DTIC FILE COPY

Technical Report 746

Project A
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel

AD-A193 343

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1985 Fiscal Year

John P. Campbell, Editor
Human Resources Research Organization

Selection and Classification Technical Area
Manpower and Personnel Research Laboratory

DTIC
ELECTE
APR 04 1988
S D
E



U. S. Army
Research Institute for the Behavioral and Social Sciences
June 1987

Approved for public release; distribution unlimited.

88 4 1 093

REPORT DOCUMENTATION PAGE

A193 343

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS:
2a. SECURITY CLASSIFICATION AUTHORITY:	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE:	
4. PERFORMING ORGANIZATION REPORT NUMBER(S):	5. MONITORING ORGANIZATION REPORT NUMBER(S): ARI Technical Report 746
6a. NAME OF PERFORMING ORGANIZATION: Human Resources Research Organization; 6b. OFFICE SYMBOL (If applicable): HumRRO	7a. NAME OF MONITORING ORGANIZATION: U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS (City, State, and ZIP Code): 1100 South Washington Street, Alexandria, VA 22314-4499	7b. ADDRESS (City, State, and ZIP Code): 5001 Eisenhower Avenue, Alexandria, VA 22333-5600
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: -- ; 8b. OFFICE SYMBOL (If applicable):	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: MDA 903-82-C-0531
8c. ADDRESS (City, State, and ZIP Code): --	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. 63731A; PROJECT NO. 2Q263 731A792; TASK NO. 232.C.1.; WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification): Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1985 Fiscal Year

12. PERSONAL AUTHOR(S): John P. Campbell, Editor (HumRRO)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final Report	FROM Oct 1984 TO Sep 1985	June 1987	473

16. SUPPLEMENTARY NOTATION: Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel (Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, U.S. Army Research Institute)

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Army-wide measures, Criterion measures, First-tour evaluation, Personnel classification, Personnel selection, Predictor measures, ~~Project A~~, Rating scales, Soldier effectiveness.

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

This report describes the research performed during the third year (FY85) of Project A, the Army's long-term program to develop a complete personnel system for selecting and classifying all entry-level Army enlisted personnel; it also summarizes the developmental work done during the first 2 years of the project. During the third year a wide variety of criterion and predictor measures were pilot tested, refined, and field tested. These efforts resulted in the Trial Battery being used in the "Concurrent Validation" phase, in which testing was begun during FY85 with a sample of several thousand soldiers. This report is supplemented by an ARI Research Note (in preparation), which contains a number of technical papers prepared during the year on various aspects of the project.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL: Lawrence M. Hanser	22b. TELEPHONE (Include Area Code): (202) 274-8275; 22c. OFFICE SYMBOL: PERI-RS

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE: UNCLASSIFIED

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

Dr. Deirdre Knapp
Dr. Clint B. Walker
Dr. Leonard A. White
Dr. Michael G. Rumsey

DTIC
ELECTE
APR 04 1980
S D
E

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.

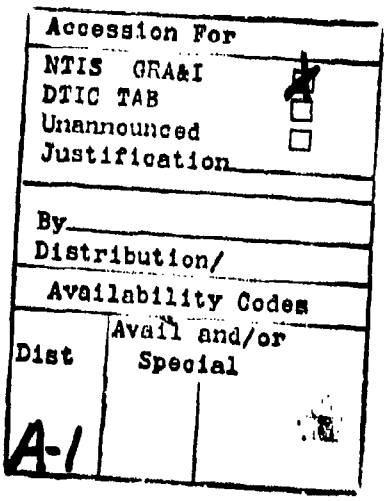
NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents

Technical Report 746

Project A
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1985 Fiscal Year

John P. Campbell, Editor
Human Resources Research Organization



Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, 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

June 1987

Army Project Number
2Q263731A792

Manpower and Personnel

Approved for public release; distribution unlimited.

This document describes the third year of research and summarizes earlier 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 Army Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training and performance.

Project A is being conducted under contract to the Selection and Classification Technical area (SCTA) of the Manpower and Personnel Research Laboratory (MPRL) at the U.S. Army Research Institute for the Behavioral and Social Sciences. 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." This research supports the MPRL and SCTA mission to improve the Army's capability to select and classify its applicants for enlistment or reenlistment by ensuring that fair and valid measures are developed for evaluating applicant potential based on expected job performance and utility to the Army.

Project A was authorized through a Letter, DCSOPS, "Army Research Project to Validate the Predictive Value of the Armed Services Vocational Aptitude Battery," effective 19 November 1980; and a Memorandum, Assistant Secretary of Defense (MRA&L), "Enlistment Standards," effective 11 September 1980.

In order to ensure that Project A research achieves its full scientific potential and will be maximally useful to the Army, a governance advisory group comprised of Army General Officers, Interservice Scientists, and experts in personnel measurement, selection, and classification was established. Members of the latter component provide guidance on technical aspects of the research, while general officer and interservice components oversee the entire research effort, provide military judgment, provide periodic reviews of research progress, results, and plans, and coordinate within their commands. Members of the General Officers' Advisory Group include MG Porter (DMPM) (Chair), MG Briggs (FORSCOM, DCSPER), MG Knudson (DCSOPS), BG Franks (USAREUR, ADCSOPS), and MG Edmonds (TRADOC, DCS-T). The General Officers' Advisory Group was briefed in May 1985 on the issue of obtaining proponent concurrence of the criterion measures prior to administration in the concurrent validation. Members of Project A's Scientific Advisory Group (SAG), who guide the technical quality of the research, include Drs. Milton Hakel (Chair), Philip Bobko, Thomas Cook, Lloyd Humphreys, Robert Linn, Mary Tenopyr, and Jay Uhlaner. The SAG was briefed in October 1984 on the results of the Batch A field test administration. Further, the SAG was briefed in March 1985 on the contents of the proposed Trial Battery.

A comprehensive set of new selection/classification tests and job performance/training criteria have been developed and field tested. Results from the Project A field tests and subsequent concurrent validation will be used to link enlistment standards to required job performance standards and to more accurately assign soldiers to Army jobs.



EDGAR M. JOHNSON
Technical Director

This Project A Annual Report for Fiscal Year 1985 has a different form than the reports for previous years. It is intended to be a comprehensive and reasonably detailed summary of the first 3 years of the Army Selection and Classification Project (Project A). The first 3 years are noteworthy because they encompass all the development work on the broad array of selection/classification tests and performance criteria upon which the concurrent and longitudinal validations will be based. Consequently, this report is meant to be an account of instrument development, from the conceptualization of the domains to be assessed to a description of the final revisions of the measures themselves.

Three years may seem like a long time to spend on development work but we hope that by the end of the report the reader will be convinced that Project A is of a different order of magnitude than most personnel research projects and that 3 years was a bare minimum. Future annual reports and contract deliverables will report on the validation results, estimates of classification efficiency, and results bearing on general issues in ability measurement and job performance assessment.

The bulk of this report consists of edited and abridged material from a series of field test reports produced by the project's individual research teams. Consequently, while the current volume includes considerable detail, an even more detailed account can be found in the individual field test reports, which are supplemented by extensive appendixes issued separately.

In general, the primary sources for the major sections were as follows:

- Part I is largely based on the Annual Reports for FY83 and FY84:

  Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, by Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute, ARI Research Report 1347, 1983. (AD A141 807)

  Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Technical Appendix to the Annual Report, Newell K. Eaton and Marvin H. Goer (Eds.), ARI Research Note 83-37, 1983. (AD A137 117)

  Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report Synopsis, 1984 Fiscal Year, by Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute, ARI Research Report 1393, 1984. (AD A173 824)

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 (Eds.), ARI Technical Report 660, 1984. (AD A178 944)

Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Appendices to Annual Report, 1984 Fiscal Year, by Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute, ARI Research Note 85-14, 1984.

- Part II is composed of edited and abridged material from the following predictor field test reports:

  Development and Field Test of the Trial Battery for Project A, Norman Peterson (Ed.), ARI Technical Report 739, 1987. Authors of individual chapters include Norman Peterson, Jody Toquam, Leaetta Hough, Janis Houston, Rodney Rosse, Jeffrey McHenry, Teresa Russell, VyVy Corpe, Matthew McGue, Bruce Barge, Marvin Dunnette, John Kamp, and Mary Ann Hanson. In preparation.

  Development and Field Test of the Trial Battery for Project A: Appendixes to ARI Technical Report 739, Norman Peterson (Ed.), ARI Research Note in preparation.

- Part III is primarily drawn from the series of field test reports dealing with criterion development:

  Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS, by Robert H. Davis, Gregory A. Davis, John N. Joyner, and Maria Veronica de Vera, ARI Technical Report 757, 1987. In preparation.

  Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS: Appendixes to ARI Technical Report 757, by Robert H. Davis, Gregory A. Davis, John N. Joyner, and Maria Veronica de Vera, ARI Research Note in preparation.

  Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program, Elaine D. Pulakos and Walter C. Borman (Eds.), ARI Technical Report 716, 1986. Authors of individual chapters include Walter C. Borman, Sharon R. Rose, and Elaine D. Pulakos. (AD B112 857)

  Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program: Appendixes to ARI Technical Report 716, Elaine D. Pulakos and Walter C. Borman (Eds.), ARI Research Note 87-22, 1987. In preparation.

Development and Field Test of Task-Based MOS-Specific Criterion Measures, by Charlotte H. Campbell, Roy C. Campbell, Michael G. Rumsey, and Dorothy C. Edwards, ARI Technical Report 717, 1986. In preparation.

Development and Field Test of Task-Based MOS-Specific Criterion Measures: Appendixes to ARI Technical Report 717, by Charlotte H. Campbell, Roy C. Campbell, Michael G. Rumsey, and Dorothy C. Edwards, ARI Research Note in preparation.

Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS, by Jody L. Toquam, Jeffrey J. McHenry, VyVy A. Corpe, Sharon R. Rose, Steven E. Lammlein, Edward Kemery, Walter C. Borman, Raymond Mendel, and Michael J. Bosshardt, ARI Technical Report in preparation.

Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS: Appendixes to ARI Technical Report, by Jody L. Toquam, Jeffrey J. McHenry, VyVy A. Corpe, Sharon R. Rose, Steven E. Lammlein, Edward Kemery, Walter C. Borman, Raymond Mendel, and Michael J. Bosshardt, ARI Research Note in preparation.

The Development of Administrative Measures as Indicators of Soldier Effectiveness, by Barry J. Riegelhaupt, Carolyn DeMeyer Harris, and Robert Sadacca, ARI Technical Report 754, 1987. In preparation.

The introduction to Part III is a creation of the editor. The description of the criterion field test procedures and the general summary of criterion field test results are edited versions of material from a 1985 paper:

Criterion Reduction and Combination via a Participative Decision-Making Panel, by John P. Campbell and James J. Harris, paper presented at the convention of the American Psychological Association, Los Angeles, 1985.

The descriptions of the development and testing of the combat performance prediction scales are based primarily on material supplied by Barry J. Riegelhaupt and Robert Sadacca.

- Part IV was assembled by the editor with assistance from James Harris and Laurie Wise. Much of the material comes from briefing materials developed by Harris, Wise, and the editor.

Various technical papers prepared during FY85 on specialized aspects of the Project A research are made available in a supplement to this report: Improving the Selection, Classification, and Utilization of Army

Enlisted Personnel: Annual Report, 1985 Fiscal Year--Supplement to ARI Technical Report 746, ARI Research Note in preparation. These papers are listed in Appendix A of the present report.

Additional editorial assistance was provided by Lola M. Zook. Barbara Hamilton cut, spliced, typed, and retyped many versions of the original manuscript.

John P. Campbell

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL: ANNUAL REPORT, 1985 FISCAL YEAR

EXECUTIVE SUMMARY

Requirement:

Project A is a comprehensive, long-range U.S. Army program to develop an improved personnel selection and classification system for enlisted personnel. The system encompasses 675,000 persons and several hundred military occupational specialties (MOS). The objectives are to (a) validate existing selection measures against both existing and project-developed criteria, and to develop new measures; and (b) validate early criteria (e.g., performance in training) as predictors of later criteria (e.g., job performance ratings), to improve reassignment and promotion decisions.

Procedure:

Under the sponsorship of the U.S. Army Research Institute (ARI), work on the 9-year project was begun in 1982. The research involves an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria).

