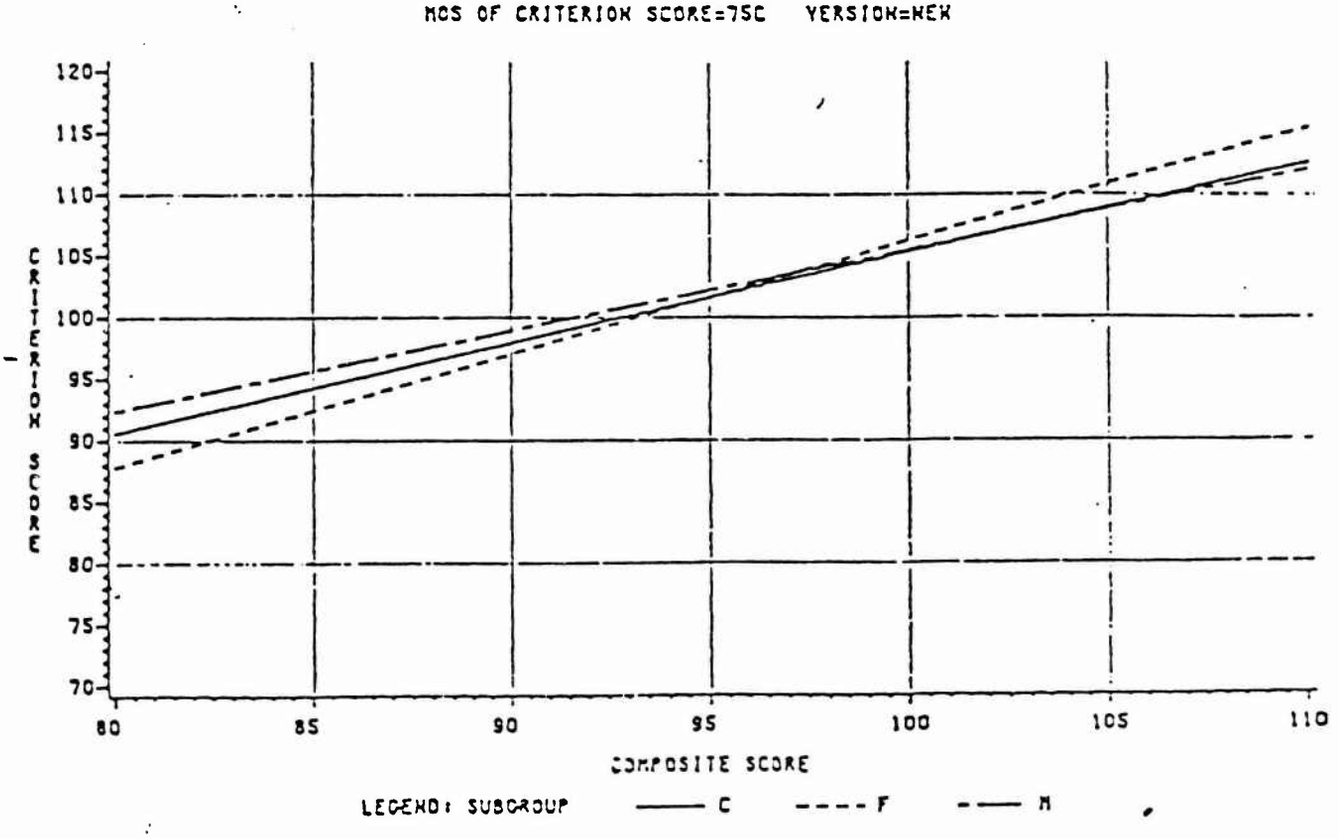
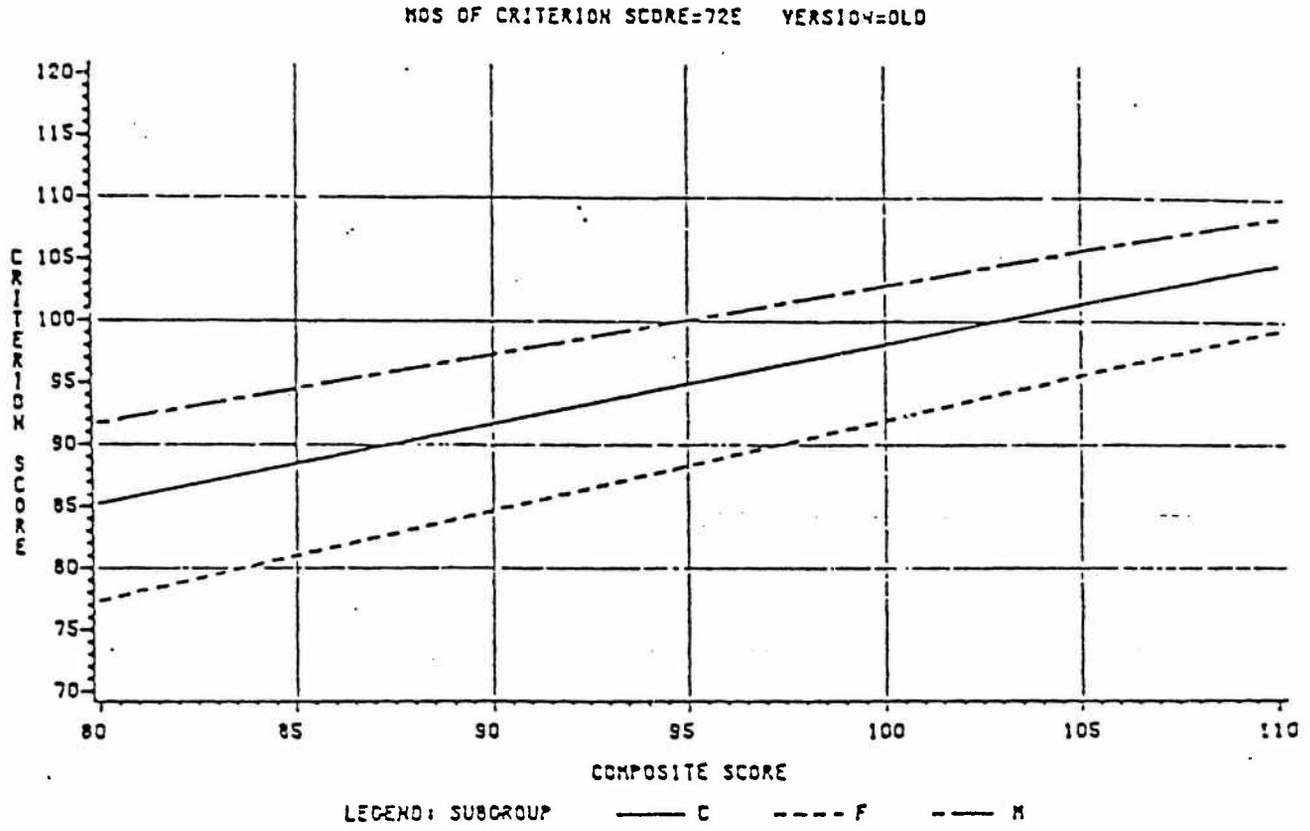


Figure 25 Regression lines for Current and Alternative (old and new) Composites for MOS 75C, by Gender



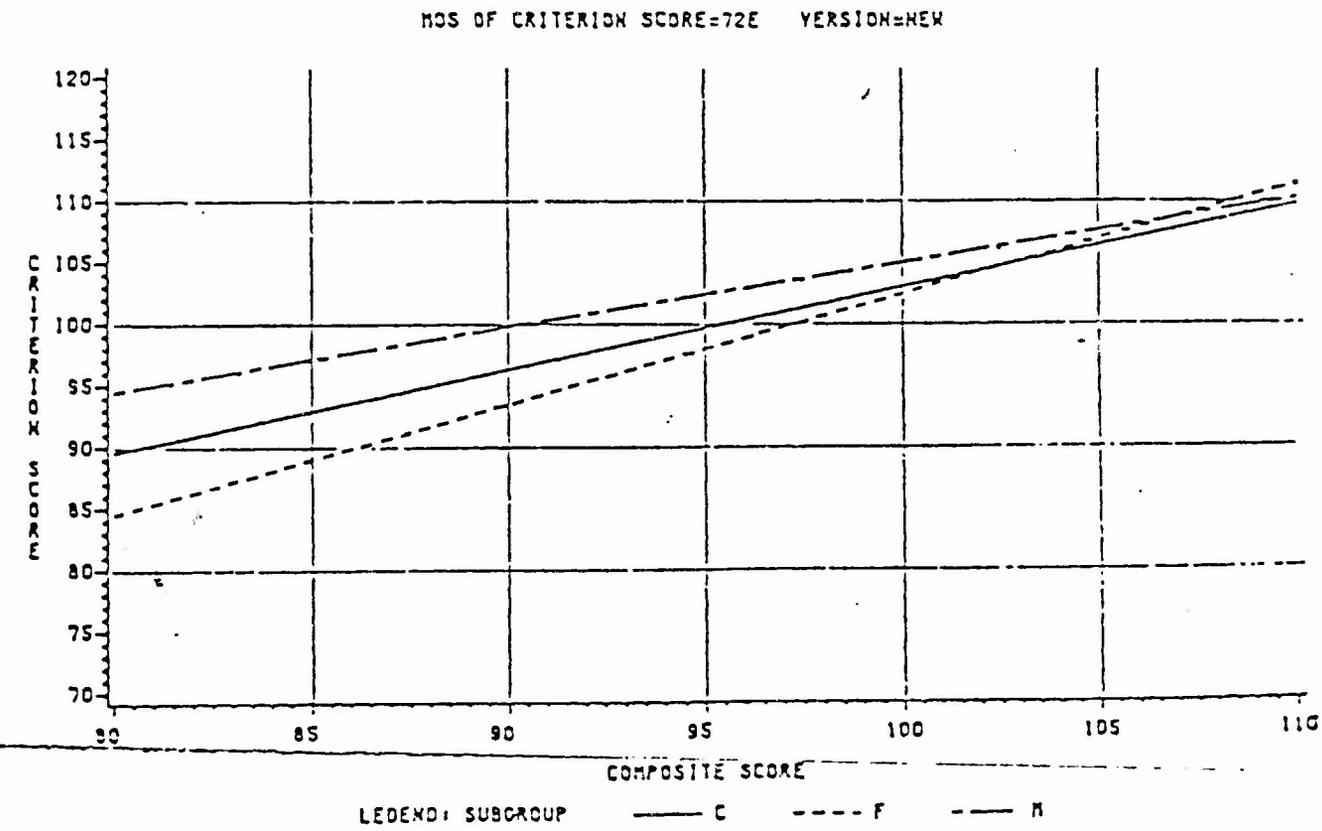
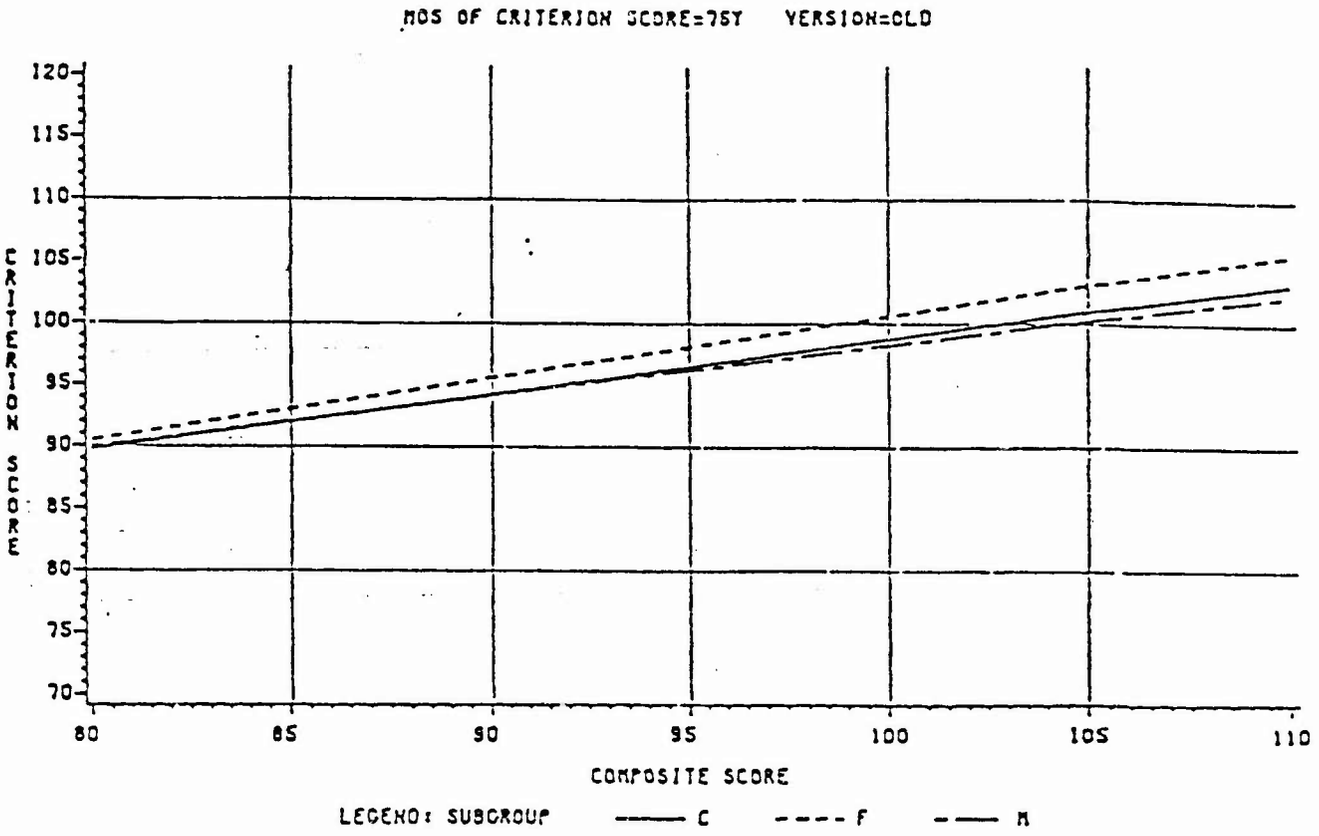