In the first stage, file data from FY81/82 Army accessions were used to explore the relationships between the scores applicants made on the Armed Services Vocational Aptitude Battery (ASVAB) and their later performance in training and first-tour skill tests. The second stage is being executed with FY83/84 accessions; the 19 MOS in the sample were selected as representative of the Army's 250+ entry-level MOS and account for 45% of Army accessions. A preliminary battery of perceptual, spatial, temperament, interest, and biodata predictor measures was tested on several thousand soldiers as they entered four MOS; subsequent versions were pilot tested and field tested with nine MOS. The resulting predictor battery, along with a comprehensive set of job knowledge tests, hands-on job samples, and performance ratings, is being administered to 19 MOS. In the third stage, all of the measures, refined from experience, will be used to test about 50,000 soldiers across 19 MOS in the FY86/87 predictor battery administration and subsequent measurement of first-tour performance. About 3,500 are expected to be available for second-tour performance measurement in FY91.

Findings:

The wide variety of predictor and criterion measures under development were extensively field tested during FY84 and the first half of FY85, the

third year of effort in the project. These tests resulted in the Trial Battery's being used in the "Concurrent Validation" phase begun in FY85.

Utilization of Findings:

The full array of selection/classification measures of job and training performance from Project A is being utilized in current and long-range research programs expected to make the Army more effective in matching first-tour enlisted manpower requirements with available personnel resources.

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL: ANNUAL REPORT, 1985 FISCAL YEAR

CONTENTS

Part I
OVERVIEW OF PROJECT A AND SUMMARIES OF FISCAL YEAR 1983 AND FISCAL YEAR 1984 ACTIVITIES

Section		Page
1	**ORIGINS AND FORMULATION OF PROJECT A.**	I-2
	The Selection/Classification System for Army Enlisted Personnel	3
	Origins of the Project	5
	Specific Objectives of Project A	6
	Project A Organization	7
	The Research Plan and Integrated Master Plan	11
	Summary of Research Design and Sample Selection	11
2	**OVERVIEW OF FISCAL YEAR 1983.**	I-18
	Planning Activities	18
	Criterion Development	19
	Predictor Selection	20
	Data Base Management/Validation Analysis	20
	In Conclusion	21
3	**OVERVIEW OF FISCAL YEAR 1984.**	22
	Project Administration	22
	School and Job Performance Measurement	23
	Predictor Measurement	27
	Data Base Management/Validation Analyses	29
	In Conclusion	33
4	**FISCAL YEAR 1985.**	34
	Project Administration	34
	Research Activities	34
	Organization of the Report	36

Part II
PREDICTOR DEVELOPMENT

Section		Page
1	INTRODUCTION TO PREDICTOR DEVELOPMENT	II-2
	Objective	2
	General Research Design and Organization	3
	Literature Review	5
	Expert Forecasts of Predictor Construct Validities	8
	Development and Administration of the Preliminary Battery	14
	Initial Computer-Administered Battery Development	17
	Identification of Pilot Trial Battery Measures	23
	Pilot Tests and Field Tests of the Pilot Trial Battery	23
2	SUMMARY OF PILOT TESTS PROCEDURES	I-25
	Pilot Test 1: Fort Carson	25
	Pilot Test 2: Fort Campbell	26
	Pilot Test 3: Fort Lewis	26
	Summary of Pilot Tests	30
3	DEVELOPMENT OF COGNITIVE PAPER-AND-PENCIL MEASURES	31
	General Issues	31
	Construct - Spatial Visualization	32
	Construct - Field Independence	40
	Construct - Spatial Orientation	42
	Construct - Induction/Figural Reasoning	49
	Summary of Pilot Test Results for Cognitive Paper-and-Pencil Measures	53
4	DEVELOPMENT OF COMPUTER-ADMINISTERED TESTS	58
	Construct - Reaction Time (Processing Efficiency)	60
	Construct - Short-Term Memory	63
	Construct - Perceptual Speed and Accuracy	68
	Construct - Psychomotor Precision	76
	Construct - Multilimb Coordination	79
	Construct - Number Operations	82
	Construct - Movement Judgment	83
	Summary of Pilot Test Results for Computer-Administered Tests	84

Page

5 DEVELOPMENT OF NON-COGNITIVE MEASURES 89

Description of ABLE Constructs/Scales 93
ABLE Revisions Based on Pilot Test Results 97
Description of Interest (AVOICE) Constructs/Scales 102
AVOICE Revisions and Scale Statistics Based on Pilot Tests 109
Summary of Pilot Test Results for Non-Cognitive Measures 114

6 FIELD TESTS OF THE PILOT TRIAL BATTERY 115

Cognitive Paper-and-Pencil and Computer-Administered Field Tests 115
Field Test of Non-Cognitive Measures (ABLE and AVOICE) 128
Summary of Field Test Results 144

7 TRANSFORMING THE PILOT TRIAL BATTERY INTO THE TRIAL BATTERY . . 145

Changes to Cognitive Paper-and-Pencil Tests 145
Changes to Computer-Administered Tests 146
Changes to Non-Cognitive Measures (ABLE and AVOICE) 150
Description of the Trial Battery and Summary Comments . . . 152

Part III
CRITERION DEVELOPMENT

Section

1 INTRODUCTION TO CRITERION DEVELOPMENT III-2

Modeling Performance 2
Unit vs. Individual Performance 6
Plan for Part III 6

2 DEVELOPMENT OF MEASURES OF TRAINING SUCCESS 8

The Measurement Model 8
Test Development Procedure 11
Field Test Instruments 26

		Page
3	DEVELOPMENT OF TASK-BASED MOS-SPECIFIC CRITERION MEASURES	28
	Objectives	28
	Development Procedure	28
	Field Test Instruments	48
4	DEVELOPMENT OF MOS-SPECIFIC BEHAVIORALLY ANCHORED RATING SCALES (BARS)	50
	Development Procedure	51
	Field Test Versions of MOS-Specific BARS	61
5	DEVELOPMENT OF ARMY-WIDE RATING SCALES	62
	Development of Army-Wide Behavior Rating Scales	62
	Development of Army-Wide Common Task Dimensions	66
	Field Test Instruments	66
6	DEVELOPMENT OF THE COMBAT PERFORMANCE PREDICTION RATING SCALE	67
	Scale Development	68
	Field Test Version of Combat Effectiveness Prediction Scale	76
7	ADMINISTRATIVE/ARCHIVAL RECORDS AS ARMY-WIDE PERFORMANCE MEASURES	78
	Identification of Administrative Indexes	78
	Results of Analysis of MPRJ Data	86
	Criterion Field Test: Self-Reports of Administrative Actions	93
8	CRITERION FIELD TESTS: SAMPLE AND PROCEDURE	94
	The Sample	94
	The Criterion Measures	96
	Procedure	99
	Planned Analysis	102
	Interpretation and Use of the Field Test Results	103
9	FIELD TEST RESULTS: TRAINING ACHIEVEMENT TESTS	105
	Descriptive Statistics for Field Tests	105
	Revisions to Training Achievement Tests	105
	Some Lessons Learned	110
	Summary and Discussion	111

CONTENTS (Continued)

		Page
10	FIELD TEST RESULTS: TASK-BASED MOS-SPECIFIC CRITERION MEASURES	113
	Item/Scale Analyses	113
	Intercorrelations Among Task Performance Measures	121
	Revision of Task-Specific Performance Measures	127
	Proponent Agency Review	129
	Summary and Discussion	131
11	FIELD TEST RESULTS: MOS-SPECIFIC RATINGS (BARS)	132
	Rating Scale Adjustments	132
	Descriptive Statistics for MOS BARS Ratings	135
	Revision of the MOS-Specific BARS for Administration to the Concurrent Validation Sample	135
12	FIELD TEST RESULTS: ARMY-WIDE RATING MEASURES	151
	Statistics From Field Test	151
	Revision of the Army-Wide Scales	155
	Summary and Conclusions	165
13	RATER ORIENTATION AND TRAINING	167
	Components of Rater Training	167
	Batch B Rater Training Experiment	169
	Results	172
	Summary and Conclusions	174
14	FIELD TEST RESULTS: COMBAT PERFORMANCE PREDICTION SCALE	175
	Statistics From Field Test	175
	Revision of Scale for Concurrent Validation	175
15	FIELD TEST RESULTS: ARCHIVAL/ADMINISTRATION INDICATORS	184
	Results From Batch A	184
	Procedural Changes for Batch B	188
	Results From Batch B	191
	Revisions for Concurrent Validation	191
16	FIELD TEST RESULTS: CRITERION INTERRELATIONSHIPS	197
	Representative Criterion Intercorrelations	197
	True Score Relationships	207
	Summary	209
	A Plausible Model	209

Part IV
CONCURRENT VALIDATION

Section		Page
1	CONCURRENT VALIDATION: PREDICTOR AND CRITERION VARIABLES	IV-2
2	CONCURRENT VALIDATION: SAMPLES AND PROCEDURES	5
	Cross-Validation Sample	6
	Data Collection Team Composition and Training	9
	Data Collection Procedure	10
3	CONCURRENT VALIDATION: ANALYSIS PLAN	15
	General Steps	15
	Data Preparation	15
	Predictor Score Analyses	17
	Criterion Score Analyses	18
	Predictor/Criterion Interrelationships	19
	Criterion Composite Scores	21
	Performance Utility	22
EPILOGUE		1
REFERENCES		1
APPENDIX A. Project A FY85 Technical Papers		A-1

LIST OF TABLES

			Page
Table	I.1	Project A Military Occupational Specialties (MOS)	I-17
	I.2	FY83/84 Soldiers with Preliminary Battery and Training Data	I-30
	II.1	Tests of Pilot Trial Battery Administered at Fort Carson (17 April 1984)	II-27
	II.2	Pilot Tests Administered at Fort Campbell (16 May 1984)	II-28
	II.3	Pilot Tests Administered at Fort Lewis (11-15 June 1984)	II-29
	II.4	Summary of Pilot Testing Sessions for Pilot Trial Battery	II-30
	II.5	Cognitive Paper-and-Pencil Measures: Summary of Fort Lewis Pilot Test Results	II-53
	II.6	Intercorrelations Among the 10 Cognitive Paper-and-Pencil Measures: Pilot Test Data	II-55
	II.7	Rotated Orthogonal Factor Solution for Four Factors on Cognitive Paper-and-Pencil Measures: Pilot Test Data	II-56
	II.8	Reaction Time Test 1 (Simple Reaction Time): Fort Lewis Pilot Test	II-61
	II.9	Reaction Time Test 2 (Choice Reaction Time): Fort Lewis Pilot Test	II-64
	II.10	Memory Search Test: Fort Lewis Pilot Test	II-66
	II.11	Overall Characteristics of Perceptual Speed and Accuracy Test: Fort Lewis Pilot Test	II-69
	II.12	Scores from Perceptual Speed and Accuracy Test: Fort Lewis Pilot Test	II-71

		Page
II.13	Intercorrelations Among Perceptual Speed and Accuracy Test Scores: Fort Lewis Pilot Test . . .	II-71
II.14	Target Identification Test: Fort Lewis Pilot Test	II-75
II.15	Target Tracking Test 1: Fort Lewis Pilot Test . .	II-77
II.16	Target Shoot Test: Fort Lewis Pilot Test	II-80
II.17	Target Tracking Test 2: Fort Lewis Pilot Test . .	II-81
II.18	Means, Standard Deviations, and Split-Half Reliability Coefficients for 24 Computer Measure Scores Based on Fort Lewis Pilot Test Data	II-85
II.19	Intercorrelations of Dependent Measures Developed from Computer-Administered Tests: Fort Lewis Pilot Test	II-86
II.20	Intercorrelations of Cognitive Paper-and-Pencil Tests and Computer-Administered Tests: Fort Lewis Pilot Test	II-87
II.21	Summary of Criterion-Related Validities for Interest Inventories	II-90
II.22	Summary of Criterion-Related Validities for Biographical Inventories	II-90
II.23	Summary of Criterion-Related Validities for Temperament Constructs	II-91
II.24	Temperament/Biodata Scales (by Construct) Developed for Pilot Trial Battery: ABLE - Assessment of Background and Life Experiences . . .	II-93
II.25	ABLE Scale Statistics for Fort Campbell and Fort Lewis Pilot Samples	II-99
II.26	ABLE Scale Intercorrelations: Fort Campbell Pilot Test	II-100
II.27	Correlations Between ABLE Constructs and Scales and Personal Opinion Inventory (POI) Marker Variables: Fort Campbell Pilot Test	II-101

		Page
II.28	Varimax Rotated Principal Factor Analyses of 10 ABLE Scales: Fort Lewis Pilot Test	II-103
II.29	Holland Basic Interest Constructs, and Army Vocational Interest Career Examination (AVOICE) Scales Developed for Pilot Trial Battery	II-104
II.30	Additional AVOICE Measures: Organizational Climate/Environment and Expressed Interests Scales	II-105
II.31	AVOICE Scale Statistics for Total Group: Fort Lewis Pilot Test	II-110
II.32	AVOICE Means and Standard Deviations Separately for Males and Females: Fort Lewis Pilot Test . . .	II-112
II.33	Race and Gender of the Fort Knox Test Sample for the Pilot Trial Battery	II-117
II.34	Means, Standard Deviations, Reliability Estimates for the Fort Knox Field Test of the Ten Paper-and-Pencil Cognitive Tests	II-118
II.35	Gains on Pilot Test Battery for Persons Taking Tests at Both Time 1 and Time 2	II-119
II.36	Characteristics of the 19 Dependent Measures for Computer-Administered Tests: Fort Knox Field Tests	II-121
II.37	Effects of Practice on Selected Computer Test Scores	II-123
II.38	Intercorrelations Among the ASVAB Subtests and the Pilot Trial Battery Cognitive Paper-and-Pencil and Perceptual/Psychomotor Computer-Administered Tests: Fort Knox Sample	II-125
II.39	Mean Correlations, Standard Deviations, and Minimum Correlations Between Scores on ASVAB Subtests and Pilot Trial Battery Tests of Cognitive, Perceptual, and Psychomotor Abilities . . .	II-126

		Page
II.40	Principal Components Factor Analysis of Scores of the ASVAB Subtests, Cognitive Paper-and-Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer-Administered Tests	II-127
II.41	ABLE Scale Score Characteristics: Fort Knox Field Test	II-130
II.42	ABLE Test-Retest Results: Fort Knox Field Test . .	II-131
II.43	AVOICE Scale Score Characteristics: Fort Knox Field Test	II-132
II.44	ABLE Factor Analysis: Fort Knox Field Test	II-133
II.45	AVOICE Factor Analysis: Fort Knox Field Test . . .	II-134
II.46	Honesty and Faking Effects, ABLE Content Scales: Fort Bragg	II-137
II.47	Honesty and Faking Effects, ABLE Response Validity Scales: Fort Bragg	II-138
II.48	Effects of Regressing Out Response Validity Scales (Social Desirability and Poor Impression) on Faking Condition ABLE Content Scale Scores: Fort Bragg	II-139
II.49	Comparison of Fakability Results from Fort Bragg (Honest), Fort Knox, and MEPS (Recruits) ABLE Scales	II-141
II.50	Effects of Faking, AVOICE Combat Scales: Fort Bragg	II-142
II.51	Effects of Faking, AVOICE Combat Support Scales: Fort Bragg	II-143
II.52	Summary of Changes to Paper-and-Pencil Cognitive Measures in the Pilot Trial Battery . . .	II-146
II.53	Summary of Changes to Computer-Administered Measures in the Pilot Trial Battery	II-149

		Page
II.54	Summary of Changes to Pilot Trial Battery Versions of Assessment of Background and Life Experiences (ABLE) and Army Vocational Interest Career Examination (AVOICE)	II-151
II.55	Description of Measures in the Trial Battery . . .	II-152
III.1	Military Occupational Specialties (MOS) Included in Batches A, B, and Z	III-9
III.2	Number of Subject Matter Experts Participating in Training Achievement Test Reviews, and Locations of Reviews	III-15
III.3	Distribution of Soldiers in Four Race Categories, Army-Wide and Among Subject Matter Expert Reviewers for Training Achievement Tests . .	III-16
III.4	Mean Item Importance Ratings by Job Incumbents for Three Scenarios (Initial Item Pool for Training Achievement Tests)	III-18
III.5	Mean Item Importance Ratings by Trainers (Initial Item Pool for Training Achievement Tests)	III-20
III.6	Number and Percentage of Items Rated Relevant to Job and Training (Initial Item Pool for Training Achievement Tests)	III-22
III.7	Results from Training Achievement Tests Administered to Trainees	III-23
III.8	Mean Item Importance Ratings by Job Incumbents for Three Scenarios (Field Test Version of Training Achievement Tests)	III-25
III.9	Mean Item Importance Ratings by Trainers (Field Test Version of Training Achievement Tests)	III-26
III.10	Number and Percentage of Items Rated Relevant to Job and Training (Field Test Version of Training Achievement Tests)	III-27

CONTENTS (Continued)

		Page
III.11	Military Occupational Specialties (MOS) Selected for Criterion Test Development	III-29
III.12	Effects of Domain Definition on MOS Task Lists . .	III-33
III.13	Participants in MOS-Specific BARS Workshops	III-51
III.14	Locations and Dates of MOS-Specific BARS Workshops	III-52
III.15	BARS Performance Incident Workshops: Number of Participants and Incidents Generated by MOS and by Location - Batch A	III-54
III.16	BARS Performance Incident Workshops: Number of Participants and Incidents Generated by MOS and by Location - Batch B	III-55
III.17	BARS Retranslation Exercise: Number of Forms Developed for Each MOS and Average Number of Raters Completing Each Form	III-57
III.18	Behavioral Examples Reliably Retranslated into Each Dimension on the BARS Measures	III-58
III.19	Participants in Behavioral Analysis Workshops for Army-Wide Rating Scales	III-63
III.20	Soldier Effectiveness Examples Generated for Army-Wide Behavior Rating Scales	III-63
III.21	Behavioral Examples Reliably Retranslated into Each Dimension for Army-Wide Behavior Rating Scales	III-65
III.22	Combat Performance Workshop Participants and Examples Generated	III-71
III.23	Number of Edited Examples of Combat Behavior . . .	III-71
III.24	Agreement by Discriminability Item Distribution . .	III-73
III.25	Combat Prediction Agreement by Discriminability Item Distribution for Reduced Dimensional Set and Redefined Agreement Criteria	III-75

		Page
III.26	Items Selected for Field Test of Combat Performance Prediction Scale	III-75
III.27	Preliminary List of Administrative Measures Indicative of Soldier Effectiveness	III-79
III.28	Expanded List of Administrative Measures Indicative of Soldier Effectiveness	III-82
III.29	MOS x Post Populations in Study of Military Personnel Records Jackets	III-83
III.30	Number of Military Personnel Records Jackets Requested and Received at Each Post	III-83
III.31	List of Created Variables in Study of Administrative Measures	III-85
III.32	Frequency Distributions for Selected Variables in MPRJ-OMPF Comparison	III-87
III.33	Frequency Distributions for Selected Variables in MPRJ/EMF Comparison	III-88
III.34	Frequency and Percentage Distributions for Administrative Variables	III-89
III.35	Means, Standard Deviations, and Correlation Coefficients of Administrative Variables	III-91
III.36	Summary of Univariate and Multivariate Analyses of Administrative Variables	III-93
III.37	Field Test Sample Soldiers by MOS and Location . .	III-95
III.38	Field Test Sample Soldiers by Sex and Race	III-95
III.39	Results From Training Achievement Field Tests Administered to Incumbents	III-106
III.40	Number of Items in Training Achievement Tests at Each Stage of Development: Batch A	III-107
III.41	Number of Items in Training Achievement Tests at Each Stage of Development: Batch B	III-108

		Page
III.42	Number of Items in Training Achievement Tests at Each Stage of Development: Batch Z	III-109
III.43	Summary of Item Difficulties (Percent Passing) and Item-Total Correlations for Knowledge Test Components in Nine MOS	III-114
III.44	Means, Standard Deviations, and Split-Half Reliabilities for Knowledge Test Components for Nine MOS	III-115
III.45	Coefficient Alpha of Knowledge Tests Appearing in Multiple MOS	III-116
III.46	Means, Standard Deviations, and Split-Half Reliabilities for Hands-On Test Components for Nine MOS	III-118
III.47	Means, Standard Deviations, Number of Raters, and Interrater Reliabilities of Supervisor and Peer Ratings Across 15 Tasks for Nine MOS	III-120
III.48	Correlations Between Hands-On and Knowledge Test Components for MOS Classified by Type of Occupation	III-125
III.49	Reported Correlations Between Hands-On (Motor) and Knowledge Tests	III-126
III.50	Summary of Testing Mode Array for MOS Task Tests Before Proponent Review	III-128
III.51	Ratio of Raters to Ratees, Before and After Screening, for Supervisors and Peer Ratings on MOS-Specific BARS	III-134
III.52	Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS	III-136
III.53	Average Intercorrelations of MOS-Specific BARS for Supervisors, for Peers, and Between Supervisors and Peers	III-145

		Page
III.54	Summary of Reliability Estimates of MOS-Specific BARS for Supervisor and Peer Ratings	III-147
III.55	Summary of Grand Mean Values for Unadjusted and Adjusted BARS Ratings by MOS	III-148
III.56	Frequency Distributions (Percent) of Ratings Across the Seven Points of the Army-Wide Measures	III-152
III.57	Means and Standard Deviations of Selected Army-Wide Rating Measures	III-153
III.58	Intraclass Correlation Coefficients for Selected Army-Wide Rating Measures	III-154
III.59	Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS	III-156
III.60	True Score Matrix for Vignette Ratees on Six Army-Wide Dimensions	III-171
III.61	Interrater Reliabilities by Training Condition Across All MOS	III-173
III.62	Means and Standard Deviations for Rater-Ratee Pairs on the Combat Performance Prediction Scale	III-176
III.63	Intraclass Correlations for Estimating Reliabilities for the Combat Performance Prediction Scale	III-177
III.64	Coefficient Alpha for the Combat Performance Prediction Scale	III-178
III.65	Item Statistics Used in Selecting Combat Prediction Scale Items	III-179
III.66	Corrected Intraclass Correlations for Estimating Reliabilities of Best 40 Items on Combat Performance Prediction Scale	III-181
III.67	Comparison of Reenlistment Eligibility Information Obtained From Self-Report and 201 Files: Batch A	III-185

		Page
III.68	Comparison of Promotion Rate Information Obtained From Self-Report and 201 Files: Batch A	III-185
III.69	Comparison of Awards Information Obtained From Self-Report and 201 Files: Batch A	III-186
III.70	Comparison of Letters/Certificates Information Obtained From Self-Report and 201 Files: Batch A	III-186
III.71	Comparison of Articles 15/FLAG Information Obtained From Self-Report and 201 Files: Batch A	III-187
III.72	Comparison of Military Training Information Obtained From Self-Report and 201 Files: Batch A	III-187
III.73	Correlations Between Army-Wide Supervisor Ratings and Administrative Measures: Batch A	III-189
III.74	Correlations Between Army-Wide Peer Ratings and Administrative Measures: Batch A	III-190
III.75	Comparison of Letters/Certificates Information Obtained From Self-Report and 201 Files	III-192
III.76	Comparison of Awards Information Obtained From Self-Report and 201 Files: Batch B	III-192
III.77	Comparison of Article 15/FLAG Information Obtained From Self-Report and 201 Files: Batch B	III-193
III.78	Comparison of M16 Qualification Information Obtained From Self-Report and 201 Files: Batch B	III-193
III.79	Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS	III-198
III.80	Intercorrelation Among Selected Criterion Measures for Infantryman (MOS 11B)	III-208

		Page
III.81	Intercorrelations Among Selected Criterion Measures for Administrative Specialist (MOS 71L)	III-208
IV.1	Project A MOS Used in the Concurrent Validation Phase	IV-2
IV.2	Summary of Predictor Measures Used in Concurrent Validation: The Trial Battery	IV-3
IV.3	Summary of Criterion Measures Used in Batch A and Batch Z Concurrent Validation Samples	IV-4

LIST OF FIGURES

Figure	I.1	Project A Organization as of 30 September 1983	I-9
	I.2	Governance Advisory Group as of 30 September 1983	I-10
	I.3	The Overall Data Collection Plan	I-13
	I.4	Governance Advisory Group as of 30 September 1984	I-23
	I.5	Project A Organization as of 30 September 1984	I-24
	I.6	Predictive Validities System for Nine and Four Aptitude Area Composites	I-32
	I.7	A Comparison of Current and Alternative Aptitude Area Composites	I-33
	I.8	Project A Management Group as of 30 September 1985	I-35
	I.9	Project A Organization as of 30 September 1985	I-36
	II.1	Flow Chart of Predictor Measure Development Activities of Project A	II-4
	II.2	Factors Used to Evaluate Predictor Measures for the Preliminary Battery	II-7

CONTENTS (Continued)