Figure 26 Regression lines for Current and Alternative (old and new) Composites for MOS 72E, by Gender



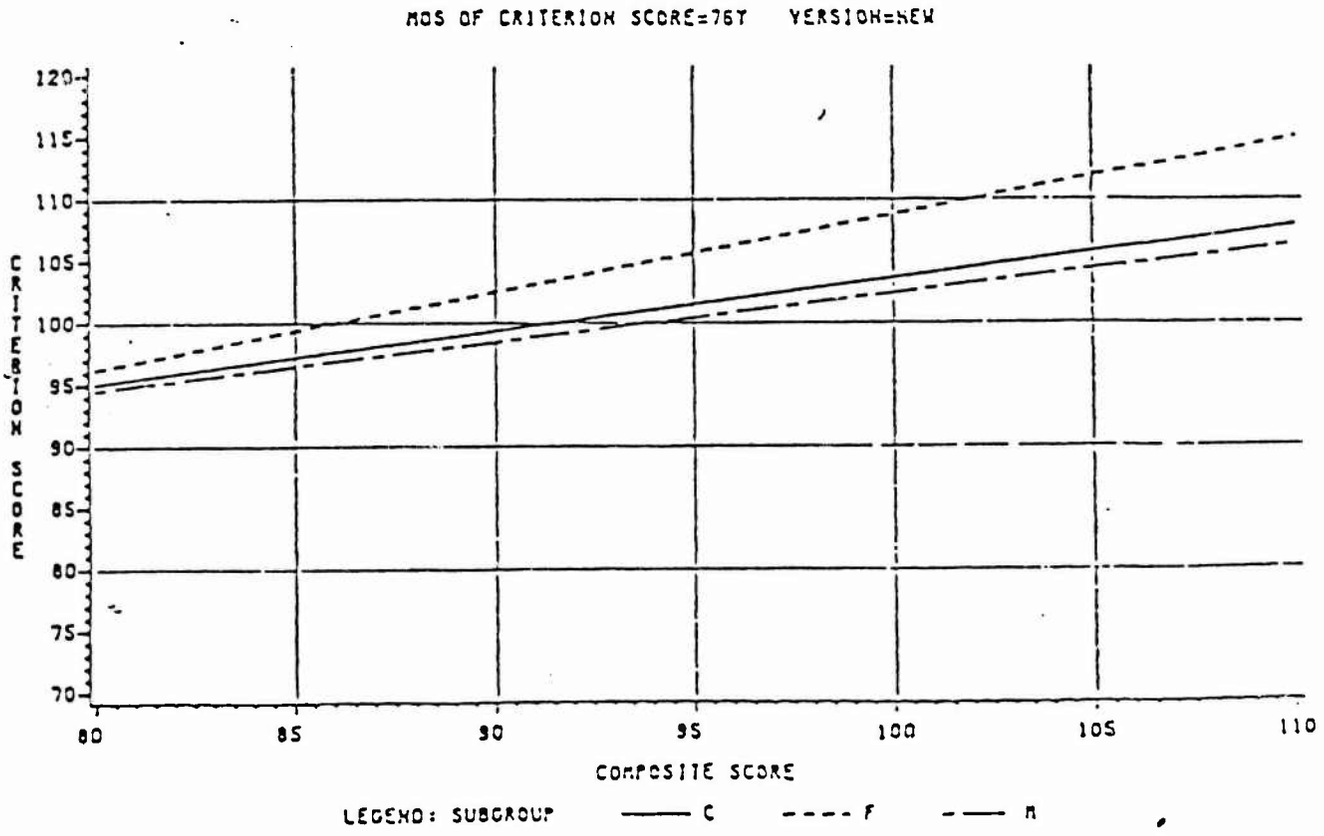


Figure 27 Regression lines for Current and Alternative (old and new) Composites for MOS 76Y, by Gender

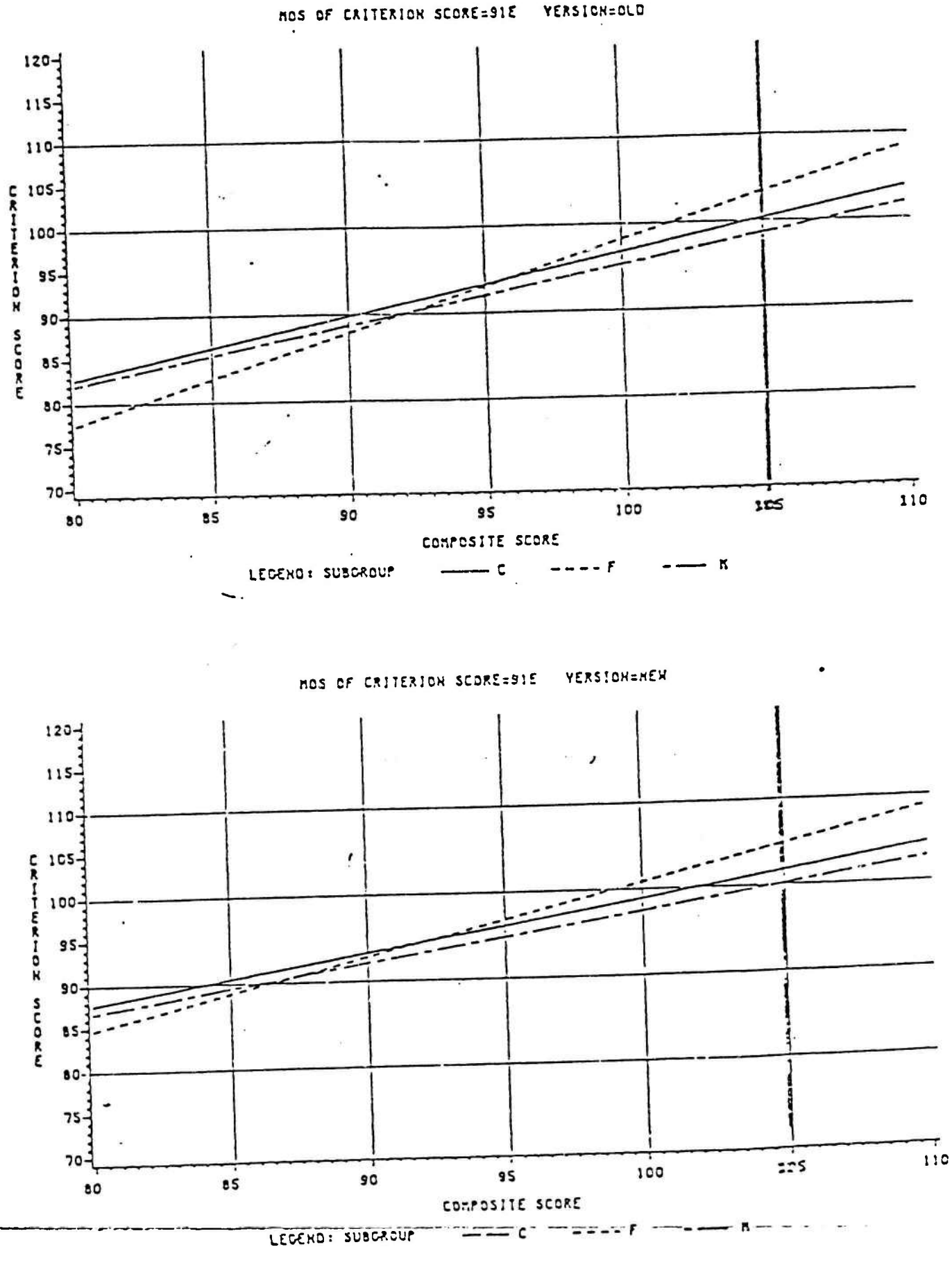
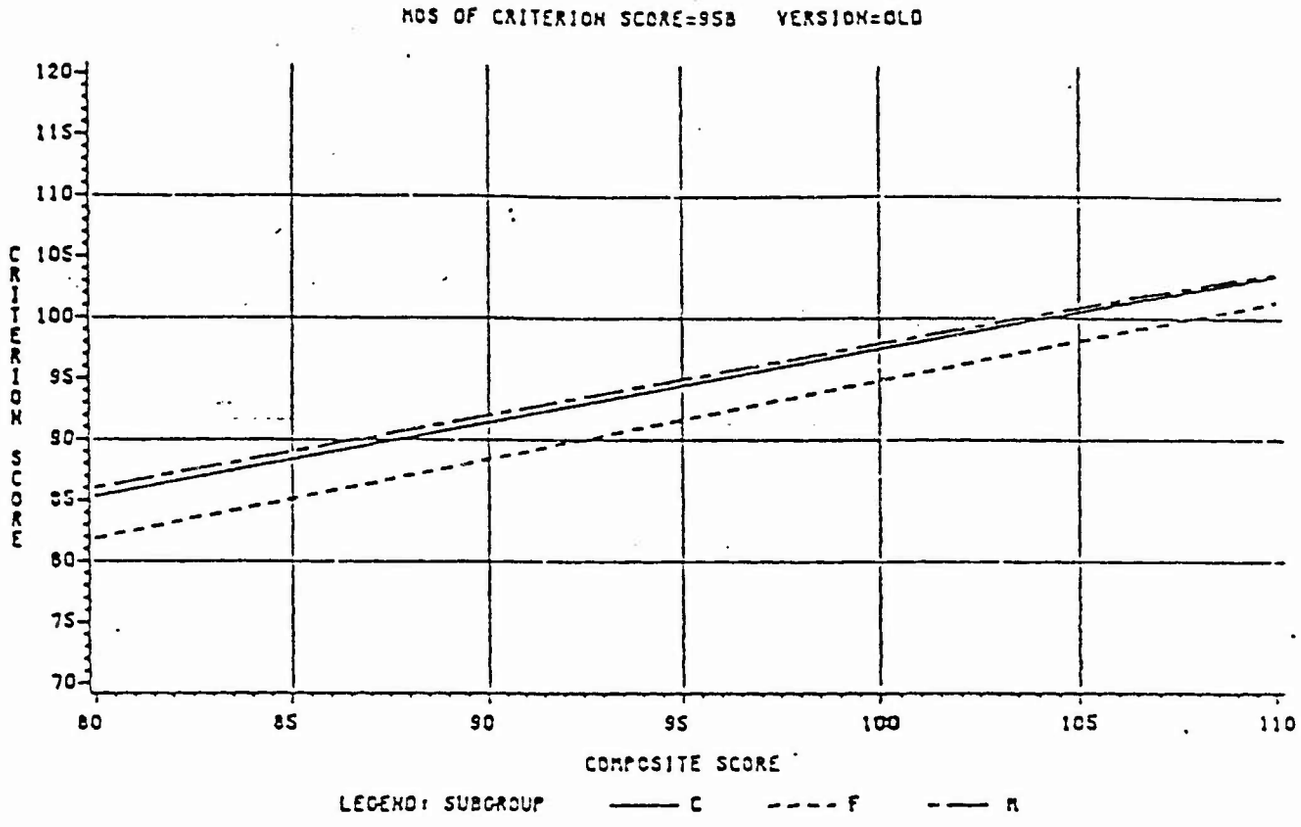


Figure 28 Regression lines for Current and Alternative (old and new) Composites for MOS 91E, by Gender



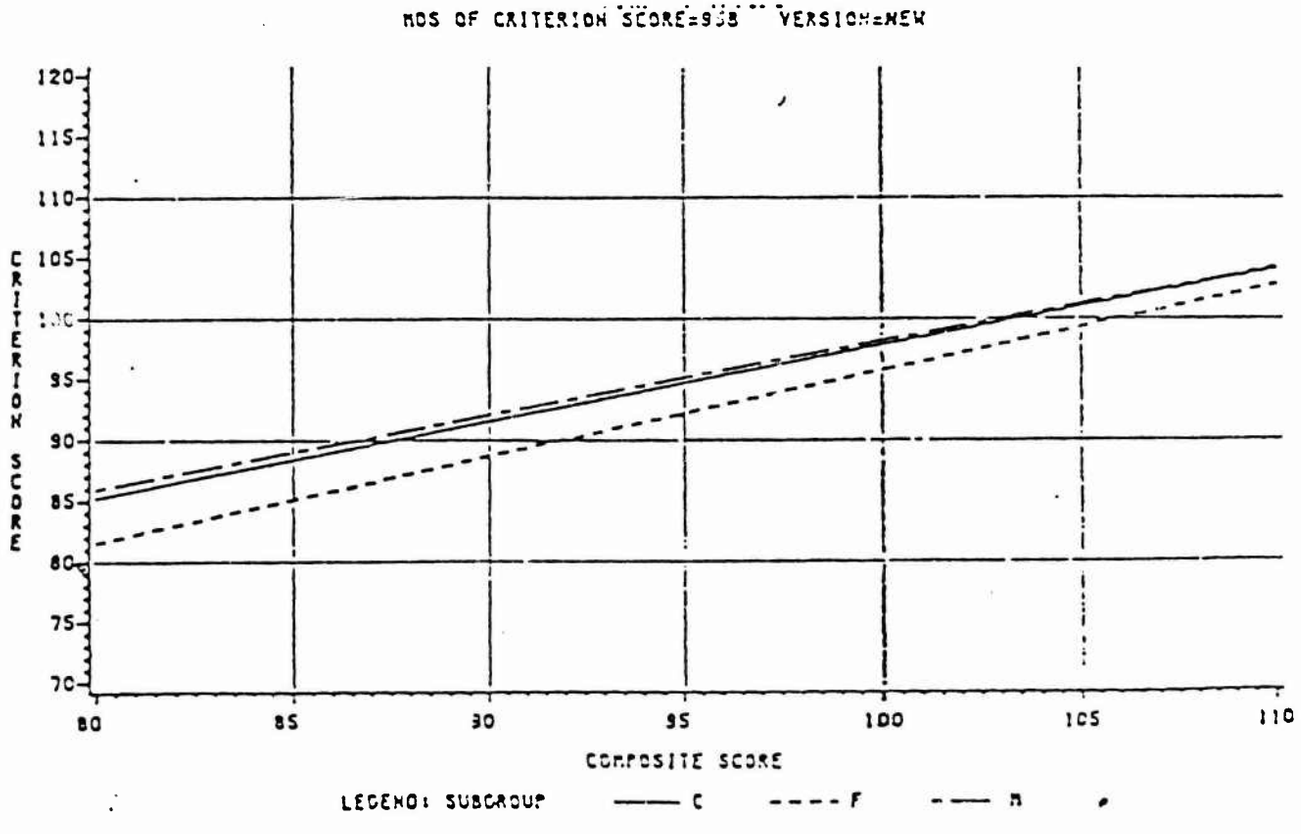


Figure 29. Regression lines for Current and Alternative (old and new) Composites for MOS 95B, by Gender

Table 11

Predicted Score Differences (Diff.) and Standard Errors (SE) of the Difference between Blacks (B) and Whites (W) for Particular MOS: Current Operational Composites

Cluster	N B	N W	Cutoff	Composite Score 85 Diff.	85 SE	95 Diff.	95 SE	105 Diff.	105 SE
CL Cluster									
MOS 71L	1229	1322	95	-4.69	1.44	-5.13	1.07	-5.51	.82
MOS 75D	481	205	95	-2.47	3.32	2.40	2.29	7.26	1.66
CO Cluster									
MOS 11H	122	769	85	2.14	1.89	.46	2.04	-1.21	2.31
MOS 11B	1146	4174	85	1.94	.74	2.46	.71	2.97	.77
EL Cluster									
MOS 31M	563	1185	95	4.38	1.33	4.18	1.03	3.98	.97
MOS 36C	214	132	90	-2.53	3.50	-1.93	2.36	-1.32	2.98
FA Cluster									
MOS 13F	125	657	100	7.74	2.00	8.10	1.67	8.46	1.65
MOS 13B	1814	2471	85	2.08	.80	4.12	.65	6.16	.63

changes as the composition of the composites changes. In this MOS, underprediction tends to occur for lower composite scores using the current composite. With the alternative composite, underprediction is observed for higher scores of the AA composite. The other MOS that showed a noticeable change among the regression lines with the alternative composites was 76Y. In this case the alternative composite tends to show somewhat more underprediction of female performance than does the current composite. While the average difference (3.25 points) in underprediction of the alternative versus the current composite for 76Y is small relative to the criterion standard deviation, the difference does approach statistical significance. This finding suggests that as new criterion data become available further attention and research be devoted to analyzing the differences between male and female soldiers in MOS 76Y. In most cases, however, as with the comparisons based upon race, the change to the alternative composite should not result in substantial underprediction of subgroup performance. The new composites could be used operationally without an increase in predictive bias in the selection and classification system.

Table 12

Predicted Score Differences (Diff.) and Standard Errors (SE) of the Difference between Females (F) and Males (M) for Particular MOS: Current Operational Composites

Cluster	N F	N M	Cutoff	Composite Score 85 Diff.	85 SE	95 Diff.	95 SE	105 Diff.	105 SE
CL Cluster									
MOS 76Y	248	888	95	-.97	3.11	-1.92	1.98	-2.86	1.41
MOS 75C	149	168	95	-2.59	2.88	.04	1.73	2.67	1.40
SC Cluster									
MOS 05C	260	1711	95	-5.44	3.05	-1.14	1.78	3.16	1.26
MOS 72E	237	325	90	13.53	2.72	11.80	1.78	10.07	1.60
ST Cluster									
MOS 95B	426	3269	100	3.84	2.45	3.28	1.55	2.71	.95
MOS 91E	117	184	95	2.71	4.89	-1.02	3.00	-4.74	2.17

The second finding of these analyses is that, as expected, the large MOS (eg. 11B, 13B, etc.) show patterns of under- and overprediction that are quite similar to the summary data presented earlier at the cluster level. For example, the large MOS in Tables 11 and 13 all show overprediction of black performance for both sets of composites. Such results are also presented in Tables 3 and 4 in the section discussing differences in the regression lines for the two races at the cluster or composite level. This result is not surprising since the MOS statistics were weighted by sample size when they were pooled to obtain the cluster data.

The third finding from these analyses is that within a cluster it appears that differences in the subgroup proportions can result in major changes in the pattern among the regression lines. For example, within the SC cluster, use of the alternative composites would result in underpredicting female criterion scores in MOS 05C where the ratio of males to females is 6.6. However, in MOS 72E where this ratio is only 1.4, use of the same composites would result in overprediction of female performance.

This finding suggests that it may be necessary to evaluate predictive bias at the MOS level. Each MOS in the sample could be analyzed using the Johnson-Neyman technique (See Rogosa, 1980) to determine whether a significant difference between the subgroup regression lines exists for any value of the composite. If a region of significance exists and includes the cutoff score for that MOS, further investigation of that MOS would be warranted. The aggregation of results to the cluster level might

Table 13

Predicted Score Differences (Diff.) and Standard Errors (SE) of the Difference between Blacks (B) and Whites (W) for Particular MOS: Four Alternative Composites

Cluster	N B	N W	Cutoff	Composite Score 85 Diff.	85 SE	95 Diff.	95 SE	105 Diff.	105 SE
CL Cluster									
MOS 71L	1229	1322	95	-1.47	1.35	-1.00	1.01	-.55	.77
MOS 75D	481	205	95	-1.44	3.14	1.07	2.17	3.57	1.57
CO Cluster									
MOS 11H	122	769	85	-.83	1.89	.99	2.03	2.82	2.30
MOS 11B	1146	4174	85	.79	.73	2.08	.71	3.38	.76
EL Cluster									
MOS 31M	563	1185	95	1.59	1.32	2.03	1.03	2.47	.97
MOS 36C	214	132	90	-2.35	3.51	-1.63	2.37	.90	2.99
FA Cluster									
MOS 13F	125	657	100	.45	1.99	3.38	1.67	6.31	1.65
MOS 13B	1814	2471	85	1.16	.79	2.64	.65	4.13	.63

best be done qualitatively. For example, the population of MOS within the cluster that show significant differences around the cutoff score could be reported.

Summary

The current and proposed alternative AA composites were investigated for possible subgroup bias in a number of ways, including analyses of predictive validities, comparisons of subgroup regression lines, and plotting the relationship of the subgroup regressions and the common regression line. All subgroups were found to be well predicted by the composites. Both sets of composites were found to show some small differences in predictive validity as a function of racial background and gender. The comparisons of regression lines indicated that while some MOS require further research (ie. 76Y), in general the use of either set of composites to select and classify enlisted personnel for the Army should not result in increased bias against blacks or women.

ARI Technical Report 651*
VALIDATION OF CURRENT AND ALTERNATIVE ASVAB AREA COMPOSITES,
BASED ON TRAINING AND SQT INFORMATION ON
FY1981 AND FY1982 ENLISTED ACCESSIONS

D.H. McLaughlin, P.G. Rossmeissl, L.L. Wise,
D.A. Brandt, Ming-mei Wang

This report describes a large-scale research effort to validate and improve the ASVAB Aptitude Area (AA) composites now used by the Army to select and classify enlisted personnel. Data were collected from existing Army sources on over 60,000 soldiers and over 60 MOS. The research had three major components: first, the composites now being used by the Army were validated; second, a new set of composites were derived empirically; finally both sets were compared on the basis of predictive validity, differential validity, and possible prediction bias. Both sets of composites were found to perform well, with the alternative set of four composites doing slightly better than the nine now in operational use.

---

* In press. To be available Defense Technical Information Center, 5010 Duke Street, Alexandria, VA 22314. Phone: (202) 274-7633.

Reports and papers on methodological issues

Methodological problems continued to receive especially careful consideration during validation work in Project A's second year because of the importance of establishing a definitive database and analytical guidelines for materials on which much of the research in the project's later years will be built. The wide scope and complex interrelationships involved in and among the various lines of inquiry complicated many methodological decisions. While most of the reports prepared during the year had methodological aspects of interest, three were primarily concerned with validation methodology for particular topical areas.

(1) Decisions made with regard to the scope, content, and organization of the data base system are described by Rossmeissl, Wise, and Wang. The RAPID data base management system was chosen to meet the demands that will be placed on the project's Longitudinal Research Data Base, which must provide many research teams with access to a vast amount of interelated data that will be assembled over the remaining years of the project.

(2) Application of meta-analytic techniques in estimating criterion-related validity of cognitive tests, with reference to selection and classification, is discussed by Rossmeissl and Stern. Primary attention is given to three of the possible sources of error in validity estimates: sampling bias, unreliability of the criterion measure, and restriction in range of the predictors.

(3) Adjustments for the effects of range restriction on the validity of the current ASVAB composites are described in a paper by Brandt, McLaughlin, Wise, and Rossmeissl. The results indicate that, in general, the composites provide information that is relevant to predicting performance in training and on the job.

A Data Base System for Validation Research

Paul G. Rossmeissl
Army Research Institute

Lauress L. Wise and Ming-mei Wang
American Institutes for Research

October 1983

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Presented at the 25th Annual Conference of the Military Testing Association at Gulf Shores, Alabama.

A Data Base System for Validation Research

Paul G. Rossmeissl
U.S Army Research Institute

Lauress L. Wise
and
Ming-mei Wang
American Institutes for Research

INTRODUCTION

The Army Research Institute is currently engaged in two large scale research projects in order to develop a new selection and classification system that will improve the efficiency of personnel utilization within the Army. The purposes of the first project, Development of Improved Army Selection and Classification Systems, are: to validate current predictors, to develop new or improved predictors and performance measures, and to conduct a longitudinal validation of current and newly developed classification measures for prediction of the enlistee's performance from training through the second tour of duty. The second project, Development of an Enlisted Personnel Allocation System (EPAS), is to develop a state-of-the-art personnel assignment system, to facilitate the initial enlistment decisions.

The research progress of these projects will depend to a large extent on the vast amount of interrelated data that must be assembled in a manner that will provide access to the many research teams involved and still maintain the integrity and privacy of the data. This paper describes the system planned for maintaining these data that will be needed during the several years it will take to complete this research.

THE DATABASE MANAGEMENT SYSTEM

At the time of the initiation of these projects a database management system (DBMS) named RAPID (Turner, Hammond, & Cotton; 1979) was identified as the most cost effective DBMS that could meet the needs of the projects. RAPID is a relational data base system developed by Statistics Canada and was especially designed to accommodate large statistical data sets. The decision to use RAPID was based upon three important features of the system.

The first of these features was that RAPID was designed to provide for a significant degree of data compression. This feature means that it is feasible to store much more of the data on disks or mass storage units rather than on tape, which will greatly increase the speed in which data can be retrieved.

The second advantage of the RAPID lies in its storage and access mode. RAPID uses a "transposed file" organization, which means that it stores together all the information on a single variable rather than all of the information on a single "case" or observation. Data are stored in direct access files with appropriate indices so that the system can read selected variables without having to read through the entire file. The standard statistical packages, like SAS and SPSS, in contrast, employ a sequential access mode and store data case by case. With such a system, even when only a few cases and variables are required, the entire file must read in order to select the desired information. Most other common DBMSs do use direct access files, but still store the information by case so that they only add additional overhead in accessing selected variables.

In order to estimate the true value of these two advantages, tests were run comparing the RAPID system with SAS to determine the storage and retrieval cost-effectiveness of each. To accomplish these tests; files of Army accession data of various sizes were entered into each system. Each file consisted of forty variables. A file of sample size n=o was also loaded into each system to estimate the amount of "overhead" (data descriptions, etc.) that each system requires prior to storing the data. The storage space comparison for the two systems is given in Table 1.

Table 1

Rapid vs SAS
Storage Space Comparison

file size	no. of tracks used RAPID	SAS	% savings
n=0 (overhead)	12	5	-
n=500	23	37	.38
n=2000	90	130	.31
n=20000	824	1255	.34

While RAPID initially requires more overhead than SAS it appears that using RAPID to store real data will result in space savings of about 34% in comparison to SAS.

To estimate the advantages of RAPID's transposed file structure tests, were run to determine the cost of creating a workfile

(a SAS data set containing the desired variables) from both of the two systems. The results of this comparison are given in Table 2. In all situations RAPID required less I/O than SAS to

Table 2

RAPID vs SAS
Creation of a Workfile

	I/O Count Index	
	RAPID	SAS
File size		
n=500		
no. of variables in workfile		
1	81	166
8	95	182
n=2000		
no. of variables in workfile		
1	88	260
8	134	272
n=20000		
no. of variables in workfile		
1	362	1652
8	1429	1710

create the workfile. This advantage would be particularly important for computer systems that place a heavy charge on I/O procedures.

The final advantage of RAPID is that it provides convenient interfaces with both SAS and SPSS (as well as some other) statistical packages. This feature facilitates the creation of special work files and allows the use of SAS to manipulate data to be loaded into the data base.

SECURITY PROCEDURES

Whenever a large amount of data on individuals is maintained and stored, it is necessary to protect that data from compromise. The security of the Project A and B data base is particularly important for a number of reasons. Some of the data collected on individual soldiers, such as promotions, paygrade, or disciplinary actions, will be private in nature, and the privacy of that information must be protected. Since many researchers will be accessing the data base for a variety of uses, the integrity of the data must be maintained in a manner that insures the data remain accurate and consistent across uses. Finally, it is necessary to secure the data base so that the Army maintains complete ownership of the data, to insure that the data within the data base are used only for authorized project A and B research.

ARI proposes to protect the security of the data base in three ways. Soldier social security numbers (SSNs) will be routinely encrypted to insure the privacy of each soldier's records. Access to the data base will be carefully controlled both to further protect soldier privacy and to insure proper use of the data. Finally, a log will be maintained for the system that will note each attempted access of the data base and whether or not the access was authorized. Each of these procedures will be outlined in turn below.