		Page
II.3	Hierarchical Map of Predictor Space	II-13
II.4	Predictor Categories Discussed at IPR in March 1984, Linked to Pilot Trial Battery Test Names	II-24
II.5	Sample Items From Assembling Objects Test	II-34
II.6	Sample Items From Object Rotation Test	II-36
II.7	Sample Items From Path Test	II-38
II.8	Sample Items From Maze Test	II-40
II.9	Sample Items From Shapes Test	II-42
II.10	Sample Items From Orientation Test 1	II-44
II.11	Sample Items From Orientation Test 2	II-46
II.12	Sample Items From Orientation Test 3	II-48
II.13	Sample Items From Reasoning Test 1	II-50
II.14	Sample Items From Reasoning Test 2	II-52
II.15	Response Pedestal for Computerized Tests	II-59
II.16	Graphic Displays of Example Items From the Computer-Administered Target Identification Test	II-74
II.17	Linkages Between Literature Review, Expert Judgments, and Preliminary and Trial Battery on Non-Cognitive Measures	II-92
II.18	Organizational Climate/Environment Constructs, Scales Within Constructs, and an Item From Each Scale	II-108
III.1	Development Process For Tests of Achievement in Training	III-12

		Page
III.2	Alternative Scenarios Used for Judging Importance of Tasks and Items for Training Achievement Tests	III-17
III.3	Scenarios Used in SME Ratings of Task Importance for Task-Based MOS-Specific Tests	III-35
III.4	Infantryman (MOS 11B) Tasks Selected for Hands-On/Knowledge Testing	III-39
III.5	Administrative Specialist (MOS 71L) Hands-On Performance Test Sample	III-41
III.6	Infantryman (MOS 11B) Hands-On Performance Test Sample	III-43
III.7	Sample Behavioral Summary Rating Scale for Military Police (95B)	III-60
III.8	Preliminary Set of Combat Performance Dimensions	III-69
III.9	Revised Set of Combat Performance Dimensions	III-72
III.10	Sample of Combat Performance Prediction Scale	III-77
III.11	Typical Field Test Administration Schedule	III-101
III.12	Average Correlations Between Task Measurement Methods on Same Tasks and Different Tasks for Nine MOS	III-122
III.13	Reliabilities and Correlations Between Task Measurement Methods Across Tasks for Nine MOS	III-124
III.14	Sample Performance Rating Scale Before and After Modifications for Military Police (95B) MOS-Specific BARS	III-150
III.15	Sample Items From Combat Performance Prediction Rating Scale	III-182
III.16	Results of Outlier Comparison From Self-Report Information	III-194

		Page
III.17	Self-Report Form for Use in Concurrent Validation	III-195
III.18	Job Performance--A Proposed Structural Mode. . . .	III-211
IV.1	Concurrent Validation Schedule	IV-5
IV.2	Concurrent Validation Sample Soldiers by MOS by Location	IV-8
IV.3	Concurrent Validation Test Outline	IV-10
IV.4	Sample Schedule for Concurrent Validation Administration	IV-11
IV.5	Predictor Variable/Criterion Factor Matrix	IV-20
IV.6	Analysis Plan for Predictor Variable x Criterion Factor Matrix	IV-20

PART I

OVERVIEW OF PROJECT A AND SUMMARIES OF FISCAL YEAR 1983 AND FISCAL YEAR 1984 ACTIVITIES

Part I describes the origins and objectives of Project A, the project's organizational structure, and the overall design of the research. The central activities and accomplishments of the first 2 years are then summarized, along with the plan for integrating these materials with the description of the project's third year in the remainder of this report.

Section 1

ORIGINS AND FORMULATION OF PROJECT A[1]

"Project A" (Improving the Selection, Classification, and Utilization of Army Enlisted Personnel) is perhaps the largest personnel research and development project ever undertaken. Its general purpose is to develop an improved selection/classification system for all entry-level positions in an organization that annually recruits 400,000-500,000 people, selects 100,000-120,000 of them, and assigns each individual to 1 of more than 250 job classifications. The full design for Project A covers a span of 9 years and we now have completed the third year.

The project is so large that it could not be executed by one research organization or university group. Consequently, the "contractor" is actually a consortium of three research firms: Human Resources Research Organization (HumRRO) of Alexandria, Virginia; Personnel Decisions Research Institute (PDRI) of Minneapolis, Minnesota; and American Institutes for Research (AIR) of Washington, D.C. The contract is administered and monitored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), which also contributes a sizable proportion of scientific and technical resources to the project.

A parallel effort to Project A is Project B (Development of a Computerized Personnel Allocation System). Project B is responsible for modeling the labor supply and labor demand components of a fully functioning personnel allocation system, and for developing the computer algorithms and software to integrate information on supply, demand, and classification validity.

If both Project A and Project B are successful, the final product will be composed of the following elements:

- A labor supply forecasting model and procedures for estimating the parameter values of the model.
- A model for forecasting the Army's long- and near-term personnel needs (labor demand) and procedures for estimating the parameter values of the model.
- A new set of selection/classification tests which, together with the Armed Services Vocational Aptitude Battery (ASVAB), optimize the balance between the costs of testing and the gain in classification utility.
- A metric and procedure for estimating the utility of performance within and across jobs.

---

[1]Much of the material in Section 1 is drawn from the Project A annual report for the 1983 fiscal year (ARI Research Report 1347) and the 1984 fiscal year (ARI Research Report 1393) and associated documents.

- A set of computerized algorithms (e.g., linear programming) that integrates demand information, supply information, and validity information in such a way that, for any designated period, the overall utility of personnel assignments is maximized.

All of this is an ambitious undertaking. The following report is a summary of the first 3 years of Project A's contributions to the effort as well as a detailed report of activities during the third year.

The Selection/Classification System for Army Enlisted Personnel

The Current System

Each year more than 100,000 new recruits are selected, classified, trained, and assigned to perform the hundreds of jobs required for an effective Army. The system currently used for making the initial selection and classification decision has a long history. The development of the primary selection measure, ASVAB 8/9/10, can be traced through earlier forms--the Army Classification Battery (ACB), the Army Qualification Battery (AQB), the Armed Forces Qualification Test (AFQT), the Army General Classification Test (AGCT)--back to the original Army Alpha.

To be qualified for initial enlistment into the Army by the present system, applicants must meet a number of eligibility criteria, including age, moral standards, physical standards, and "trainability." The latter determination, the most relevant in the current context, is based upon a combination of two sets of criteria: scores attained on the ASVAB, and educational attainment.

The ASVAB is currently administered as an entry test at Military Entrance Processing Stations (MEPS) or at Mobile Examining Team (MET) sites. It is also administered by MET to high school juniors and seniors. Scores from this test are used for guidance counseling and are also provided to Army recruiters as a means of identifying qualified recruitment prospects. In addition to ASVAB, non-high school graduates are administered a short biographical questionnaire, the Military Applicant Profile (MAP), which has been found to be a useful tool for identifying the individuals who are likely to be poor risks in terms of probability of completing Army initial entry training.

For applicants who have not previously taken the ASVAB and whose educational/mental qualifications appear to be marginal in terms of the Army's trainability standards, a short enlistment screening test may be administered to assess an applicant's prospects of passing the ASVAB test. Applicants who appear, upon initial recruiter screening, to have a reasonable prospect of qualifying for service are referred either to a MEP site for administration of the ASVAB, or directly to a MEPS. MEPS staff complete all aspects of the screening process, including administration of the mental and physical examination. On the basis of the information assembled, those found qualified for enlistment are classified by Military Occupational Specialty (MOS) and assigned to a particular training activity.

About 80% of Army enlistees enter the Army under a specific enlistment option that guarantees choices of initial school training, career field assignment, unit assignment, or geographical area. For these applicants, the initial classification and training assignment decision must be made prior to the entry into service. This is accomplished at the MEPS by referring applicants who have passed the basic screening criteria (mental, physical, moral) to an Army guidance counselor, whose responsibility it is to match the applicant's qualifications and preferences to the Army's current skill training requirements, and to "make reservations" for training assignments, consistent with the applicant's enlistment option.

The classification and training reservation procedure is accomplished by the Recruit Quota System (REQUEST), which was implemented in 1973. REQUEST is a computer-based system that coordinates the information needed to reserve training slots for volunteers. One major limitation is that REQUEST uses minimum qualifications for accessions control. Thus, to the extent that an applicant may minimally qualify for a wide range of courses or specialties, based on aptitude test scores, the initial classification decision is governed by (a) his or her own stated preference (often based upon limited knowledge about the actual job content and working conditions of the various military occupations), (b) the availability of training slots, and (c) priorities/needs of the Army. Numerous procedures for improving the system are under development. These include "MOS Match Module" and the previously mentioned Project B Computerized Personnel Allocation System, as well as other smaller efforts.

This review of current practice suggests that the present selection and classification procedures could be improved by taking advantage of recent technological advances and developments in decision theory. There is a need for developing a formal decision-making procedure that is aimed at maximizing the overall utility of the classification outcomes to the Army. However, this decision process must allow for the potentially adverse impacts on recruitment if the enlistee's interests, work values, and preferences are not given sufficient consideration. There are clear trade-offs that must be evaluated between the procedures necessary to (a) attract qualified people, and (b) put them into the right slots.

Modifications Needed in the Current System

The current Army personnel system has a number of features that must be addressed in Project A:

1. Current selection measures cover a fairly limited range of individual characteristics. The ASVAB is an excellent measure of general cognitive abilities. However, in addition, there is a need for developing potentially relevant non-cognitive measures, such as psychomotor/perceptual abilities, vocational interests, and biographical indexes, and determining their usefulness in predicting aspects of Army-wide and MOS-specific performance.

2. No measures of job performance that can be used as criterion measures in validation research are available. Current measures of job proficiency (Skill Qualification Tests--SQT) are designed

primarily as diagnostic training tools rathe than as standardized procedures for performance appraisal.

3. The available information on selection and classification validity is based on the relationship of entrance tests to performance in training, not performance on the job.

4. The Army does not have the data system necessary to make critical personnel decisions throughout a soldier's lifecycle on the basis of accumulating information about the job performance of the individual and the needs and priorities of the Army.

5. Currently, if an applicant chooses a specific training program and meets the minimum aptitude requirements, he or she is placed into that training if an opening exists. This procedure does not take into account where that individual could best serve the needs of the Army or even where that individual could be most successful in the Army.

6. The Army does not have an efficient means of expressing needs and policies in terms of personnel goals, constraints, and trade-offs. An adaptive, self-adjusting system that can more fully support management decision making is needed.

These characteristics of the current system stem primarily from the dynamics in the labor market, the new requirements produced by emerging weapon systems, and the inevitable lag of an operational system behind the most recent technological advances in testing and personnel decision making.

Origins of the Project

In response to needs expressed by the Army and by Congress, as well as the previously mentioned professional considerations, ARI began in 1980 to develop a major new research and development (R&D) program in personnel selection, classification, and allocation. The basic requirement was to demonstrate the validity of the ASVAB as a predictor of both training and on-the-job performance.

While ARI staff were systematically reviewing that requirement, the concept of a larger project began to emerge. With only a moderate amount of additional resources, new predictors in the perceptual, psychomotor, interest, temperament, and biodata domains could be evaluated as well. A longitudinal research data base could be developed to accumulate information on a variety of predictor/criterion relationships from enlistment, through training, first-tour assignments, reenlistment decisions, and for some, to their second tour. Also, the data could be the basis for making near-real-time decisions on the best match between characteristics of an individual enlistee or reenlistee and the requirements of available Army Military Occupational Specialties (MOS).

To address the selection and classification portion of the effort, solicitation MDA 903-81-12-R-0158, "Project A: Development and Validation of Army Selection and Classification Measures," was issued 21 October 1981.

This document is the "official" starting point of Project A. The solicitation initially outlined a 7-year program designed to provide the information necessary for implementing a state-of-the-art selection and classification system for all U.S. Army enlisted personnel.

While the contract Statement of Work (SOW) and the Request for Proposals (RFP) were being developed, certain structural changes were made within ARI to accommodate the project. A new manpower and personnel laboratory was created with Joyce L. Shields as director, and a selection and classification technical area was established headed by Newell K. Eaton. To execute the in-house research and to monitor the contract, it was also necessary to recruit additional professional staff.

In response to the RFP, the Human Resources Research Organization, American Institutes for Research, and Personnel Decisions Research Institute, formed a consortium to develop a research proposal for Project A. HumRRO assumed the responsibility as the prime contractor, and the consortium's proposal was submitted in January 1982. The contract was awarded to the HumRRO-AIR-PDRI consortium 30 September 1982.

Specific Objectives of Project A

The project has two principal kinds of objectives. The first type pertains to the operational needs of the Army. They constitute the basic purposes for which the project is funded and supported. Specifically, Project A is to:

1. Develop new measures of job performance that can be used as criteria against which to validate selection/classification measures. The new criterion measures will use a variety of methods to assess both job-specific measures of task performance and general performance factors that are not job specific.

2. Validate existing selection measures against both existing and project-developed criteria.

3. Develop and validate new and/or improved selection and classification measures.

4. Validate proximal criteria, such as performance in training, as predictors of later criteria, such as job performance ratings, so that more informed decisions about reassignment and promotion can be made throughout the individual's tour.