The key procedure in guaranteeing the privacy of individual soldier data is the coding or encrypting of each soldier's identifier. This encryption is accomplished by scrambling each soldier's SSN in an unpredictable manner. The specific algorithm that does the encrypting (and if needed, decrypting) is known only to the data base administrators, with a printed copy of the algorithm being securely maintained. All of the data files of the data base that are routinely accessed and any project workfiles that are generated from the larger data base files will use only the encrypted SSN as an identifier.

Data integrity and accuracy are maintained by controlling the access to the large files or relations within the data base. This procedure also helps contribute to the privacy protection of soldier records. The system uses the RACF procedure at NIH to restrict the access of selected files to authorized users. Under RACF different degrees of access can be structured. By specifying a "universal access" of "NONE", access can be restricted to only those users granted specific exceptions. Users would then have to provide an eight character RACF password (different for each user) in order to read the datafiles for which they are authorized. Using the provisions of RACF, a series of access "levels" has been installed to provide timely access to relevant data needed by project researchers and yet protect the data from compromise.

At the highest level of access are the database administrators who will have access to all of the files and relations within the data base, and are the only personnel who will be able to enter data into the data base or to modify data already stored in the data base. Thus, the data base administrators have responsibility for all data entry and editing. It is also the duty of the data base administrators to create workfiles based upon database files and relations as they are needed by other project researchers.

Project personnel with the next highest data base access authority will be able to read data from all of the files within the data base, with the exception of the Link File which contains all of the basic identifying information for each soldier. This exception is being made to help maintain soldier privacy. Only a few of the project staff will be at this access level and their primary responsibility with regard to the data base will to be to back up the data base administrators.

Most project researchers will have some level three data base access. Researchers at this level will have direct access to those files that are generated by the particular issues they are investigating. They will also have direct access to files created by other research that should have a direct influence upon their research. For example, researchers investigating the development of new preinduction predictors of performance will have direct access to the task analysis data that is being collected by those investigating criterion development so that the new predictors that are developed will address areas of the criterion space not currently covered by ASVAB.

However, the most common way in which researchers will obtain data base records is through the creation of workfiles. By requesting the creation of a workfile, a scientist will be able to obtain data from all of the large files in the data base (once again with the exception of the Link File which will always be kept secure and private). The key aspect of workfiles from the aspect of data security is that the scientist will only receive the data that he or she requested, and there will be a precise record of who requested what data. When a workfile is needed the researcher will submit a data request form which will ask questions like: who needs the data, what variables are needed, why are they required, and what will be done with the data after its current use is completed. In addition, each data request form will remind the scientist seeking the data that all data base records are the property of the Army and are to be used only for legitimate project research.

As a final security practice the procedure used to execute the RAPID DBMS's data retrieval programs has been modified to log a record of each access or attempted access to the data base. This access log will be routinely reviewed to assure that no inappropriate access has been attempted. In addition, the monthly accounting information of each project user will be monitored for any indication of unauthorized access to the data base. These audit trails will serve as a second level of protec-

tion against unauthorized use of the data by anyone who manages to obtain the necessary RACF passwords. They will not directly prevent access to the data base, but the threat of exposure should serve as a deterrent to attempts at unauthorized data base retrieval. The log will also help the data base administrators decide which project files should be stored on disk rather than tape by providing information as to how frequently data are requested from any given file.

Any set of procedures designed to store data electronically needs to balance the ease with which data can be accessed against the security of the data base. The procedures described here tend to favor the security aspect of this balance. The number of data files that most project scientists will be able to access directly will be small in comparison to the total amount of stored data. Furthermore, only the data base administrators will access to the true soldier identifying information and be able to add or modify data. However, these limitations should not prove to be too restrictive since prompt creation and efficient use of workfiles should provide each scientist with the data that he or she needs to perform the required research.

REFERENCE

Turner, M. J., Hammond, R., & Cotton P. A DBMS for Large Statistical Databases, Paper presented at the 42nd Biennial Congress of the International Statistical Institute, Manila, Phillipines, December 1979.

The Application of Meta-Analytic Techniques in Estimating Selection/Classification Parameters

Paul G. Rossmeissl
Brian M. Stern

Army Research Institute

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Paper presented at the Psychonomics Society, San Diego, November 1983.

The Application of Meta-Analytic Techniques in Estimating Selection/Classification Parameters

Paul G. Rossmeissl
Brian M. Stern

U.S. Army Research Institute

One of the oldest problems confronting research scientists is that of combining findings from a number of research settings. Recently Hunter, Schmidt, and Jackson (1982) proposed a set of meta-analytic techniques to serve as the basis for a quantitative integration of research findings across different experimental settings. The first purpose of this paper is to illustrate some of these techniques as they might be applied to the investigation of the criterion-related validity of cognitive tests. The results of these analyses will then be used to draw conclusions about cognitive test validities and how they should be interpreted.

The key concept underlying the meta-analytic approach is that the variance of any statistic, in this case validity coefficients for different job-test combinations, can be divided into components corresponding to true and error variance. In this manner the approach is conceptually similar to traditional analysis of variance. A meta-analytic approach goes beyond this basis by using statistically sound procedures to estimate the magnitude of the various error components and then estimate the true test validity by correcting for these error components While Schmidt, Hunter, and Pearlman (1981) enumerate seven possible sources of error in validity estimates, the focus of this paper will be on the three that traditionally account for most of the error variance: sampling bias, unreliability in the criterion measure, and restriction in range of the predictors.

Method

The predictor tests.

The predictors used in this research were all based upon the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is a cognitive test battery that is employed by all of the military services to make selection and classification decisions. It is composed of ten subtests General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Auto/Shop Information (AS), Mathematical Knowledge (MK), Mechanical Comprehension (MC), Electronics Information (EI), and Coding Speed (CS). These ten subtests are combined by the U.S. Army in various ways to form composites. The composition of these composites is shown in Table 1. The Armed Forces Qualification Test (AFQT) is used as a

Table 1

The Composition of the ASVAB Composites

Operational Army Composites

AFQT	= VE + AR + .5NO
Electronics (EL)	= AR + EI + MK + GS
Operators/Foods (OF)	= NO + VE + MC + AS
Surveillance/Communications (SC)	= NO + CS + VE + AS
Motor Maintenance (MM)	= NO + EI + MC + AS
Clerical (CL)	= NO + CS + VE
Skilled Technical (ST)	= VE + MK + MC + GS
Combat (CO)	= AR + CS + MC + AS
Field Artillery (FA)	= AR + CS + MC + MK
General Technical (GT)	= VE + AR
General Maintenance (GM)	= MK + EI + GS + AS

screen for selection into the Army, while the other composites are used to determine whether or not an enlistee is qualified for a job in one of nine broad aptitude areas.

Sample and criterion.

The examination of ASVAB test validities was based upon 11 jobs or military occupational specialties (MOS). The MOS included in the research are given in Table 2. The 11 MOS range

Table 2

MOS Included in the Research

MOS	Name	N
05G	Signal/Security Specialist	91
16P	Short Range Missile Crewman	101
16S	MANPADS Crewman	514
32D	Tech Controller	120
33S	Electronic Warfare Systems Repairer	103
61B	Watercraft Operator	92
61C	Watercraft Engineer	150
67Y	Attack Helicopter Repairer	137
68J	Attack Fire Control Repairer	128
71D	Legal Clerk	96
76P	Material Control/Accounting Specialist	613

from legal clerk to missile crewman and represent jobs in six of the nine aptitude areas.

The criterion that was used as the basis for the test validities was the end-of-training scores for soldiers in each of these MOS. This score was a written test or a combination of written tests that is used by the Army to determine whether or not the soldier has been adequately trained to perform in that MOS.

Analyses.

The sample validity coefficients were obtained using standard statistical techniques. The meta-analytic procedures that were used to correct for the experimental artifacts are described in detail in Hunter, Schmidt, and Jackson (1982), and will only be outlined here in the sequence in which they were employed. The first step was to calculate the weighted average of the validities. This calculation is given in equation 1. Where the r is a particular predictor MOS correlation and N

$$(1) \qquad \bar{r} = \sum (N_i r_i) / \sum N_i$$

is the number of soldiers in that MOS. The second step was to use equation 2 to estimate the variance across MOS again weighted

$$(2) \qquad S_r^2 = \sum (N_i (r_i - \bar{r})^2) / \sum N_i$$

by sample size. The observed variance of the distributions of correlations is in part a function of sampling error. The variance that could be attributed to sampling error was examined in order to estimate the degree which sampling error was contributing to the observed variance in validities. This calculation is provided by equation 3 when N/(N-1) approaches

$$(3) \qquad S_e^2 = K(1 - \bar{r})^2 / \sum N_i$$

one. Here K is the number of validity coefficients and N is the total number of soldiers in all of the MOS that were investigated. Finally the mean true validity can be estimated by equation 4 where s^2 is the sample variance, S^2 the population variance, G = s/S or the level of range restriction in the test scores, and cr the reliability of the criterion measure (in this

$$(4) \qquad \hat{\rho}_{xy} = \left(1/G (\bar{r}) / \sqrt{[(1/G)^2 - 1](\bar{r})^2 + 1} \right) / \sqrt{CR}$$

case assumed to be .8).

Results and Discussion

Uncorrected validities

The observed uncorrected validities are presented in Table 3 for the ten ASVAB and in Table 4 for the operational Army

Table 3

Uncorrected Validities
for ASVAB Subtest Scores

MOS	GS	AR	PC	WK	NO	CS	AS	MK	MC	EI
05G	.45	.50	.40	.43	.07	.23	.26	.33	.35	.41
16P	.20	.20	.02	.02	.10	.05	.29	.11	.09	.16
16S	.16	.17	.18	.18	-.01	.12	.20	.19	.19	.24
32D	.34	.42	.46	.40	.27	.30	.16	.35	.26	.19
33S	.39	.39	.29	.41	.21	.08	.42	.49	.46	.49
61B	.41	.40	.32	.44	-.02	.08	.38	.35	.39	.26
61C	.33	.46	.26	.21	.10	.14	.29	.52	.27	.37
67Y	.28	.26	.14	.09	.24	.11	.28	.19	.15	.20
68J	.26	.29	.24	.19	.15	.07	.34	.33	.44	.36
71D	.29	.31	.39	.41	.03	.00	.22	.30	.18	.29
76P	.20	.40	.25	.24	-.02	.11	.15	.42	.26	.19

Table 4

Uncorrected Validities
for Operational Army Composites

MOS	AFQT Uncorrected	Army Composite Uncorrected
05G	.55	(SC) .48
16P	.15	(OF) .21
16S	.17	(OF) .23
32D	.44	(EL) .43
33S	.46	(ST) .56
61B	.49	(MM) .45
61C	.45	(OF) .45
67Y	.29	(MM) .39
68J	.28	(EL) .44
71D	.38	(CL) .27
76P	.40	(CL) .26

composites. Based upon these two tables alone one would not place much faith in the ability of cognitive tests to predict Training performance. While some of the observed validities are large many are quite small. Furthermore, there is considerable variance among the validity estimates across MOS. This picture changes considerably, however, after one performs the meta-analytic calculations.

Meta-analysis corrections

The results of solving equations 1 through 4 are presented in Tables 5 and 6. Table 5 gives the results of the application to the data from the ASVAB subtests and Table 6 presents the findings for the Army composites. In both tables the results are integrated across the 11 MOS that were investigated.

Two major findings emerge from an examination of these tables. First, a large portion of the original observed variance in validity coefficients can be attributed to sampling error. In the case of the ASVAB subtests the average percentage of the observed variance attributable to sampling error is 58%, while 40% of the observed variance in composite validities could be credited to sampling bias. This finding underscores the importance of large Ns if one is to draw conclusions about validity coefficients.

The second major finding concerns the true validities of the cognitive tests as shown in the last column of both Tables 5 and 6. These validities are in fact quite high. The average estimated true validity for the ASVAB subtests was .56, and the Army composites showed an average estimated true validity of .65. Together these findings indicate that cognitive tests can be very accurate predictors of training success. These data also illustrate the value of combining the subtests into composites. The estimated true validites for the composites were about 18% higher than the true validities that were estimated from the subtests.

References

Hunter, J.E., Schmidt, F.L., & Jackson, G. Meta-Analysis: Cumulating Research Findings Across Studies. Beverly Hills, CA: Sage Publications, 1982.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 1981, 66, 166-185

Table 5
Validity Generalization Results
ASVAB Subtests

Subtest	Observed Average r	Observed Variance in r	Percentage Variance Due to Sampling Error	Residual Variance in r	Estimated Mean True Validity
General Sciences	.2449	.0073	57.5	.0031	.6054
Arithmetic Reasoning	.3222	.0124	31.4	.0085	.6480
Word Knowledge	.2412	.0119	35.3	.0077	.5865
Paragraph Comprehension	.2433	.0095	45.3	.0052	.5659
Numerical Operations	.0725	.0079	62.5	.0030	.4863
Auto/Shop Information	.2235	.0059	73.6	.0016	.4875
Mathematical Knowledge	.3205	.0143	28.0	.0103	.6320
Mechanical Comprehension	.2485	.0079	54.2	.0036	.5623
Electronic Information	.2500	.0064	65.6	.0022	.5829
Coding Speed	.1191	.0039	125.9	-.0114	.4366

Table 6
Validity Generalization Results
Army Composites

	Observed Average r	Observed Variance in r	Percentage Variance Due to Sampling Error	Residual Variance in r	Estimated Mean True Validity
Composites					
AFQT	.3324	.0155	24.0	.0118	.6694
Clerical	.2238	.0066	67.7	.0021	.5848
Motor Maintenance	.3214	.0111	33.1	.0074	.6490
Operators/ Foods	.3248	.0100	36.9	.0063	.6637
Electronics	.3770	.0116	29.7	.0081	.6968
Surveillance/ Communications	.2903	.0065	61.8	.0025	.6403
Skilled Technical	.3463	.0130	27.8	.0094	.6877

Adjustments for the Effects of Range Restriction on Composite Validity

David Brandt
Donald H. McLaughlin
Lauress L. Wise
American Institutes for Research

Paul G. Rossmeissl
Army Research Institute

August 1984

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

Paper presented at the Annual Convention of the American Psychological Association at Toronto, Canada.