5. Determine the relative utility to the Army of different performance levels across MOS.

6. Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

A second set of objectives has to do with questions of a more scientific nature. This second set of questions is being addressed with essentially the same data as the first. That is, the project does not have two parts with

one having to do with basic research and the other focused on applied research. Instead, the scope of the project and the attempt to consider an entire system at one time make it possible to concurrently address a number of more basic research objectives. Some of these are as follows:

1. Identify the basic variables (constructs) that constitute the universe of information available for selection/classification into entry-level-skilled jobs.

2. Develop a comprehensive model of performance for entry-level-skilled jobs that incorporates both a theoretical latent structure and linkages to state-of-the-art measurement.

3. Describe the utility functions and the utility metrics that individuals actually use when estimating "utility of performance."

4. Describe the degree of differential prediction across (a) major domains of abilities, personality, interests, and personal history, (b) major factors of job performance, and (c) different types of jobs. The project will collect a large sample of information from each of these three populations (i.e., individual differences, performance factors, and jobs).

5. Determine the extent of differential prediction across racial and gender groups for a systematic sample of individual differences, performance factors, and jobs.

6. Develop new statistical estimators of classification efficiency.

Each of the above objectives, both applied and basic, breaks down into a number of more specific questions that will be touched on in later sections.

Project A Organization

Task Structure

For purposes of an orderly division of labor, Project A is organized into five major research tasks:

Task 1. Data Base Management and Data Analysis. Task 1 has two major components. The first component deals with designing, generating, and maintaining the data base. By the end of the project the data base will contain several hundred thousand records taken from three Army troop cohorts, three major validation samples, and numerous pilot samples. The second component is concerned with providing the analytic capability for (a) analyzing field test and validation data and (b) evaluating the existing set of predictors against the new performance measures, to determine whether the new predictors have incremental validity over and above the present system. These two components must be accomplished using state-of-the-art technology in methods for analyzing personnel selection research data.

Task 2. Development of Predictors of Job Performance. To date, a large proportion of the efforts of the Armed Services in this area has been concentrated on improving the ASVAB, which is now a well-researched, valid measure

of general cognitive abilities. However, many critical Army tasks depend on psychomotor and perceptual skills for their successful performance. Further, neither biodata nor motivational variables are now comprehensively evaluated. It is perhaps in these four non-cognitive domains that the greatest potential for adding valid independent dimensions to current classification instruments is to be found. The objectives of Task 2 are to develop a broad array of new and improved selection measures and to administer them to three major validation samples. A critical aspect of this task is the demonstration of the incremental validity added by new predictors.

Task 3. Measurement of School/Training Success. The objective of Task 3 is to derive school and training performance indexes that can be used (a) as criteria against which to validate the initial predictors, and (b) as predictors of later job performance. Comprehensive job knowledge tests will be developed for the sample of MOS investigated, and their content and construct validity will be determined. Two additional purposes for developing training criterion measures are to determine the relationship of training performance to job performance and to find out whether validating against each of these kinds of criteria selects the same or different predictor measures.

Task 4. Assessment of Army-Wide Performance. In contrast to performance measures that may be developed for a specific Army MOS, Task 4 will develop measures that can be used across all MOS (i.e., Army-wide). The intent is to develop measures of first- and second-tour job performance against which all Army enlisted personnel may be assessed. A major objective for Task 4 is to develop a model of soldier effectiveness that specifies the major dimensions of an individual's contribution to the Army as an organization. Another important objective of Task 4 is to develop measures of performance utility.

Task 5. Development of MOS-Specific Performance Measures. The focus of Task 5 is the development of reliable and valid measures of specific job task performance for a selected set of MOS. This task may be thought of as having three major components: job analysis, construction of job performance measures, and validation of the constructs for the new measures. While only a subset of MOS will be analyzed during this project, the Army may in the future wish to develop job performance measures for a larger number of MOS. For this reason, it is intended that the methods used will apply to all Army MOS.

In addition, Task 6 deals with administrative management of the project.

The Consortium/ARI Organization

The initial project organization is depicted in Figure I.1. The principal consortium investigators are shown, with their respective organizations, in the lower row. The principal ARI staff are shown in the upper row. Within the project, consortium and ARI investigators undertake both independent and joint research activities. ARI staff also have the administrative role of contract oversight.

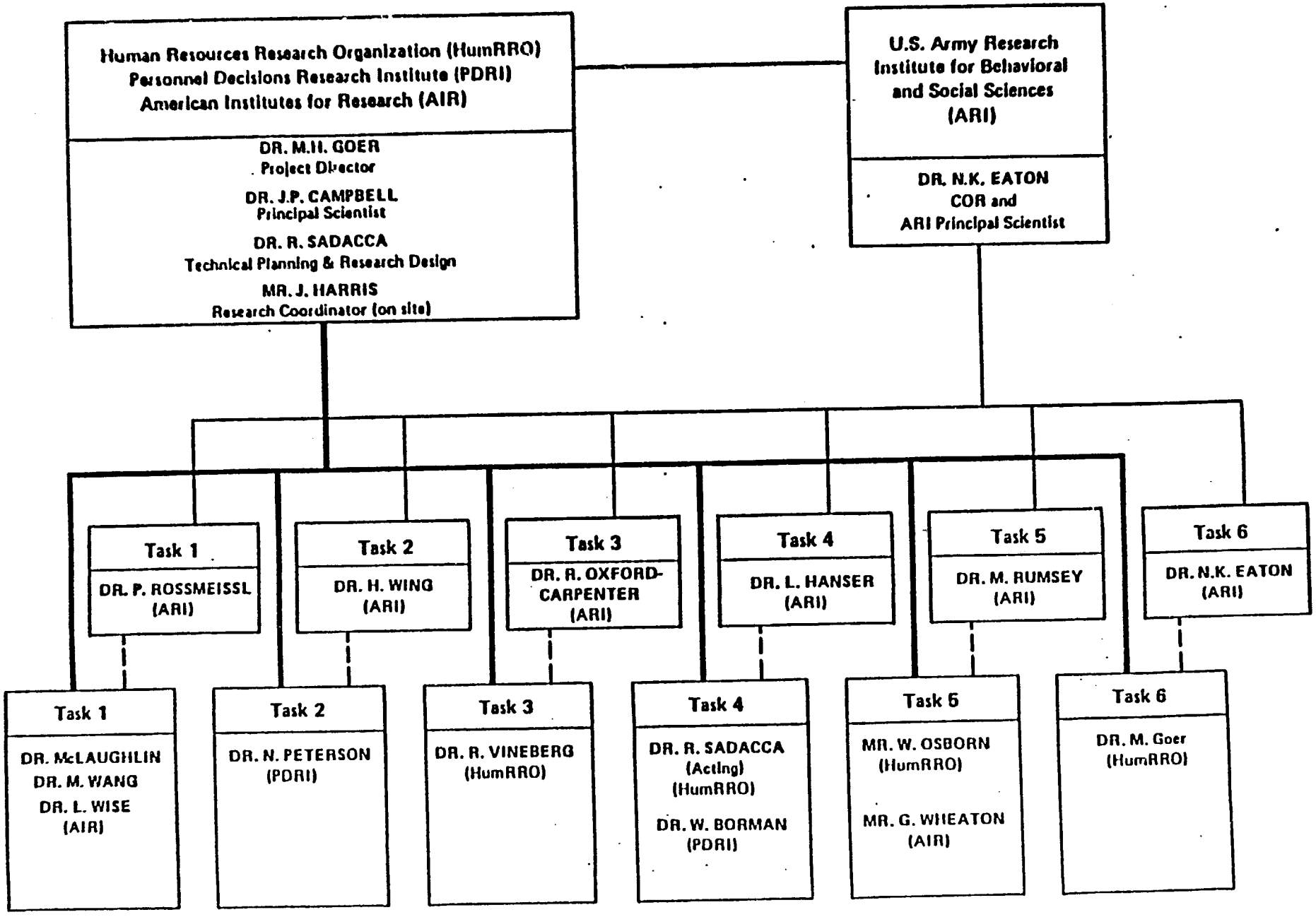


Figure I.1. Project A organization as of 30 September 1983.

During the first 3 years, technical and management oversight has been the responsibility of Newell K. Eaton, the contracting officer's technical representative (COR). He has been the ARI principal scientist, with the responsibility for technical review and guidance. Consortium management has been the responsibility of Marvin H. Goer, the managing project director. Within the consortium, John P. Campbell has been the principal scientist responsible for overall scientific quality. Robert Sadacca has been the assistant for technical planning and research design. James Harris has been primarily responsible for the day-to-day coordination of the project's multiple activities.

The Advisory Group Structure

To ensure that Project A is consistent with other ongoing research programs being conducted by the other armed services, a mechanism was needed for maintaining close coordination with the other military departments, as well as with the Department of Defense. A procedure also was needed to assure that the research program is technically sound, both conceptually and methodologically. Finally, a method was needed to receive feedback on priorities and objectives, as well as to identify current problems before they become too large to fix.

The method used to meet these needs was to establish a series of advisory groups. Figure I.2 shows the structure and membership of the Governance Advisory Group, which is comprised of the Scientific Advisory

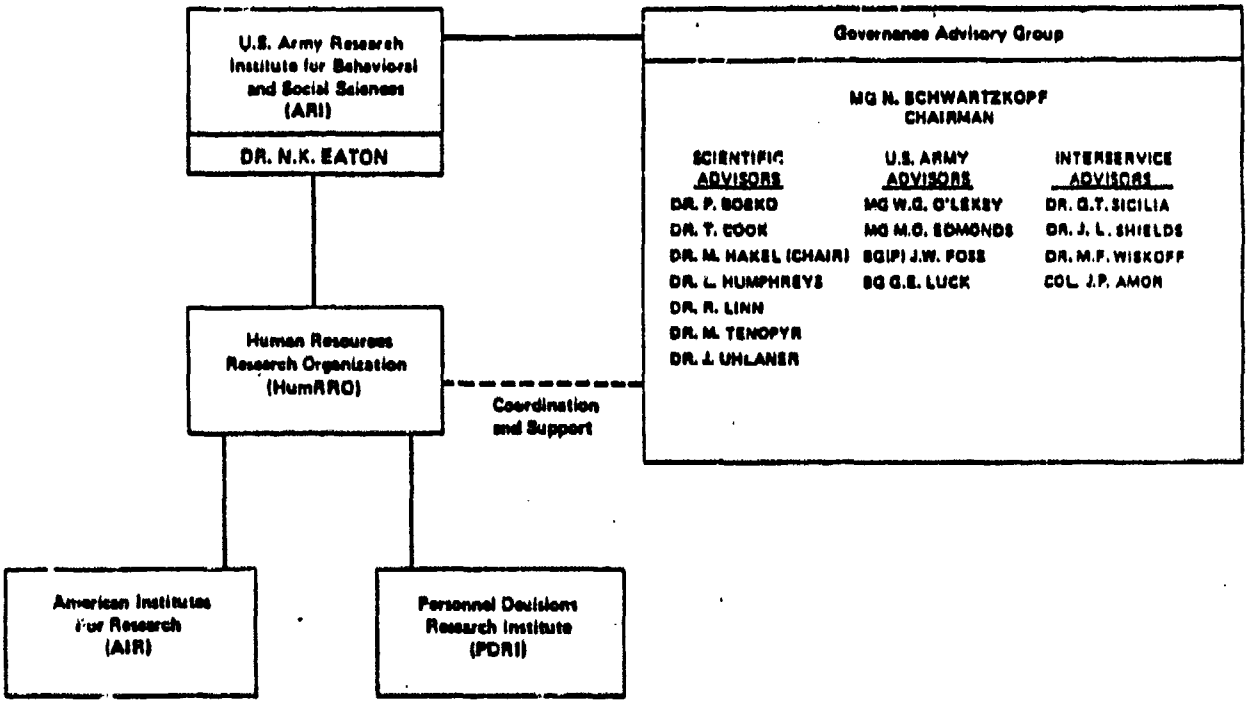


Figure I.2. Governance Advisory Group as of 30 September 1983.

Group (SAG), Army Advisory Group (GAG), and Interservice Advisory Group (ISAG) components.

The SAG comprises nationally recognized authorities in psychometrics, experimental design, sampling theory, utility analysis, and applied research in selection and classification, and in the conduct of psychological research in selection in the Army environment. The ISAG comprises the Laboratory Directors for applied psychological research in the Army, Air Force, and Navy, and the Director of Accession Policy from the DoD Office of Assistant Secretary of Defense for Manpower and Reserve Affairs.