This paper discusses the adjusted validities of the current ASVAB composites. This work was part of a larger effort to investigate the validity of the Armed Forces Vocational Aptitude Battery (ASVAB) and to identify a set of "optimal" composites for the selection and classification of recruits into the Army. This paper reports only on the validities of the set of nine composites currently in use by the Army.

The ASVAB is a written test battery that includes ten subtests. These subtests are:

- Paragraph Comprehension (PC)
- Word Knowledge (WK)
- General Science (GS)
- Arithmetic Reasoning (AR)
- Mathematics Knowledge (MK)
- Mechanical Comprehension (MC)
- Auto Shop Information (AS)
- Electronics Information (EI)
- Coding Speed (CS)
- Numerical Operations (NO)

Paragraph Comprehension and Word Knowledge are combined into a single measure called Verbal Skills (VE).

The military currently uses a combination of four of these subtests as an overall determinant of qualification for military service. This composite, known as the Armed Forces Qualifying Test (AFQT), consists of VE, AR, and NO, with units weights given to VE and AR and a weight of one-half to NO. For classification, each service uses additional ASVAB composites to qualify recruits for particular military occupational specialities (MOS). For nearly all of the entry-level Army MOS, a soldier must have a passing score on one (or in a few cases, two) of nine ASVAB composites. This set of nine was developed in a previous analysis of the relationship of ASVAB scores to criterion measures (Maier & Grafton, 1981). They are:

• Clerical/Administrative	CL	VE+NO+CS
• Combat	CO	AR+CS+AS+MC
• Electronics Repair	EL	GS+AR+MK+EI
• Field Artillary	FA	AR+CS+MK+MC
• General Maintenance	GM	CS+AS+MK+EI
• Mechanical Maintenance	MM	NO+AS+MC+EI
• Operators/Food	OF	VE+NO+AS+AS
• Surveillance/Communications	SC	VE+NO+CS+AS
• Skilled Technical	ST	GS+VE+MK+MC

Each of these nine composites has been assigned to a particular subset of MOS. With the exception of a few unusual MOS (e.g. those comprising the Army Band) and non-entry level MOS (e.g., guidance counselor), this consitutes a complete partitioning of all Army MOS into nine clusters. To qualify for a particular MOS, a recruit must score above an established cutoff on the composite associated with the particular MOS cluster. These cutoff levels vary from MOS to MOS within cluster. They also fluctuate to some extent as a function of supply and demand.

This paper presents the adjusted validities of the nine ASVAB composites currently in operational use by the Army. Our discussion focuses on several interrelated issues:

- The levels of predictive validity, after adjustment;
- Differences among composites in ability to predict performance;
- Differences among MOS clusters of predictability of performance.

The predictive validity coefficients indicate the extent to which the composites cover the skills necessary to obtain proficiency in the corresponding MOS, as measured by training outcomes and SQT scores. Although the primary interest is in the composite associated with each particular MOS, there is also some interest in the entire matrix of composites by MOS, in order to address the question of whether a lower than average validity in a particular MOS is due to the nature of the composite or to the relation of the criterion in that MOS to the ASVAB in general. Differences among MOS in the overall relation of the ASVAB criterion can be interpreted as indicators of either (1) needs for greater criterion reliability or (2) areas in which the skills covered by the ASVAB need to be broadened.

METHOD

In any validation analysis it is essential to correct for the effects of selection. The validity coefficient is meant to be descriptive of the relationship between a predictor and a criterion measure in the population of job applicants. However, criterion data are only available on the subset of applicants who qualify for each job and retain it long enough to be tested. In general, this implies that the subset of cases for which complete data is available is a biased sample drawn from the applicant population. Presumably, the job incumbents are more able than those who were initially rejected. It is therefore necessary to adjust sample correlations for this bias. Because the problem of selection bias is fundamentally a "missing data" problem, some assumptions about the nature of the missing data must be made.

For the purposes of ASVAB validation, the multivariate adjustment due to Lawley (1943) and described by Lord and Novick (1968, pp. 146ff) was used. We assumed that explicit selection was being made on all ASVAB subtests. This methodology makes the key assumption that the multiple regression of the criterion on the subtests is the same in both the applicant and selected populations. More sophisticated ways of correcting for selection bias are available (Heckman, 1979), but these methods were designed for the simple case of selection rather than selection/classification. Little is known about the behavior of these methods in selection/classification research.

We chose to adjust to the FY81/82 applicant population rather than the 1980 Reference Population on the assumption that this population was more representative of the applicant pool presently available to the Army.

Summary statistics for the FY81/82 applicant population were obtained in the following way. The complete file of FY81/82 applicants was first constructed. Because this file included over 800,000 cases, summary statistics were not obtained on the whole file. Instead, a systematic sample was drawn by concatenating the FY81 and FY82 files. This resulted in a file of 19,027 cases. This was used as the basic population to which the validities would be adjusted. Table 1 gives the intercorrelations among the Army composites in this population.

Table 1

Intercorrelations among the Current Composites: Applicant Population

Composite	Composite								
CL	CL	CO	EL	FA	GM	MM	OF	SC	ST
CL	100								
CO	80	100							
EL	73	89	100						
FA	84	94	91	100					
GM	67	90	96	84	100				
MM	75	93	88	84	93	100			
OF	83	94	88	88	91	97	100		
SC	96	91	82	87	82	88	94	100	
ST	76	89	96	90	94	87	92	84	100

We report first on the validities using the so-called "combined" criterion. This is followed by a discussion of the validities of the composites using the Training and SQT criteria.

Results

Combined Criterion

Table 2 gives the adjusted validities for the nine current composites for each of the MOS clusters. The validities were obtained by averaging the validities for the individual MOS within each cluster and weighting by the number of soldiers in each MOS in the FY81/82 cohort. The main diagonal of Table 2 gives the validities of the composites associated with each cluster of MOS.

Table 2
Adjusted Validities of the Current Composites:
Combined Criteria

Cluster of MOS	N	Composite CL	CO	EL	FA	GM	MM	OF	SC	ST	Average
CL	10368	48	51	53	54	49	46	50	50	53	50
CO	14266	36	44	43	43	43	42	44	40	44	42
EL	5533	38	47	47	46	47	46	47	44	47	45
FA	5602	39	49	48	48	49	49	49	45	44	47
GM	2571	39	48	46	46	47	48	48	45	47	46
MM	7073	36	48	46	45	48	48	48	43	46	45
OF	8704	38	48	47	45	48	47	48	44	48	46
SC	3729	39	49	48	47	48	47	48	45	49	47
ST	7061	51	56	57	57	55	54	56	54	58	55
Average		40	49	48	48	48	47	49	46	48	47

The most striking feature of the data in Table 2 is the uniformity of the validities. All of the entries are between .36 and .58, and the mean of the validities for the set of operational composites is .47. Except for the CL composite, whose validites range from .36 to .51, the composites all perform about the same. In every instance, a given MOS cluster is predicted about as well from its own composite as from several of the others. One MOS cluster, ST, appears to be slightly more predictable than the others; and another cluster, CO, appears to be slightly less predictable. The remaining MOS clusters show very little variance.

Of the current composites, only CL consistently shows validities in the 30's. We will discuss the possible weaknesses of the CL composite and the speeded tests in the discussion section.

Training Data

Table 3 presents the average adjusted validities using the Training criterion. It is apparent from Table 3 that there is no great variation in the average effectiveness of the composites. Except for the clerical composite, which is slightly less predictive than the remaining composites, the average validities across all MOS clusters in the Army are within two points of each other.

Table 3

Average Adjusted Validities: Training Criterion

MOS Cluster	N	Composite CL	CO	EL	FA	GM	MM	OF	SC	ST	Average
CL	5272	40	43	45	46	42	39	42	42	45	43
CO	2879	30	36	33	35	33	34	35	34	34	34
EL	2610	35	42	40	41	39	40	41	39	40	40
FA	1759	27	37	34	35	35	37	36	32	33	34
GM	1944	42	52	51	50	52	52	52	49	50	50
MM	5426	33	44	42	41	44	44	44	40	42	42
OF	4626	28	35	34	33	35	34	35	33	35	34
SC	1463	33	35	35	36	33	32	34	34	35	34
ST	3181	46	52	53	51	52	50	53	51	54	51
Average		35	41	40	43	40	39	40	38	40	40

Performance in some MOS clusters is appreciably less well predicted than in others. The CO, OF, FA, and SC clusters are well below the overall mean. Each of these MOS clusters is composed of MOS that involve a substantial amount of physical and psychomotor skill. It may be that the reason that the ASVAB does relatively poorly for these MOS is that especially important predictors are absent from the battery. Regrettably, in each instance there is no composite that predicts that cluster relatively well.

SQT Criterion

Table 4 displays the weighted average adjusted validities of the clusters of MOS in the sample using the SQT as the criterion. As was the case with the training data, there is little variability among composites within a cluster. Except for the CL and SC composites, the average validities of the composites are within a couple of points of one another when collapsed across AA clusters. There is greater variability in the predictability of clusters, and the pattern is slightly different than for training data. GM, MM, and CO are most poorly predicted by the ASVAB, while CL and ST are best predicted. As before, the CL composite predicts performance in the CL cluster better than would be expected, but it must be remembered that the CL composite performs worst overall. Generally, composites that include GS, MK, AS, and MC tend to have higher validities than other composites.

Table 4

Average Adjusted Validities: SQT Criterion

MOS Cluster	N	Composite CL	CO	EL	FA	GM	MM	OF	SC	ST	Average
CL	8006	49	52	55	55	51	48	52	51	55	52
CO	15970	36	44	44	43	43	43	44	40	44	42
EL	5960	35	45	45	43	45	44	45	41	45	43
FA	6964	36	46	46	45	46	46	46	42	46	44
GM	1304	33	41	40	40	40	40	41	38	41	39
MM	4309	32	44	43	41	45	45	44	39	43	42
OF	4724	40	51	51	48	51	49	50	46	51	49
SC	3649	40	52	51	49	52	51	51	47	52	49
ST	6915	48	54	55	55	53	51	54	52	55	53
Average		39	48	48	47	47	46	47	44	48	46

Discussion

From these results a few general trends emerge. Among the composites, CL appears to be the least adequate. Alternative composites that included a quantitative component consistently did better for the MOS in which CL is operational. The FA composite, which includes both AR and MK, was consistently better than the CL composite for the Clerical MOS. Maier (1982) presents data that show that adding more mathematical content to CL does increase its validity. Our data are consistent with his findings.

The relatively weak performance of the CL composite observed here is also consistent with the findings of Sims and Hiatt (1981). Their adjusted validity coefficients for ASVAB 6/7 show the same pattern. Sims and Hiatt recommended groupings of MOS using a combination of empirical evidence and face validity. They recommended that CL be used in nine MOS included in their sample. In every instance, both the FA and the ST composites had the same or higher validities than CL. Thus, it seems clear, on the basis of training data, that some composite that includes a quantitative component will predict training success in a clerical MOS better than CL.

Our findings regarding the pattern of validities for the Clerical cluster are also consistent with the results of the the investigation of CL composite carried out by Weltin and Popelka (1983). They obtained

ASVAB and end-of-course grades for 3,984 new trainees entering the Army for clerical training in FY81 in twelve MOS. They evaluated the current CL composite (CL=VE+CS+NO) by comparing its adjusted validity with the adjusted validity of a revised composite suggested by a multiple regression of the ASVAB subtests on the training criterion. Results for the twelve MOS were quite similar in that all suggested that a quantitative subtest (either MK or AR) consistently accounted for the most variance in the criterion. Also, a revised composite consisting of unit weighted AR and VE predicted as well or better than the current composite in all twelve MOS. They reported that this composite correlated significantly higher than the operational composite. Thus, a clear message in both our assessment and the Weltin-Popelka analysis is that substituting a math subtest for the speeded subtests appreciably increases the validity of the CL composite.

Several authors have speculated on the poor performance of the CL composite. The major factors singled out by other workers are:

- The failure to adhere to uniform testing conditions has a greater effect on the speeded tests than the other tests in the ASVAB. When the timing of the test is not rigidly enforced, extra items can be marked by examinees. Weltin and Popelka also reported that examinees tested under military conditions have lower scores than those tested under civilian conditions.

- Scores on the speeded tests can be improved appreciably by practice. McCormick, Dunlap, Kennedy, and Jones (1982) found that when applicants were permitted to take the ASVAB repeatedly, scores of the speeded tests showed the greatest improvement. Thus, if these skills are relevant to job performance in MOS, it may be that they are sufficiently trainable that variance is removed by the time that criterion data are collected.

In general, the results of these analyses indicate that the current ASVAB area composites provide information relevant to the prediction of performance in training and on the job. Composites that included both speeded tests performed below average. There is little variability in validity coefficients within a given MOS cluster. However, they fall short of the ideal of targeting specific jobs for individuals. There is little evidence that these composites capture skills specific to the MOS with which they are associated.

References

Heckman, J. J. (1979) Sample selection bias as a specification error. Econometrica, 47 (1), 153-160.

Lawley, D. (1943). A note on Karl Pearson's selection formulae. Royal Society of Edinburgh, Proceedings, Section A. 62, 28-30.

Lord, P., & Novick, M. (1968). Statistical theory of mental test scores. Reading, MA: Addison-Wesley Publishing Company, Inc.

Maier, M. H., & Grafton, F. C. (1981). Aptitude composites for ASVAB 8, 9, and 10. (Research Rep. No. 1308). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Maier M. H. & Truss, A. R. (1983). Validity of ASVAB Forms 8, 9, and 10 for Marine Corps training courses: Subtests and current composites. (Center for Naval Analyses Memorandum No. 83-3107). Alexandria, VA: Marine Corps Operations Analysis Group.