The GAG includes representatives from the Office of Deputy Chief of Staff for Personnel (DCSPER), Office of Deputy Chief of Staff for Operations (DCSOPS), Training and Doctrine Command (TRADOC), Forces Command (FORSCOM), and U.S. Army Europe (USAREUR). These senior officers have a significant interest in the project planning and priorities. They also represent the elements that provide the necessary and substantial troop support.

The Research Plan and Integrated Master Plan

The first 6 months of the project were spent planning, documenting, reviewing, modifying, and redrafting research plans, troop support requests, administrative support plans, and budgetary plans, as well as executing initial research efforts. Drafts of the plans were provided to the SAG and ISAG. Their comments, provided orally during meetings and subsequently written in response to draft documents, were incorporated in the research plan.

The culminating review was conducted in April 1983 by the Army Advisory Group, with representatives from the Scientific and Interservice Advisory Groups. In that meeting the advisors reviewed the entire research program, research design, sampling strategy, main cohort and focal MOS recommendation, and troop support implications. They incorporated changes to reduce the troop support burden and distribute it more equitably among the three participating commands (FORSCOM, TRADOC, USAREUR). All three components of the Governance Advisory Group endorsed the research program.

In May 1983, ARI issued ARI Research Report 1332, Improving the Selection, Classification, and Utilization of Army Enlisted Personnel--Project A: Research Plan. In June 1983, the "Project A: Integrated Master Plan" (HumRRO FR-PRD-83-8) was issued, providing detailed budget allocations, schedules, and specifications of contract deliverables.

Summary of Research Design and Sample Selection

The overall design of Project A is described in detail in the Master Research Plan (June 1983). Again, the overall objectives are to develop and validate an experimental battery of new and improved selection measures against a comprehensive array of job performance and training criteria. The validation research must produce sample estimates of the parameters necessary to implement a computerized selection and classification system for all first-tour enlisted MOS.

Research Design

To meet these objectives, a design was developed that uses two predictive and one concurrent validation on two major troop cohorts (FY83/84 accessions and FY86/87 accessions), and one file data validation on the FY81/82 cohort. That is, in addition to collecting data from new samples, the project is making use of existing file data that have been, or can be, accumulated for 1981 and 1982 accessions. A schematic of the data collection plan is shown in Figure I.3.

The logic of the design is straightforward. Existing file data on the FY81/82 cohort provided the first opportunity to revalidate the ASVAB against existing training criteria and against the SQT. As described in a separate report (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984), and summarized later in this report, the results of the analyses of FY81/82 file data were used to suggest operational changes in ASVAB composites. The file sample consisted of approximately 90,000 records distributed over 120 MOS in sufficient numbers to permit analysis. The FY81/82 data also provide a benchmark against which to compare the additional validation data to be collected.

The FY83/84 cohort provided the first opportunity to obtain validation data using new predictor tests and new performance measures. Two samples have been taken from this cohort. First, a "preliminary" predictor battery of predominately off-the-shelf tests chosen to represent major constructs was administered to soldiers in four MOS (31C, 19E/K, 63B, 71L) as they entered the Army during the last half of FY83 and the first half of FY84. A total of 11,000 personnel in the four MOS were tested. Besides looking at the relationship of the Preliminary Battery constructs to the existing ASVAB, we followed a portion of this sample during the summer and fall of 1985 with a broad array of criterion measures (described later). The follow-up of the Preliminary Battery sample was part of a much larger Concurrent Validation sample drawn from 1985 job incumbents who entered the Army during FY83/84.

Results from the administration of the preliminary predictor battery sample (described later under predictor development) were used to help develop the trial predictor battery for use in the major Concurrent Validation during the summer and fall of 1985. Immediately prior to the Concurrent Validation, all predictors and all criterion measures were put through a series of field tests. For example, all criterion measures were field tested on approximately 150 incumbents in each of nine MOS. The test battery used during the predictor field tests was labeled the Pilot Trial Battery. Both the Preliminary Battery sample and the field tests were used to develop the Trial Battery for use in the Concurrent Validation.

The Trial Battery is being validated in a sample of 19 MOS against an array of newly developed training and job performance measures. For each MOS, 500-700 incumbents are being tested. As noted above, a subset of the Concurrent Validation sample took the Preliminary Battery approximately 18 months earlier, which will permit a longitudinal validation of the off-the-shelf tests that were selected to represent major ability and personality constructs.

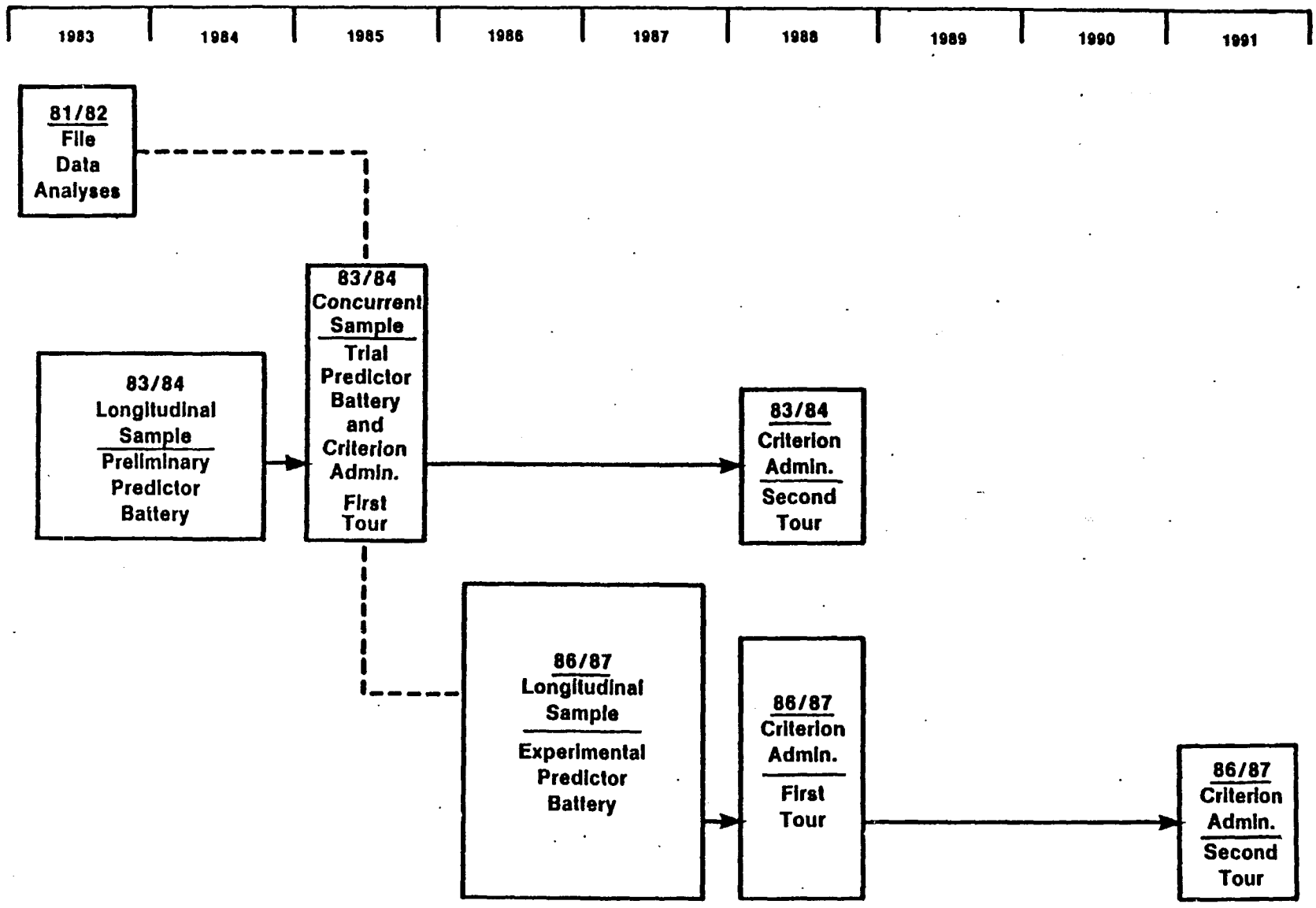


Figure I.3. The overall data collection plan.

Analysis of the Trial Battery data will result in further revision of the predictor battery. The revised version will then be called the Experimental Battery, which will be used with a longitudinal validation sample selected from people who enter the Army in FY86 and FY87. The Experimental Battery will be administered at the time of entry to approximately 50,000 people distributed across 19 MOS. The training measures will be administered at the conclusion of each individual's Advanced Individual Training (AIT) course and the job performance criterion data will be collected approximately 18 months later. In addition, both general (Army-wide) and job (MOS)-specific performance measures will be developed and administered to the surviving members of both the FY83/84 and FY86/87 cohort samples during their second tour of duty. Consequently, for both these samples the design is also a longitudinal one.

Sample Selection

The overall objective in generating the samples has been to maximize the validity and reliability of the information to be gathered, while at the same time minimizing the time and costs involved. In part, costs are a function of the numbers of people in the sample. However, costs are also influenced by the relative difficulty involved in locating and assembling the people in a particular sample, by the degree to which the unit's operations are disrupted by the data collection, by the staff costs involved in collecting the data in a particular manner, and by other such factors.

The sampling plan itself incorporated two principal considerations. First, a sample of MOS was selected from the universe of possible MOS; then, the required sample sizes of enlisted personnel (EP) within each MOS were specified. The MOS are the primary sampling units. This design is necessary because Project A is developing a system for a population of jobs (MOS), but only a sample of MOS can be studied.

Large and representative samples of enlisted personnel within each selected MOS are important because stable statistical results must be obtained for each MOS. There is a trade-off in the allocation of project resources between the number of MOS researched and the number of subjects tested within each MOS: The more MOS are investigated, the fewer subjects per MOS can be tested, and vice versa. Cost versus statistical reliability considerations dictated that 19 MOS could be studied.

To samples from all 19 MOS we have administered the new predictors (from Task 2) and collected the school and Army-wide performance data (of Tasks 3 and 4). For nine of these MOS, we have also administered the MOS-specific performance measures developed in Task 5. The nine MOS were chosen to provide maximum coverage of the total array of knowledge, ability, and skill requirements of Army jobs, given certain statistical constraints.

MOS Selection

The selection of the sample of 19 MOS proceeded through a series of stages. The guidelines that follow were used to draw an initial sample of MOS.

- High-density MOS that would provide sufficient sample sizes for statistically reliable estimates of new predictor validity and differential validity across racial and gender groups.

- Representativn coverage of the aptitude areas measured by the ASVAB area composites.

- High-priority MOS (as rated by the Army[2] in the event of a national emergency).

- Representation of the Army's designated Career Management Fields (CMF).

- Representation of the jobs most crucial to the Army's mission (e.g., the combat specialties).

This set of 19 MOS represented 19 of the Army's 30 Career Management Fields (CMF). Of the 11 CMF not represented, 2 (CMF 96 and 98) are classified, 2 (CMF 33 and 74) had fewer than 500 FY81 accessions, and 7 (CMF 23, 28, 29, 79, 81, 84, and 74) had fewer than 300 FY81 accessions. The initial set includes only 5% of Army jobs but 44% of the soldiers recruited in FY81.

Similarly, of the 15% women in the 1981 cohort, 44% are represented in the sample; of the 27% blacks, 44% are represented in the sample; and of the 5% Hispanic, 43% are represented. Although female and minority representation is high absolutely, relatively it remains about the same as in the population.

Nine of the 19 MOS were earmarked for the job-specific performance measurement phase of the project. These were selected, as a subset, with the same general criteria used in identifying the parent list of 19. Since the larger list is composed of 5 combat and 14 noncombat MOS, it seemed reasonable that these categories be proportionally represented in the subset of 9. Consequently, the 9 MOS designated for job-specific performance measurement development are:

	(1)	11B	- Infantryman
*	(2)	13B	- Cannon Crewman
	(3)	19E/K	- Tank Crewman
	(4)	05C	- Radio TT Operator
	(5)	63B	- Vehicle and Generator Mechanic
*	(6)	64C	- Motor Transport Operator
*	(7)	71L	- Administrative Specialist
	(8)	91B	- Medical Care Specialist
*	(9)	95B	- Military Police

An initial batch of four (designated on the list by asterisks) was selected and termed Batch A; the other five are Batch B. Work was begun first on Batch A and then on Batch B.

---

[2]ODCSOPS (DAMO-ODM), DF, 2 Jul 82, Subject: IRR Training Priorities.

Refinements of the MOS sample included a cluster analysis of expert ratings of MOS similarity and a review of the initial sample by the Governance Advisory Group.

MOS Cluster Analysis

To obtain data for empirically clustering MOS on the basis of their task content similarity, a brief job description was generated for each of 111 MOS from the job activities described in AR 611-201[3]. The sample of 111 MOS represents 47% of the population of 238 Skill Level 1, Active Army MOS with conventional ASVAB entrance requirements. It includes the 84 largest MOS (300 or more new job incumbents yearly), plus an additional 27 selected randomly but proportionately by CMF. Each job description was limited to two sides of a 5x7 card.

Members of the contractor research staff and ARI Army officers--approximately 25 in all--served as expert judges and were given the task of sorting the sample of 111 job descriptions into homogeneous categories based on perceived similarities and differences in job activities as described in AR 611-201. Data from the similarity scaling task were used to cluster analyze the matrix of similarities for the 111 jobs.

The results were used to check the representativeness of the initial sample of 19 MOS. That is, did the initial sample of MOS include representatives from all the major clusters of MOS derived from the similarity scaling? On the basis of these results, and guidance received from the Governance Advisory Group, two MOS that had been selected initially (62E and 31M) were replaced by MOS 51B and MOS 27E, which are in the same CMF and involve the same Aptitude Area Composites as the replaced MOS. The sample of MOS resulting from the above procedures is shown in Table I.1.

The next two sections of this report summarize in somewhat more detail, the project's activities for FY83 (year one) and FY84 (year two).

---

[3]Army Regulation 611-201, Enlisted Career Management Fields and Military Occupational Specialties.

Table I.1. Project A Military Occupational Specialties (MOS)

MOS	Title	CMF	APT Comp	Priority MOS	FY81 Accessions				Trainee Projections		Expected Number Graduates[1]
					Total	Women	Blacks	Hispanic	FY83	FY84	
05C	Radio TT Operator	31	SC	No	3175	585	898	119	2004	2200	1645
63B	Vehicle & Generator Mech	63	MM	No	4653	386	1178	242	5304	4402	4280
64C	Motor Transport Operator	64	OF	Yes	5440	774	1279	141	3706	5000	4484
71L	Admin Specialist	71	CL	No	4484	2744	1967	215	6191	4592	3859
13B	Cannon Crewman	13	FA	Yes	5783	0	2053	367	6092	3553	3572
91B	Medical Care Specialist	91	ST	Yes	3074	924	876	224	3761	unav	3621
19E/K	Tank Crewman	19	CO	Yes	3233	0	604	188	3223	3261	2912
95B	Military Police	95	ST	Yes	6073	704	624	127	5720	5300	4373
11B	Infantryman	11	CO	Yes	7028	0	1128	367	12633	13710	11338
76Y	Unit Supply Specialist	76	CL	No	4565	1179	1998	283	6636	4091	3829
94B	Food Service Specialist	94	OF	No	3859	715	1416	125	5133	5157	4600
12B	Combat Engineer	12	CO	Yes	3707	0	716	147	844	2540	1845
16S	MANPADS Crewman	16	OF	Yes	691	0	206	27	797	1015	815
55B	Ammunition Specialist	55	GM	No	662	171	283	42	620	810	762
76W	Petroleum Supply Spec	92	CL	NO	849	259	559	43	1373	1350	1234
54E	Chemical Operations Spec	54	ST	Yes	557	89	185	41	1012	1247	1068
67N	Utility Helicopter Rpr	67	MM	No	1032	33	68	29	572	465	470
51B	Carpentry/Masonry Spc	51	GM	No	602	6	136	14	120	483	341
27E	Tow/Dragon Rpr	27	EL	No	333	40	76	17	312	308	258
	Total				59800	8609	16250	2758	66053	59484	55306

[1] Weighted average of Trainee Projections (3 months of FY83 and 9 months of FY84) adjusted for expected school attrition (actual FY81 rates).

Section 2

OVERVIEW OF FISCAL YEAR 1983

During the first year of the project, detailed plans were prepared, the sample of focal MOS was selected, the sample sizes required from each were specified, and work was begun on the comprehensive predictor and criterion development that would be the basis for the later validation. In addition, the available computer file data on the FY81/82 cohort were merged from the various sources, edited thoroughly, and prepared for analysis.

Plans for the project as a whole and activities during the first year were described in the Annual Report for the 1983 fiscal year (ARI Research Report 1347) and the technical appendix to that report (ARI Research Note 83-27), both published in October 1983.

Planning Activities

In general, as previously noted, much of the first year's effort was taken up by an intensive period of planning, briefing the advisory groups, preparing the initial troop requests, and related activities.

The requirement for a detailed research plan to be produced during the first 6 months of the contract was included in the RFP. Hindsight judges it to be an even more valuable step than the authors of the RFP might have had in mind. The research staff devoted a great deal of effort to the writing of the research plan, and it was carefully reviewed by the advisory groups and by the ARI professional staff. Revisions were then made, and the completed plan was published in May 1983 under the joint authorship of the contractor and ARI staffs.

The Research Plan and the accompanying Master Plan lay out, in detail, the specific steps to be taken in each subtask in the project, the schedule to be followed, and the budget allocations to be made to each subtask during each contract period. These two documents have become the blueprint for the project. They have also proven invaluable as a mechanism for developing a consensus and facilitating communication among contractor staff and between the contractor and ARI.

The detailed planning and review that went into the development of the Research Plan and Master Plan made it possible to specify, clearly and precisely, the troop support the project would need during its first 2 years. Consequently, the project staff has experienced relatively little difficulty in communicating project needs to the appropriate Army organizations and in gaining their support. The cooperation we have received has been outstanding.

Criterion Development

During FY83 the development of performance measures proceeded through the steps described below.

MOS Task Descriptions

Because the information had not been generated for personnel research purposes, the Army's MOS job analysis data needed considerable modification before they could be used by Project A for criterion development. Consequently, a great deal of effort in FY83 was devoted to refining and integrating task descriptions from Soldier's Manuals and from the Comprehensive Data Analysis Program (CODAP) occupational survey questionnaires. For each MOS, a data bank of task statements was accumulated from all available sources, and the individual task statements were edited to determine if they indeed focused on observable job tasks, if they were redundant or overlapped with other tasks, and if they were at the same level of generality. Subject matter experts (SMEs) were consulted to determine whether the edited pool of task descriptions provided a complete picture of the MOS content. The SMEs also judged the relative criticality of each task.

The resulting task descriptions provided the principal basis for the development of hands-on performance measures and job knowledge tests.

Assessment of Training Performance

A major objective of Project A is to use a comprehensive and standardized test construction procedure to develop a measure of training success for each focal MOS, in which the item content represents both the content of training and the content of the job. That is, the items will sample the job content representatively and will be further identified as being covered in training vs. not being covered in training. When this is accomplished, a measure of direct learning in training (scores on items that match training content) and a measure of indirect learning (scores on items <u>not</u> directly related to training content) can be related to a variety of job performance criteria, with and without ability (as measured by predictor tests) controlled.

On the way to developing norm-referenced training achievement tests for each of the 19 MOS, the staff visited each Proponent school and developed a description of the objectives and content of the training curriculum. They also used Army Occupational Survey Program (AOSP) information to develop a detailed task description of job content for each MOS. After low-frequency elements were eliminated, SME judgments were used to rate the importance and error frequency for each task element. Approximately 225 tasks were then sampled proportionately from MOS duty areas.

What was produced was a thorough analysis of the objectives, curriculum, and assessment procedures for the key schools. The process of describing MOS job content and matching it with training content was begun in FY83 and completed during FY84.

Assessment of Job Performance

The initial model of soldier effectiveness which we developed was perhaps a bit crude. We said essentially that both specific task performance and the general factors of commitment, morale, and organizational socialization comprised the total domain.

During FY83 the task descriptions for the four MOS in Batch A were completed and those for Batch B were in progress. Virtually all the critical incident workshops necessary for constructing MOS-specific task performance factors were completed. This most likely has been the most massive effort ever undertaken to apply Behaviorally Anchored Rating Scales (BARS) methods to criterion development. There now exist accounts of hundreds of critical incidents of specific task performance within each focal MOS, and thousands of critical incidents describing performance behaviors that have a general, not MOS-specific, referent. These large samples of job behaviors were used to identify MOS-specific and MOS-general performance factors and (during FY84) to develop rating scales to assess individual performance on these factors. This process produced a revised and expanded model of the criterion space to be used to generate further criterion development work.

An additional important outcome of the interaction between developing the model and describing tasks/behavior was the identification of an array of MOS-specific task performance factors intended to encompass the unique task content of all MOS in the enlisted personnel job structure. Although it was only a first cut, it provides the basis for the further development of a standardized set of task descriptors that can be applied to any MOS to describe its content. Such a standardized measure will make it possible to answer a number of important questions that could not have been addressed previously. For example, how similar are any two MOS in terms of their job content? Should they have a common selection algorithm? How different should their training schools be?

Predictor Selection

A major objective that had to be accomplished during the first contract year was to select the preliminary predictor battery for administration to the FY83/84 longitudinal sample and to lay the groundwork for the development of the trial predictor battery. To do this, the project staff carried out a massive literature search. The result was (a) a description of the specific measures that might be useful in any selection or classification effort, (b) a summary of the empirical evidence attendant to each one, and (c) an explication of the latent variables, or constructs, that seem to best represent the content of the operational measures or tests.

Data Base Management/Validation Analysis

Project A will generate a large amount of interrelated data that must be assembled into an integrated data base that can be accessed easily by the research teams for analytical purposes. Therefore, a major task was to establish and maintain the longitudinal research data base (LRDB), which links data on diverse measures gathered in the various tasks of Project A and

incorporates existing data routinely collected by the Army. Such a comprehensive LRDB will enable Project A to conduct a full analysis of how information gathered at each stage of the enlistee's progress through his or her Army career can add to the accuracy of predicting later performances.

In accordance with the Project A Research Plan, the LRDB will contain three major sets of data. The first set consists of existing data on FY81/82 accessions, including accession information (demographic/biographical data, test scores, and enlistment options), training success measures, measures of progress or attrition taken from the Enlisted Master File (EMF), and specific information on SQT scores. This first set of data is to be employed to validate the current version of the Armed Services Vocational Aptitude Battery (ASVAB), insofar as that can be done with available criteria. It will be used to investigate major methodological and conceptual issues. The second and third major sets of data will involve the new data collection efforts of the FY83/84 and FY86/87 cohorts.

A significant portion of the first year's LRDB activities involved planning the data base contents and procedures for the duration of the project. The main result of this activity was the draft and final LRDB plan. Other planning accomplishments included installing the RAPID data storage and retrieval system, developing workfile generation and data set documentation programs, identifying and implementing data file integrity and security procedures, and establishing data editing procedures.

Most of the substantive LRDB results during the first year were related to the creation of the FY81/82 cohort data base for use in the preliminary validation of the current ASVAB and the evaluation of new aptitude area composites. The validity and differential validity of the existing predictors (ASVAB 8/9/10) against existing criteria (training grades, SQT, and administrative outcomes) were being determined on all MOS for which there are sufficient data. These results will serve as a benchmark against which the subsequent validations using new and/or improved predictors and criterion measures can be compared. The validity of alternative composites of ASVAB subtests can be compared with the validity of the existing composites.

In Conclusion

During its first contract year Project A stayed on schedule and within its budget. More attention than the Army's research staffs had originally envisioned was devoted to detailed planning and outside review. However, these thorough and careful preparatory steps seemed well worthwhile in terms of facilitating communication among all persons associated with the project and uncovering unresolved issues that would have plagued us at some later time.

Also, although much of the research activity during the first year was designed as essentially preparatory, some valuable first-year products include the 81/82 data file, the task banks, the critical incident banks, and the literature review of the predictor domain.

Section 3

OVERVIEW OF FISCAL YEAR 1984

During the second year of work on Project A, the major efforts were in the development of performance measures and predictor tests, evaluation of the validity of the ASVAB, and exploratory investigation of procedures for scaling the utility of performance levels across MOS.

The work performed during the second year is described in detail in the Annual Report Synopsis for FY84 (ARI Research Report 1393) and a companion detailed report (ARI Technical Report 660), which also includes technical documents that were prepared during the year to support various aspects of the program (and which is supplemented by ARI Research Note 85-14, containing additional appendix material). All three reports were published in October 1984.

Project Administration

The overall administration and structure of Project A continued without change in FY84. However, a contract amendment dealing with the scope of work was designed and implemented as envisaged in the Research Plan (ARI Research Report 1332). 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 training school data on the FY83/84 cohort for predictor and criterion development.
- Conducting validity analyses of the FY81/82 cohort data.
- Conducting additional job and task analyses to support refinements in the MOS sample.
- Preparing detailed analyses to support the sampling strategy (and the resultant Troop Support Requests).
- Developing and administering the "Preliminary Battery."
- Acquiring, using, and maintaining computerized psychomotor/perceptual test equipment.
- Expanding the utility research program.
- Extending the research schedule through 1991 to retain the objective of analyzing second-term validity data on the second (FY86/87) main cohort.