McCormick, B., Dunlap, W., Kennedy, R., & Jones, M. (1982, November). The effects of practice on the Armed Services Vocational Aptitude Battery. Paper presented at the meeting of the Military Testing Association, San Antonio, TX.

Sims, W. H., & Hiatt, C. M. (1981). Validation of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7 with Applications to ASVAB Forms 8, 9, and 10 (Center for Naval Analyses Study No. 1160). Washington, DC: U.S. Department of the Navy.

Weltin, M. M. & Popelka, B. A. (1983, August). Evaluation of the ASVAB 8/9/10 clerical (CL) composite for predicting training performance. Paper presented at the annual convention of the American Psychological Association, Anaheim, CA.

Paper on the utility of research outcomes

The utility of any selection or classification effort is an important issue, and recent years have seen a considerable increase in interest in finding ways to measure or estimate utility. Several lines of inquiry are being followed under Project A to try to evaluate the utility of selection and classification improvement activities. An estimation technique developed by Schmidt, Hunter, McKenzie, and Muldrow (1979) was used to estimate the dollar value of the change the Army is making in the CL and SC composites at $5,000,000 per year.* Recent research by Eaton, Wing, and Mitchell (in press) provided an extension to the Schmidt et al. method and a new method, both of which appear to be more appropriate in military settings. Another utility effort, designed to evaluate the relative worth of various levels of performance within and between MOS is making substantial progress. In pilot efforts the 50th percentile infantryman is being used as a standard; Table 3 in Chapter I shows some of the first results.**

(1) A paper by Eaton, Wing, and Mitchell (scheduled for 1985 publication in Personnel Psychology) describes proposed alternate methods of estimating the dollar value of performance, and compares results obtained from these methods with those from currently used utility estimation techniques. An earlier version of this paper, "Putting the 'Dollars' Into Utility Analysis," appeared in the technical appendix to the 1983 annual report, ARI Research Note 83-37.

---

* Reported by P.G. Rossmeissl in ARI Research Highlights, June 1984.

** Reported by R. Sadacca and J.P. Campbell in a paper prepared for a briefing in October 1984.

Alternate Methods of Estimating the Dollar Value of Performance

Newell K. Eaton, Hilda Wing, and Karen J. Mitchell

Army Research Institute

(Scheduled for publication in Personnel Psychology)

This paper describes research performed under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This nine-year, large scale program is designed to provide the information and procedures required to meet the military manpower challenge of the future by enabling the Army to enlist, allocate and retain the most qualified soldiers. The research is funded primarily by Army Project Number 2Q263731A792 and is being conducted under the direction of the U.S. Army Research Institute for the Behavioral and Social Sciences. Research scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences, the Human Resources Research Organization, the American Institutes for Research, and the Personnel Decisions Research Institute as well as many Army officers and enlisted personnel are participating in this landmark effort.

ABSTRACT

The standard deviation of performance quality measured in dollars, SD$, is critical to calculating the utility of personnel decisions. In one popular technique for obtaining SD$, supervisors estimate the dollar value of performance at different levels. In many cases supervisors can base estimates on the cost of contracting out the various levels of performance. Estimation problems can arise, however, where contracting out is not possible, as in government organizations without private industry counterparts, or where individual salary is only a small percentage of the value of the performance to the organization or of the equipment operated. This paper presents two strategies ("superior equivalents" and "system effectiveness") for estimating the value of performance and determining SD$ by considering the changes in the numbers and performance levels of system units that lead to improved performance. One hundred Army tank commanders provided data about their jobs for these two strategies, as well as for the currently used "supervisor estimation" and "salary percentage" strategies. The new strategies appear to provide more appropriate and acceptable values of SD$ for those complex, expensive systems where dollar values of performance are less easily estimated.

Alternate Methods of Estimating the Dollar Value of Performance

Testing to improve selection and classification decisions is a normal part of entering into employment in most large organizations and in many small ones (Friedman & Williams, 1982). Cost-benefit analyses of selection and classification procedures have been difficult to conduct, however. Implementation costs are usually couched in real dollar terms. Easily estimated costs are associated with salaries, space, overhead for test administration, fees, per diem paid to applicants, computer time and personnel for scoring, etc. Benefits accrued from implementation of personnel policies, however, are not as clearly identifiable in dollar terms. Judgments of the net positive impact of implementing given personnel policies are, therefore, difficult to make.

Brogden (1949) and Cronbach and Gleser (1965) provided the first systematic descriptions of the utility of testing programs indexed in dollars. They linked normally distributed performance levels to the dollar values estimated for those performance levels. Their formula for the gain in productivity, or utility (U$), obtained by using valid selection procedures includes (a) Ns, the number of individuals selected; (b) SD$, the standard deviation of performance, scaled in a utility metric such as dollars; and (c) the average performance expected on the criterion by the selected group as estimated from a valid predictor, given by Rxy Zx:

$$U\$ = Ns\ SD\$\ Rxy\ Zx. \qquad (1)$$

The formula was subsequently modified to account for testing costs. A more complete description of such formulations can be found in Cronbach and Gleser (1965), Hunter and Schmidt (1982), and Cascio (1982).

While the values of most of the variables on the right hand side of the Brogden-Cronbach-Gleser formulas are known, the estimation of SD$, the standard deviation of performance scaled in dollars, is problematic. A recent review by Hunter and Schmidt (1982) reports that only two published efforts have attempted the computation of SD$ using cost accounting methods.

An alternative to the cost accounting methods is to estimate the dollar values to the organization of performance at the 50th percentile level, the 85th percentile level (one standard deviation above the mean), and, sometimes, the 15th percentile level (one standard deviation below the mean). The dollar difference between 15% and 50%, and 50% and 85%, provides an estimate of SD$. This "SD$ Estimation Technique" was used by Cascio and Silbey (1979) with second level managers in food and beverage sales (Mean = $30,000, SD$ = $9,500); by Schmidt, Hunter, McKenzie, and Muldrow (1979) with computer programmers (SD$ = $10,413); by Hunter and Schmidt (1982) with budget analysts (SD$ = $11,327); by Bobko, Karren, and Parkington (1983) with insurance counselors (Mean = $16,000, SD$ = $5,550); and by Burke and Frederick (1984) with district sales managers (Mean = 75,000, SD$ = $32,284). In Bobko et al. supervisors were also asked to estimate yearly sales volume for 15%, 50%, 85%, and 97% employees. Actual sales data were also available and yielded sales-based statistics which were 91%, 22%, 15% and 25% above the estimated

sales values. Burke and Frederick provided actual salary data for their district sales managers (Mean = $30,900 and SD = $4,600), and actual sales volume (Mean = $6,020,000 and SD = $2,634,000). The average estimated worth was approximately 2.4 times average salary while the SD$ estimate slightly exceeded average salary.

The SD$ estimations reported above were derived in contexts where performance was measurable in dollars. The SD$ estimation questions developed by Hunter and Schmidt (1982) asked for estimates of the "value to the agency" of various performance levels. The questions were preceded by instructions to "consider the cost of having an outside firm provide these products and services" (Schmidt et al., 1979, for computer programmers), or to "consider what the cost would be of having an outside consulting firm produce these products and services" (Hunter & Schmidt, 1982, for budget analysts). Both Cascio and Silbey (1979) and Bobko et al. (1983) framed questions in terms of performance value and estimates of total yearly dollar sales.

Another estimation strategy has been proposed by Hunter and Schmidt (1982). In reviewing the results of a variety of studies, they note that SD$ typically falls between 40% and 70% of annual salary. They (Hunter & Schmidt, 1983) suggest that 40% of mean salary may be a quick, inexpensive, albeit approximate, estimate of SD$. This might be termed the "Salary Percentage Technique."

It occurred to us that there may be situations where SD$ Estimation and Salary Percentage estimates would be impractical, if not misleading. These could occur where the nature of the work is such that managers are more

accustomed to considering the relative productivity of employees or crews than the costs of producing given levels of output. These could also occur where employees operate very complex, expensive equipment and/or are focal to the productivity of a costly system.

Two methods came to mind which might better serve in such circumstances. The first is somewhat like the SD$ Estimation Technique. Instead of using estimates of the dollar value of 85th percentile performance, however, the technique uses estimates of how many superior (85th percentile) performers would be needed to produce the output of a fixed number of average (50th percentile) performers. This estimate, combined with an estimate of the dollar value of average performance, provides an estimate of SD$, and is the basis for the name "Superior Equivalents Technique".

The second technique is an extension of the Brogdon-Cronbach-Gleser formula and is based on changes in aggregate system performance. In a system comprised of many units, total aggregate performance may be improved by increasing the number of units or improving the performance of each unit. The value of any aggregate performance improvement due to increased performance of a fixed number of units may be indexed by the cost of the increased number of units required to yield comparable increases in aggregate system performance. We called this the "System Effectiveness Technique."

The purpose of this research was to develop the "Superior Equivalents Technique" and the "System Effectiveness Technique," and to apply both of these as well as the SD$ Estimation and Salary Percentage Techniques, to an

existing system. We chose a system where the equipment is complex and expensive, where contracting-out for employee services is impossible, and where operators and supervisors are far more accustomed to thinking about the value of performance levels in terms of operational output rather than dollar value. This system is comprised of the tanks and their crews in the U.S. Army.

Method

First, the Superior Equivalents and System Effectiveness Techniques were developed to the extent that they could be applied to tank units. Second, items to obtain estimates required by the Superior Equivalents and the more conventional SD$ Estimation Techniques were combined into one questionnaire and administered to two groups of tank commanders (TCs). Additional data required for the various techniques were obtained as described below.

Superior Equivalents Technique Development

The Superior Equivalents Technique is conceptually similar to the conventional SD$ Estimation Technique which estimates SD$ from the difference in dollar estimates of the value of 85th and 50th percentile performance. The basic concept of the Superior Equivalents Technique is that of estimating the standard deviation of performance in performance units, and then converting the estimate to dollar units. We assume that supervisors are accustomed to evaluating the relative performance of their employees, and can make accurate judgments in these terms. Accordingly, they will have little trouble estimating the number (N85) of 85th percentile employees required to equal the performance of some fixed number (N50) of average performers. Where the value

of average performance (V50) is known, or can be estimated, SD$ may be estimated by using the ratio of N50/N85 times V50 to obtain V85, and then subtracting V50. This reduces to:

$$SD\$ = V85 - V50,$$

but

$$V85 = \frac{(V50)\ (N50)}{N85}.$$

Hence,

$$SD\$ = V50 \left[\frac{N50}{N85} - 1 \right]. \qquad (2)$$

In our case we set N50 = 17 as a fixed number of tanks with average commanders, because there are 17 tanks in a tank company. We used two methods for estimating V50. First, we used an estimate based on the salary and benefits paid the average tank commander. In general, one might assume organizations pay average employees about what they are worth. Second, we used an item on our questionnaire asking the dollar value of average performance.

System Effectiveness Technique Development

The basic concept of the System Effectiveness Technique is a system comprised of performing units which all contribute to total aggregate performance. That aggregate performance is a function of the number of units (employees/machines) and the performance of the units. Improved total aggregate performance may be obtained through improved unit performance with existing

numbers of units, or by increased numbers of units with the existing level of performance. Consequently, the value of improved unit performance in obtaining higher aggregate performance is equal to the cost of the increased number of units needed to obtain that same higher aggregate performance. When the Brogdon-Cronbach-Glaser formula is recast in these terms, the resulting extension provides a utility formula in which SD$ is replaced by more readily obtained cost and performance terms. Our derivations follow.

Let the cost of a single unit in a system be Cu. Let the total aggregate performance of the system, expressed in Y units, be TY. This may be achieved with varying numbers of units depending on the performance of the units. Or

$$TY = n1\ Y1 = n2\ Y2 = \ldots = ni\ Yi, \tag{3}$$

where ni = number of units at performance level i, on a ratio scale, and Yi = mean performance of units at level i, on a ratio scale.

Examples of performance scales useable in this formula are probability of hits per firing (Army tank commander), number of convictions per year (detective), number of pupils achieving a given standard (teacher), or other frequency-type variables. In a system where improved mean performance Y2 is obtained from the initial n1 units, the overall improvement in system aggregate performance is

$$n1\ Y2 - n1\ Y1 = n1\ (Y2 - Y1). \tag{4}$$

The number of extra units (Δn) operating at the initial performance level which are needed to achieve this improved performance is

$$\Delta n = \frac{n1\ (Y2 - Y1)}{Y1}. \tag{5}$$

The dollar value of improved performance is equivalent to the extra number of lower performing units needed times the cost per unit (Cu):

$$U\$ = \frac{Cu\ n1\ (Y2 - Y1)}{Y1}. \tag{6}$$

Simply stated, the value in dollars of achieving improved aggregate system performance equals the cost of adding the number of units required to effect the improvement where those added units operate at the initial performance level.

Estimating U$ using SD in performance units. The basic Brogden-Cronbach-Glaser formula (1) works in any metric; U and SD need not be expressed in dollars. The overall improvement in aggregate system performance is

$$U = Ns\ SDy\ Rxy\ Zx. \tag{7}$$

In output units of performance, Y, from formula (4), this equals

$$n1\ (Y2 - Y1) = Ns\ SDy\ Rxy\ Zx. \tag{8}$$

Substituting into (6):

$$U\$ = \frac{Cu\ Ns\ SDy\ Rxy\ Zx}{Y1}. \tag{9}$$

Formula (9) more conveniently describes the utility in dollars of selection. This formula uses SDy, the standard deviation in output units of performance, rather than SD$, the standard deviation of performance in dollars. Either SDy and Y1, or the ratio of SDy/Y1, may be estimated easily from empirical data. Cu is also often readily available.

Estimating SD$ using cost and performance data. Setting (1) and (9) equal

$$U\$ = Ns\ SD\$\ Rxy\ Zx = \frac{Cu\ Ns\ SDy\ Rxy\ Zx}{Y1},$$

and solving for SD$ yields

$$SD\$ = \frac{Cu\ SDy}{Y1}. \tag{10}$$

Or, SD$ equals the cost per unit times the ratio of the SDy of performance to the initial mean level of performance, Y1. It is interesting to note that this parallels the Hunter and Schmidt (1982) Salary Percentage Technique where SD$ may be linked to some percentage of salary. Here, Cu is the cost of the unit in the system; it includes equipment, support, and personnel, rather than salary alone. One might note that estimates from both (9) and (10) are appropriate only when the performance of the unit in the system is largely a function of the performance of the individual in the job under investigation. To the extent that it is not, corrections to these formulae would be required.

Instrument

A questionnaire was developed in the general form used by Schmidt et al. (1979), Bobko et al. (1983), and Burke and Fredrick (1984). It was used to obtain estimates of the dollar value of average and superior tank commander performance, and the number of tanks with superior tank commanders needed to equal the performance of a standard company of 17 tanks with average commanders. Dollar value and number-of-tank items were fill-in-the-blank. The item on number of tanks is shown below:

> For the purpose of this questionnaire an "average" tank commander is an NCO or commissioned officer whose performance is better than about half his fellow TC's. A "superior" tank commander is one whose performance is better than 85% of his fellow tank commanders.
>
> The first question deals with relative value. For example, if a "superior" clerk types 10 letters a day and an "average" clerk types 5 letters a day then, all else being equal, 5 "superior" clerks have the same value in an office as 10 "average" clerks.

In the same way, we want to know your estimate or opinion of the relative value of "average" vs. "superior" tank commanders in combat.

1. I estimate that, all else being equal, _____ tanks with "superior" tank commanders would be about equal in combat to 17 tanks with "average" tank commanders.

Respondents

The questionnaire was administered to two groups of male TCs enrolled in advanced training at a Continental U.S. Army post. The median number of years experience as a tank crew member for the 53 respondents in Group One was nine, and ten years for the 47 respondents in Group Two. Such individuals serve as trainers or supervisors of new TCs.

Other Data

Both the Superior Equivalents Technique and the System Effectiveness Technique required information from sources other than the questionnaire. To obtain another value of average performance for the Superior Equivalents Technique, as well as the data required for the Salary Percentage Technique, we used published pay and allowance tables. In 1983 the base pay for Army enlisted personnel with ten years of service at the ranks expected for tank commanders ranged from $14,000 to $16,000. Non-taxable allowances for such items as housing, post exchange, recreation, and travel be[illegible]s could amount to more than $10,000 for the typical married tank commano[illegible] with dependents. An estimate of an equivalent civilian salary would be ab[illegible] $30,000 per year.

Data for the System Effectiveness Technique were obtained from technical reports of previous research and from an approximation of tank costs. Criterion-related validity research on tank crew performance (e.g., Eaton,

Dassemer, & Christiansen, 1981) suggested that meaningful values for the ratio SDy/Y1 range from .2 to .5. We selected the more conservative value of .2. This is consistent with the estimate provided by Schmidt et al. (1979). Tank costs, consisting of purchase costs, maintenance, and personnel, were estimated between $300,000 and $500,000 per year. We chose the more conservative $300,000 value.

To permit computation of U$ for purposes of example, it was necessary to identify a selection ratio and a selection procedure validity. Inspection of tank doctrine indicates that tank commanders can be chosen from tank drivers, gunners, or loaders. However, in practice only more senior crew members are considered. We chose .5 as most likely to reflect actual selection ratios. Although higher validities are often observed (Eaton, 1978; 1980), a conservative value of Rxy = .3 was chosen.

Data from the questionnaire and the information discussed above were assembled to provide the basic input to each of the four techniques. Then, for each technique SD$ was computed, and U$ was determined on an individual tank/tank commander basis, as well as for a system having 2,500 tanks with tank commanders.

Results

SD$ Estimation Technique

The estimates of value for both average and superior tank commanders were skewed and very broad for both groups. Twelve of the TCs did not provide

these estimates. Sample sizes for the average and superior dollar estimates were 48 for Group One and 40 for Group Two. For average TC value, Group One gave estimates of $17,000, $30,000, and $100,000, for the first, second, and third quartiles, while Group Two's values were $18,000, $35,000, and $100,000. Although the shapes of the distributions were similar, both suffered from considerable positive skewing. The distributions left one uncertain about the adequacy of measurement of central tendency.

The distribution of estimates of superior TC value for Group One were $30,000, $50,000, and $300,000, for the first, second, and third quartiles, and $35,000, $95,000, and $500,000, for Group Two. Again, the distributions were highly positively skewed, and indicated even less agreement on the value of a superior TC. The difference between median estimates of superior and average performance values provided approximate estimates of SD$ = $20,000 for Group One and $60,000 for Group Two.

For a selection ratio of .5, Zx = .8 (Hunter & Schmidt, 1982). Incorporating Rxy = .3, Zx = .8, and the Group One estimate of SD$, into (1), yielded U$ = $4,800 if one tank commander were selected, and U$ = $12,000,000, if 2,500 tank commanders were selected. With the Group Two SD$ estimate of $60,000, the U$ values are $14,400 and $36,000,000, respectively. These values for SD$ and U$ are shown in Table 1.

Table 1
Estimates of SD$ and Examples of Utility

	n	SD$[a]	U$ or utility[a] per tank (Ns = 1)	U$ or utility[b] per system (Ns = 2,500)
SD$ Estimation Technique				
Group 1	48	$20,000	$ 4,800	$12,000,000
Group 2	40	$60,000	$14,400	$36,000,000
Superior Equivalents Technique				
Using Pay and Allowance Estimates of V50				
Group 1	52	$26,700	$ 6,400	$16,000,000
Group 2	45	$26,700	$ 6,400	$16,000,000
Using SD$ Estimates of V50				
Group 1	52	$26,700	$ 6,400	$16,000,000
Group 2	45	$31,100	$ 7,500	$18,700,000
System Effectiveness Technique	—	$60,000	$14,400	$36,000,000
Salary Percentage Technique	—	$12,000	$ 2,900	$ 7,200,000

[a] Rounded to nearest hundred dollars.
[b] Rounded to nearest hundred thousand dollars.

Superior Equivalents Technique

The median response given for the number of superior TCs judged equivalent to 17 average TCs was 9, and the mode was 10, in both groups. Fifty-two respondents provided data for Group One and 45 responded in Group Two. The first, second, and third quartile responses were 6, 9, and 10 for Group One, and 8, 9, and 10 for Group Two. The response, "9," was judged to be a representative value of central tendency. The distribution is shown at Figure 1.

The salary of an average tank commander was estimated at $30,000, from evaluation of pay and allowances for soldiers of relevant rank and experience. Given 9 superior TCs judged equivalent to 17 average TCs by both groups and an average TC 'worth' of $30,000 per year, a superior TC would be valued at 17/9 times $30,000, or about $56,700. Thus, the estimated SD$ is $26,700. Values for U$ were obtained for one and 2,500 tank commanders selected, yielding values of $6,400 and $16,000,000, respectively. These values are also shown in Table 1. The median values provided for average TCs from the SD$ Estimation Technique ($30,000 and $35,000 for Groups One and Two, respectively) were also used as the basis of estimating SD$ and U$. These figures are shown in Table 1.

System Effectiveness Technique

The value of SD$ was computed directly from formula (10) using the values Cu = $300,000 and SDy/Y1 = .2. This yielded SD$ = $60,000. Values of U$ for

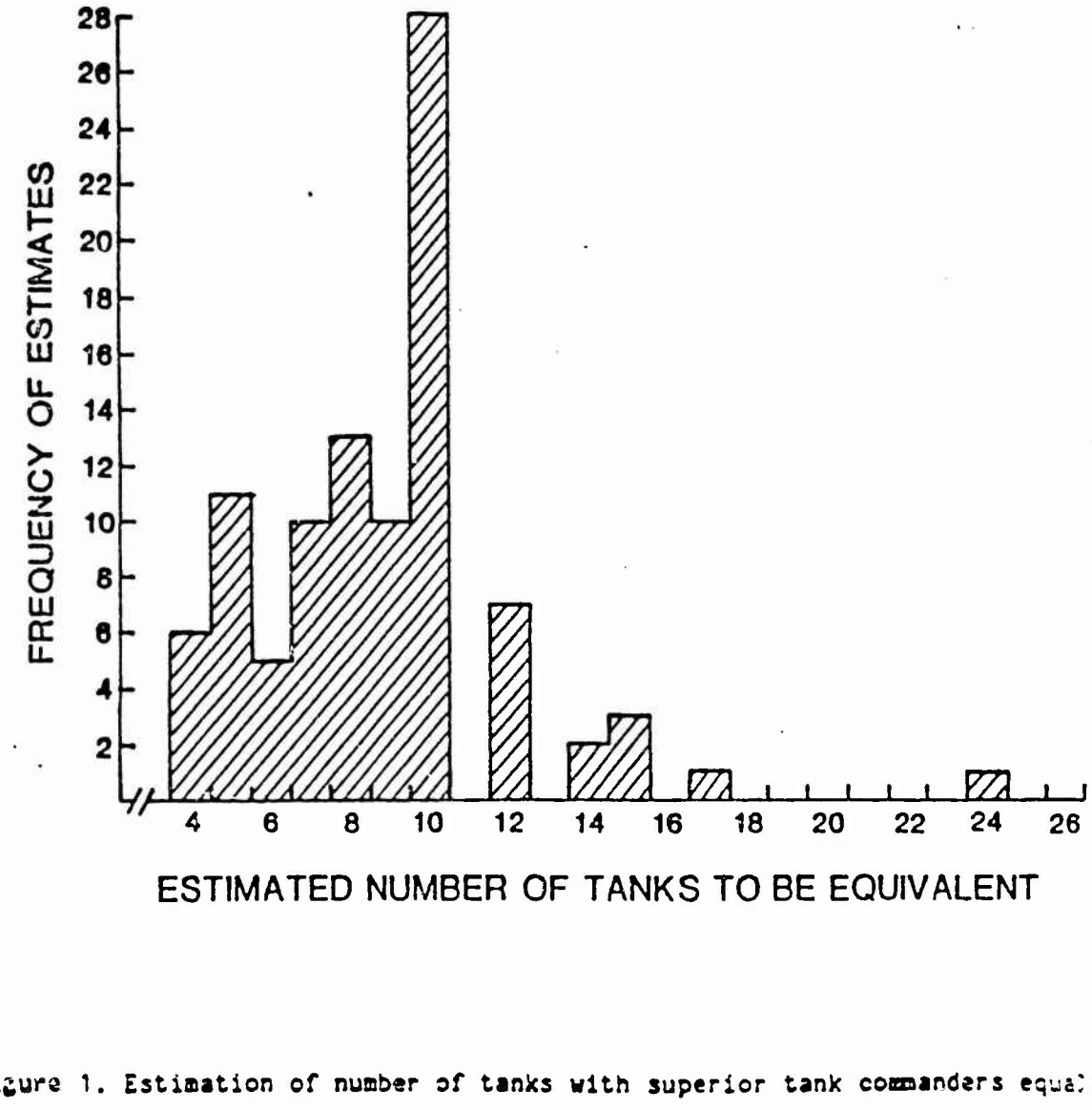


Figure 1. Estimation of number of tanks with superior tank commanders equaling 17 tanks with average tank commanders.

one and 2,500 tank commander selections were computed as $14,400 and $36,000,000, respectively. These are shown in Table 1.

Salary Percentage Technique

Finally, the value of SD$ as 40% of average annual salary ($30,000) led to U$ of about $2,900 for one TC and $7,200,000, for 2,500. These values are also shown in Table 1.

Discussion

Results from the now popular SD$ Estimation Technique were not judged satisfactory in this research. Two groups of tank commanders were asked to provide dollar estimates of the value of average and superior performance. There was minimal agreement either within or between groups for estimates of superior performance and the distributions were positively skewed. Distributions of average performance value, while consistent between groups, were also positively skewed. Together, these defects made calculation of SD$ or U$ highly suspect using the SD$ Estimation Technique. We suggest that the extreme response variability demonstrates the difficulty of making such judgments when the cost of contracting work is unknown, equipment is expensive, and/or other financially intangible factors are involved. Such is frequently the case for public employees, particularly when private industry counterparts are nonexistent.

Similar results were obtained by Bobko et al. (1983) and Burke and Frederick (1984) in contexts where estimates would appear easier to make. Both Bobko et al. and Burke and Frederick obtained skewed distributions of

performance value. The occupations studied were, respectively, insurance counselor and district sales manager. Bobko et al. suggested, and Burke and Frederick successfully implemented, a consensual feedback procedure to reduce the variability in estimates.

The Salary Percentage Technique yielded values which were much smaller than those obtained by the other methods. This was also true in Burke and Frederick (1984) where their SD$ estimates exceeded annual salary. In both cases the differences may be due to the greater responsibility inherent in the jobs. These are jobs in which an incumbent can control far more than his/her own productivity; the role provides leverage for the productivity of subordinates.

The proposed Superior Equivalents Technique seemed to work well here. Both groups of TCs were able to provide consistent estimates of the number of superior performers needed to equal the aggregate performance of a fixed number of average performers. The restricted distribution of their estimates helps support their accuracy. The requisite dollar value of average performance, obtained from pay and allowance tables, matched closely the median estimate of the TCs for average performance. Consequently, this technique provided data sufficient to make believable estimates of minimum values of SD$ and U$.

The differential success of these two techniques, we believe, directly relates to the degree of familiarity TCs have in dealing with performance in the metric of individual output rather than dollar values. The Superior Equivalents Technique appears to be the method of choice in situations where

supervisors are more accustomed to dealing with performance in output terms, or in relative output of individuals, rather than in dollar terms. Indeed, 12 of the 100 TCs refused to provide dollar estimates for average and superior performers. Many stated that soldiers' lives and combat activities were not describable in dollar terms. In such situations, supervisors' judgments of relative performance seem more credible than their estimates of the dollar value of average and superior performance. The relatively large variance and inconsistency between groups in dollar value estimates compared to superior equivalents estimates supports this point.

Despite its apparent success, the Superior Equivalents Technique may provide underestimates. The values are about half the size of the estimates from the System Effectiveness Technique. In this research, the pay and allowance estimates and the TC estimates of the dollar value of average performance were obtained separately. But, of course, tank commanders know their remuneration and that of their subordinates. Respondents here may have judged the performance of an average tank commander to be worth what he is paid because the cost of contracting the work out in the civilian market could not be estimated or because values could not easily be assigned to the intangible job components. These factors did not seem to constrain the Burke and Frederick (1984) estimates to the same extent. They found mean estimated worth to be 2.4 times greater than mean salary. Such figures are consistent with the suggestion of Schmidt, Hunter, and Pearlman (1982) that average overall worth is about twice annual salary. Perhaps a better value for $\underline{V50}$ in our formula (2)

would have been twice annual salary, or about $60,000. If so, estimates of SD$ and U$ from the Superior Equivalents Technique would have been about the same as those we obtained with the Systems Effectiveness Technique.

The higher values for SD$ and U$ obtained from the System Effectiveness Technique are based on SD$/Y1 ratios from actual field data. Our unit cost estimates, while crude, may fairly accurately reflect reality. The strengths of this technique appear to be based on the availability and interpretability of required data. In many systems such performance data are available and well understood. Cost estimates can be, and frequently are, made in accounting departments. These estimates can be adjusted if they appear unreasonable. Such performance and cost figures are subject to open examination and interpretation to a far greater extent than are supervisory estimates of the dollar worth of various performance levels.

One could argue that improved performance of tank commanders, the basis for the Superior Equivalents Technique, has only a partial impact on improving tank performance, the basis of the System Effectiveness Technique. An empirical question is the size of the contribution of the tank commander to the crew and tank. We do not yet know the answer. However, both analytical and rational judgments, as well as the lore within the armor community, suggest that the performance of the tank is largely a function of the performance of the commander. It was the assumption of the great impact of the tank commander to tank performance that led to our initial thoughts on the System Effectiveness formulae derivations.

Together, these two techniques may be useful in providing estimates which bracket true utility values. The Superior Equivalents Technique may underestimate in cases where the value of average performance is underestimated by pay and allowances and supervisors have no comparable pay data for comparison. The Systems Effectiveness Technique may provide overestimates to the extent that performance of the unit varies as a function of the performance of more than the individual whose job is being studied.

One lingering concern, that applies to both the Superior Equivalents Technique and the System Effectiveness Technique, is that performance quality in some situations may not be easily linked to a unidimensional, quantitative scale. Prospective users should question whether and how qualitative variables and multidimensional constructs are being transformed into unitary quantitative indices. For example, a police department may decide that conviction of one murderer is equivalent to the conviction of, say, five burglars. A school principal may judge that 75% above-average reading scores for an average third grade class is equivalent to 40% above-average reading scores for a below-average third grade class. We suspect that managers do, in fact, develop informal algorithms to compare performance of different individuals, perhaps on different factors. This may be what the supervisors in this and other research have done when requested to estimate the value of performance in either dollar (SD$ Estimation Technique) or unit (Superior Equivalents Technique) terms. Which terms or dimensions are most meaningful and useful will depend on important characteristics of the job under investigation.

References

Bobko, P., Karren, R., & Parkington, J. J. (1983). The estimation of standard deviations in utility analyses: An empirical test. Journal of Applied Psychology, 68, 170-176.

Brogden, H. E. (1949). When testing pays off. Personnel Psychology, 2, 171-183.

Burke, M. J., & Frederick, J. T. (1984). Two modified procedures for estimating standard deviations in utility analyses. Journal of Applied Psychology, 69, 482-489.

Cascio, W. F. (1982). Costing human resources: The financial impact of behavior in organizations. Boston: Kent Publishing Co.

Cascio, W. F., & Silbey, V. (1979). Utility of the assessment center as a selection device. Journal of Applied Psychology, 64, 107-118.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions (Second edition). Urbana: University of Illinois Press.

Eaton, N. K. (1978). Predicting tank gunnery performance (Research Memorandum 78-6). Alexandria, VA: U.S. Army Research Institute. (NTIC-ADA 077955)

Eaton, N. K. (1980). Performance motivation in armor training. JSAS Catalogue of Selected Documents in Psychology, 10, 28.

Eaton, N. K., Bessemer, D. W., & Kristiansen, D. M. (1981). Tank crew position assignment. JSAS Catalogue of Selection Documents in Psychology, 11, 62-63.

Friedman, T., & Williams, E. B. (1982). Current use of tests for employment. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences and controversies. Part II: Documentation section. Washington, DC: National Academy Press, 99-169.

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E. A. Fleishman & M. D. Dunnette (Eds.), Human performance and productivity: Volume 1. Human capability assessment. Hillsdale, N.J.: Erlbaum.

Hunter, J. E., & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work-force productivity. American Psychologist, 78, 473-478.

Schmidt, F. L., Hunter, J. E., McKenzie, R., & Muldrow, T. (1979). The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 64, 609-626.

Schmidt, F. L., Hunter, J. E., Pearlman, K. (1982). Assessing the economic impact of personnel programs on work force productivity. Personnel Psychology, 35, 333-347.

Acknowledgments

A preliminary version of this research was presented at the convention of the American Psychological Association, Los Angeles, California, in August, 1983. Lawrence M. Hanser and Donald F. Haggard provided many helpful suggestions.

Requests for reprints should be sent to Newell K. Eaton, U.S. Army Research Institute, 5001 Eisenhower Avenue, Alexandria, VA 22333.

BLANK PAGE

V. STATUS AND FUTURE DIRECTIONS OF ARMY SELECTION AND CLASSIFICATION RESEARCH

John P. Campbell and Newell K. Eaton

In the first two years of operation, the Army's Project A has provided impressive examples of ways in which to address current research problems, social issues, and policy questions of interest to military selection and classification scientists and managers. Two years' research by 50 scientists on this project have produced many empirical findings and research designs that we hope will prove fruitful during the coming years of the project and highly applicable to future research and practice in human resource management.

The principal goal of the research being conducted in Project A is to significantly improve overall enlisted performance by means of more accurate selection and classification. Together, better predictor tests and performance assessment will substantially increase classification accuracy, which in turn will mean better performance by the Army in the field. Further, Project A research will develop a wide range of new measures of enlisted job performance and further explication of the meaning of job performance in the Army. Completion of the new system is also expected to reduce personnel costs significantly and provide the Army's personnel managers with a powerful tool for evaluation and control.

Overall, the system should improve the readiness of the Army, and the performance satisfaction and career opportunities of individual soldiers. We continue to believe that these gains will be achieved most efficiently through a single, integrated research and development effort. As to future trends, it seems likely that we will have a greater opportunity to make real contributions to the productivity of our military organizations in the coming decades than in any previous time in the history of selection and classification research. We now have a much improved research technology with which to address the multitude of questions surrounding the goal of placing the right individual in the right job, to benefit both the individual and the organization.

Criterion development during FY84 resulted in the following specific accomplishments:

(1) Construction of the initial versions of the largest and most comprehensive array of job performance criterion measures in the history of personnel selection/classification research.

(2) Revision and refinement of each measure through pilot testing.

(3) Development and pilot testing of training materials for raters and test administrators.

(4) Completion of a comprehensive field test of all criterion measures, which involved two days of testing for approximately 600 job incumbents in several locations in the continental United States and in Europe.

Consequently, we have the information necessary for making final revisions and for creating the final array of criterion measures that will be used in the concurrent validation of the FY83/84 cohort during the summer of 1985.

For predictor test development FY84 may have been the most important year of the project. It was the period during which the final decisions about what to measure were made, and the full array of tests was developed, including state-of-the-art computerized measures. More than 11,000 soldiers had completed the tests that comprised the Preliminary Battery. By the end of FY84, the Pilot Trial Battery had been developed to measure a carefully identified and prioritized set of predictor constructs. This battery had been subjected to an iteration process of item construction, initial pilot tryouts, and several revision phases that resulted in a 6.5-hour battery of tests painstakingly constructed to measure as complete an array of the most relevant variables as possible. Extensive pilot test data were then collected to provide information for further refinement of the Pilot Trial Battery, especially a reduction in length.

Ultimately this process will result in the Trial Battery that will be administered to more than 12,000 soldiers in Year 3 of the project. Taking into account the 11,000 soldiers tested with the Preliminary Battery, together these two selection test batteries probably constitute the most carefully scrutinized and broadest array of selection and classification tests ever used in selection and classification research.

Also in FY84, as a first step in its many-faceted effort to improve the Army's selection and classification system, Project A completed a large-scale examination of the validity of the Aptitude Area Composite tests used by the Army as standards for selecting and classifying enlisted personnel. On the basis of these data, the Army has decided to implement the proposed alternative composites for CL (clerical) and SC (Surveillance/Communications) MOS, effective 1 October 1984. It can be estimated that these changes could lead to improved CL and SC MOS performance worth $5 million per year to the Army.

Further comment is warranted about a number of special issues bearing on criterion development that have arisen in Project A. Some have been resolved and some are still under discussion. None have precise answers or are completely scientific in nature.

Scenario Effects. At several points in Project A, raters or SMEs are being asked to make judgments about such things as (a) the relative importance of specific job tasks to an MOS, (b) the relative importance of a knowledge test item for the objectives of a particular AIT program, (c) the degree of effective job performance reflected in a particular critical incident, (d) the job proficiency of a ratee on specific performance factors, and (e) the relative value (i.e., utility) of different job performance levels across MOS.

Preliminary results indicate that "scenario" effects on judgments of importance are significant for certain kinds of tasks within some MOS. In particular, for non-combat support MOS the common tasks become more important and the MOS-specific tasks somewhat less important under a conflict rather than peacetime scenario.

Since some context effects do exist, the resolution has been to select tasks and test items that accommodate the differences. The preliminary data suggest that this should be possible within the constraints imposed by the FY83/84 concurrent validation design.

Multi-Method Measurement. In virtually any research project, measuring the major variables by more than one method is very desirable. In Project A, MOS-specific task performance is being assessed by three different methods (i.e., ratings, hands-on tests, and knowledge tests). Since testing time is not unlimited, a relevant issue is whether, for the concurrent validation, multiple measures should be retained at the expense of breadth of coverage, or vice versa. The relevant analyses that will inform this decision are not yet available, but the prevailing strategy is to do everything feasible to preserve multiple measurement.

Weighting of Criterion Components. Several measures in the criterion array are made up of component scores in the form of individual rating scales, knowledge subtests, or performance on a complete but singular task, as in the hands-on measures. A general issue concerns whether such components (e.g., the 15 separate hands-on tasks) should be differentially weighted before being combined into a total score. The same question arises when the aim is to combine specific criterion measures (e.g., ratings, knowledge tests, hands-on tests) into an overall composite for test validation.

The strategy that Project A will pursue is to compare weighted vs. unweighted criterion composites and determine whether differential weighting produces an advantage. The issue is scheduled to be considered during FY85.

Criterion Differences Across MOS. In Project A's validation of predictor measures for each of 19 MOS, the extent to which the same array of criterion measures should be used for the criterion composite in each MOS is a relevant question. This issue is being addressed directly by the continuing effort in Project A to develop an overall model of the effective soldier. In its current form, the model specifies the same set of constructs, or basic performance factors, for each MOS. In general, this means that very much the same measures would be used across MOS; however, their relative weights could vary considerably depending on the results of the MOS-specific development work and the criterion importance judgments.

These issues include some of the most central problems in selection and classification research. Prospects appear to be good that efforts under way in Project A will make substantial contributions toward resolving these, and other, significant inquiries. Three factors support this view: the administrative efficiency of large and integrated programatic efforts; the comprehensive and interrelated consideration of all of the practical, social, legal, and policy questions directed toward making the optimal use of our soldiers; and the application of the most sophisticated technology available to explore a wide range of scientific problems that offer promising prospects for effective solutions.

REFERENCES

Eaton, Newell K. & Goer, Marvin H. (Eds.) Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Technical Appendix to the Annual Report. ARI Research Note 83-37, Army Research Institute. Alexandria, VA. October 1983.

Eaton, Newell K., Goer, Marvin H., Harris, James H., and Zook, Lola M. (Eds.). Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1984 Fiscal Year. ARI Technical Report 660, Army Research Institute. Alexandria, VA. October 1984.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute. Improving the Selection, Classification and Utilization of Army Enlisted Personnel. Project A: Research Plan. ARI Research Report 1332, Army Research Institute. Alexandria, VA. May 1983.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute. Improving the Selection, Classification and Utilization of Army Enlisted Personnel: Annual Report. ARI Research Report 1347, Army Research Institute. Alexandria, VA. October 1983.

Human Resources Research Organization, American Institute for Research, Personnel Decisions Research Institute, and Army Research Institute. Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Appendices to Annual Report, 1984 Fiscal Year. ARI Research Note 85-14, Army Research Institute. Alexandria, VA. (In press.)

Hunter, J.E. & Schmidt, F.L. "Fitting People to Jobs: The Impact of Personnel Selection on National Productivity." In M.D. Dunnette and E.A. Fleishman (Eds.), Human Performance and Productivity: Human Capability Assessment. Hillsdale, N.J.: Lawrence Erlbaum. 1982. 231-284.

Peterson, N.G. & Bownas, D.A. "Skill, Task Structure, and Performance Acquisition." In M.D. Dunnette & E.A. Fleishman (Eds.), Human Performance and Productivity: Human Capability Assessment. Hillsdale, N.J.: Lawrence Erlbaum. 1982.

Schmidt, F.L., Hunter, J.E., McKenzie, R., & Muldrow, T. "The Impact of Valid Selection Procedures on Workforce Productivity." Journal of Applied Psychology, 64, 609-626, 1979.

END
5-87