Included in the changes noted above was a requirement for an extensive investigation of psychomotor/perceptual measures. Implementing this decision required the acquisition, use, and maintenance of computer-driven test equipment.

During the course of the second year there were several personnel changes in the Governance Advisory Group. These changes are reflected in Figure I.4. 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 I.5.

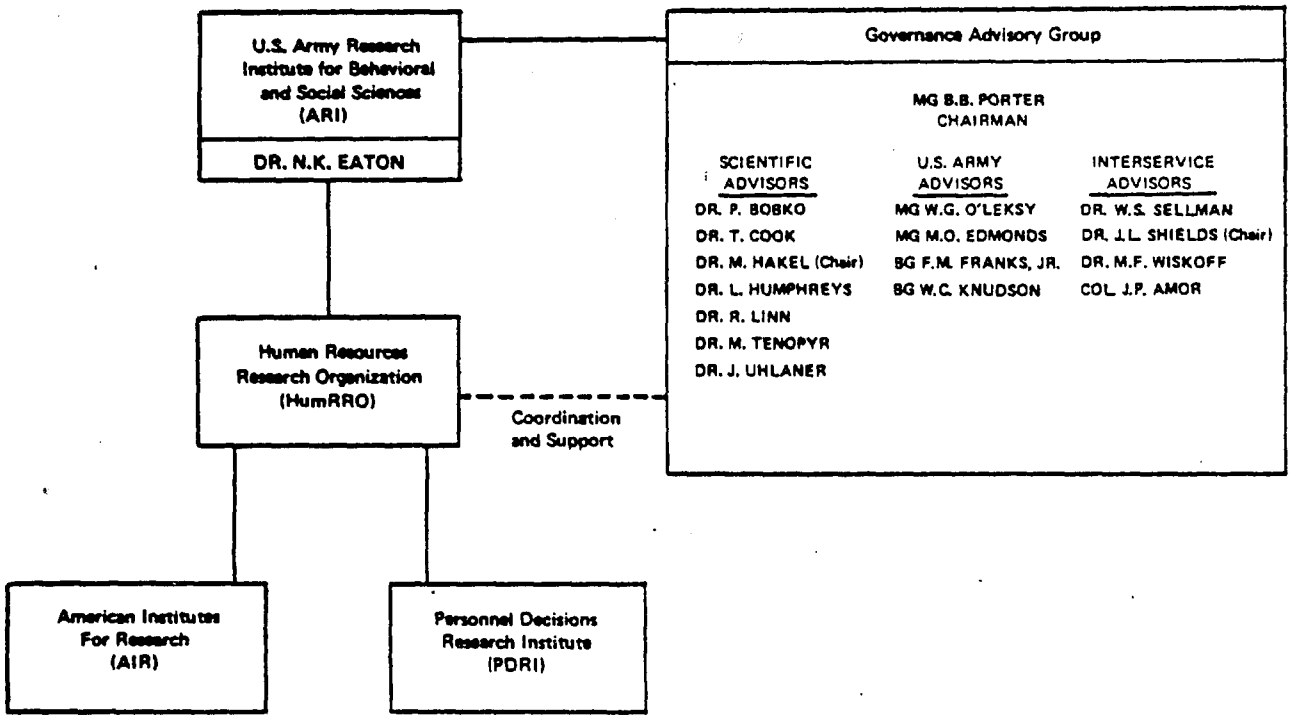


Figure I.4. Governance Advisory Group as of 30 September 1984.

School and Job Performance Measurement

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).

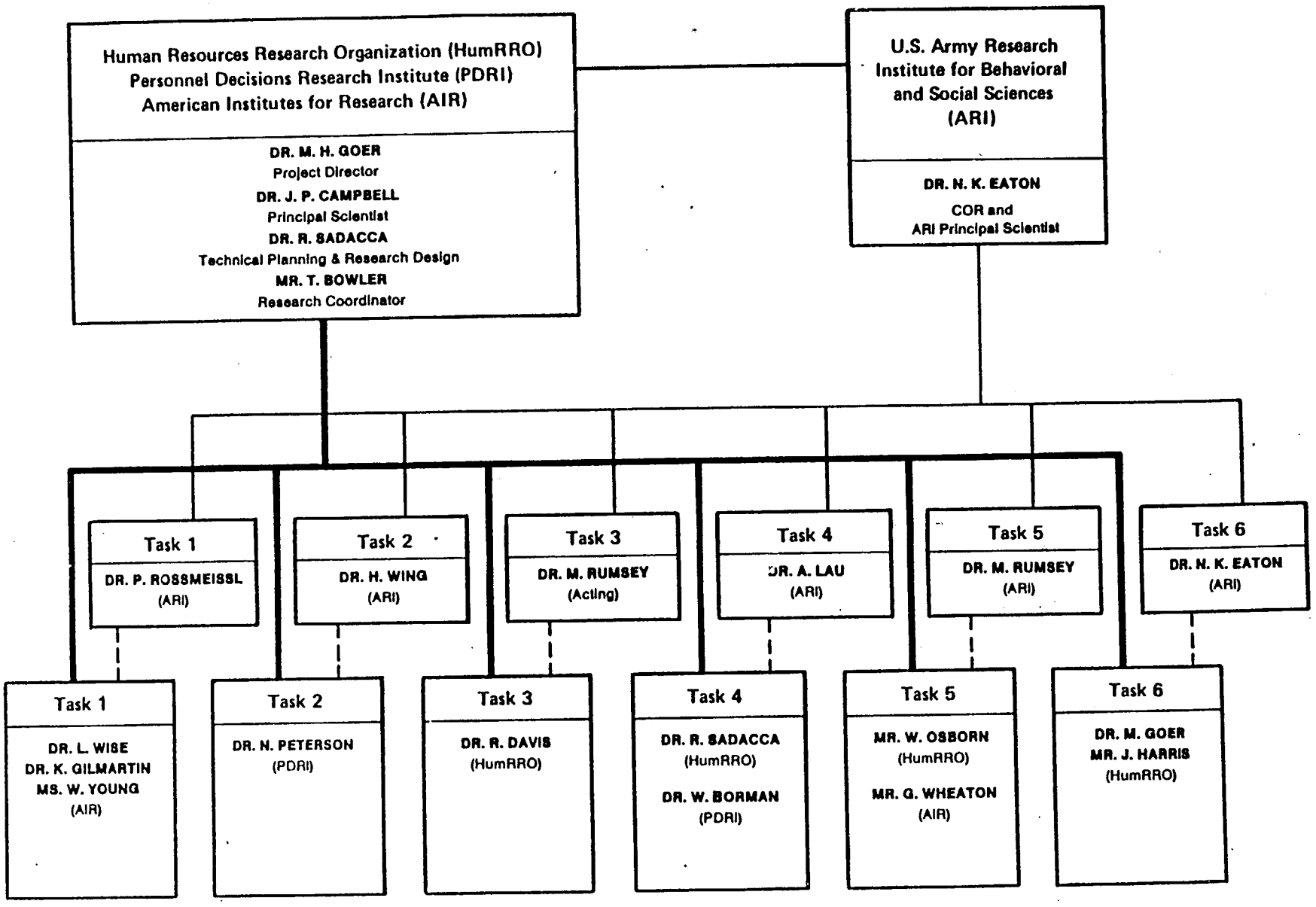


Figure I.5. Project A organization as of 30 September 1984.

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

- In working toward norm-referenced training achievement tests, 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), the Official Military Personnel File (OMPF), which is centrally stored on microfiche, and the Military Personnel Records Jacket (201 File).

The principal objectives for criterion development for FY84 were to (a) use the information developed in FY83 to construct the initial version of each criterion measure, (b) pilot test each initial version and modify as appropriate, and (c) 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). The field test continued into FY85 during which the criterion measures for the five MOS in Batch B were evaluated.

During FY84 a pilot version was developed for most, but not all, criterion measures. The specific progress on each measure is described below.

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 assignments of tasks to testing mode (i.e., hands-on job samples vs. knowledge testing).

For assignment of tasks to testing mode, each task was rated by three to five project staff on three dimensions. The extent to which a task was judged to require (a) a high level of physical skill, (b) a series of prescribed steps, and (c) 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 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 also developed. The initial hands-on measures and scorer directions were then pretested on 5 to 10 job incumbents in each MOS and revised.

MOS-Specific Job Knowledge Tests (Batch A). 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 an average of about nine 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.

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. However, task descriptions had not yet been submitted to SME judgments about difficulty, importance, and similarity. The remaining steps of task selection, Proponent review, assignment to testing mode, and test construction were carried out in FY85.

In addition, for Batch B a formal experimental procedure was used to determine the effects of scenario differences on SME judgment of task importance. The design called 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).

Training Achievement Tests (Batch A). During FY84, generation of refined task lists for each of the 19 MOS in the Project A sample continued. For each MOS in 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.

Training content SMEs and job content SMEs then 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.

Administrative (Archival) Indexes. 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 administrative indexes and compare it to actual 201 File information for the people in the field trials during FY85.

Predictor Measurement

During FY83, predictor development activities had been focused on comprehensive reviews of the literature (for cognitive, non-cognitive, and psychomotor measures respectively), visits to other personnel research laboratories, and consultations with designated experts in the field. The available literature was systematically catalogued on record forms designed for the project and the process of summarizing the information was begun.

The major activities completed during FY84 were:

- The definition and identification of the most promising predictor constructs.
- The administration and initial analysis of the Preliminary Battery.
- The development, tryout, and pilot testing of the first version of the Trial Battery, called the Pilot Trial Battery.
- The development and tryout of psychomotor/perceptual measures, using a microprocessor-driven testing device.

Each of these activities is briefly summarized below. A more complete description of new test development is presented in later sections of this report.

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 a final, prioritized list of constructs was identified.

This effort also produced a heuristic model based on factor analyses of the experts' judgments. This model 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 analysis is fully described in Wing, Peterson, and Hoffman (1984).

Preliminary Battery

Similarly, the initial analyses of Preliminary Battery data provided empirical results to guide development of Pilot Trial tests. Data were collected with the Preliminary Battery on four MOS: 05C (Fort Gordon), 19E/K (Fort Knox), 63B (Fort Dix and Fort Leonard Wood), and 71L (Fort Jackson).

The first 1,800 cases from a total sample of over 11,000 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 noncognitive inventories of temperament, interest, and biographical data. This effort is fully reported in Hough, Dunnette, Wing, Houston, and Peterson (1984).

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.

Computer-Administered 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. Several objectives were reached during 1984. An appropriate microprocessor was identified and six copies were obtained for development 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.

Data Base Management/Validation Analyses

During Project A's second year, the Longitudinal Research Data Base (LRDB) was expanded dramatically. 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 a proposal for improving the Army's existing procedures for selecting and classifying new recruits; the proposed improvements were adopted by the Army after thorough review and were implemented in the ASVAB at the beginning of FY85. A number of smaller research efforts were also supported with the expanded LRDB.

Growth of the LRDB

FY84 saw three major LRDB expansion activities:

- The enlargement 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.

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 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 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 because 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 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 MOS.

During FY84 we also accumulated archival data on training grades for soldiers in the four MOS to which the Preliminary Battery (PB) 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 I.2 shows the number of soldiers for whom training performance information is available, and the number of soldiers for whom both types of information are available.

Table I.2

FY83/84 Soldiers With Preliminary Battery and Training Data

MOS	Total PB Cases	Total[a] Training Cases	Cases With Both PB and Training Data: Total	Percent of PB Total	Percent of Training Total
05C/31C	2,411	1,951	833	35	43
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		

[a] As of FY84 year-end.

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 Concurrent Validation. The largest

accumulation of such information resulted from the Batch A combined criterion field test. The combined information led to more than 3,000 analysis variables for each of the 548 soldiers tested.

A second major field test effort during FY84 involved the Pilot Trial Battery. Scheduling conflicts postponed the data collection effort until very late in the fiscal year, so initial processing of these data had only begun by the end of FY84.

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 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 currently 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 ASVAB is composed of 10 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/Administration (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 in the Project A analyses as indexes of soldier performance were end-of-course training grades and SQT scores. While both have some limitations, they were the best available measures of soldier performance. These two criteria 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 phase of the research was the large size of the samples used in the analyses. 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.

The validities obtained in this research for the current nine AA composites are given in Figure I.6. 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 for the existing predictors are about as high as one is likely to find in measuring test validities.

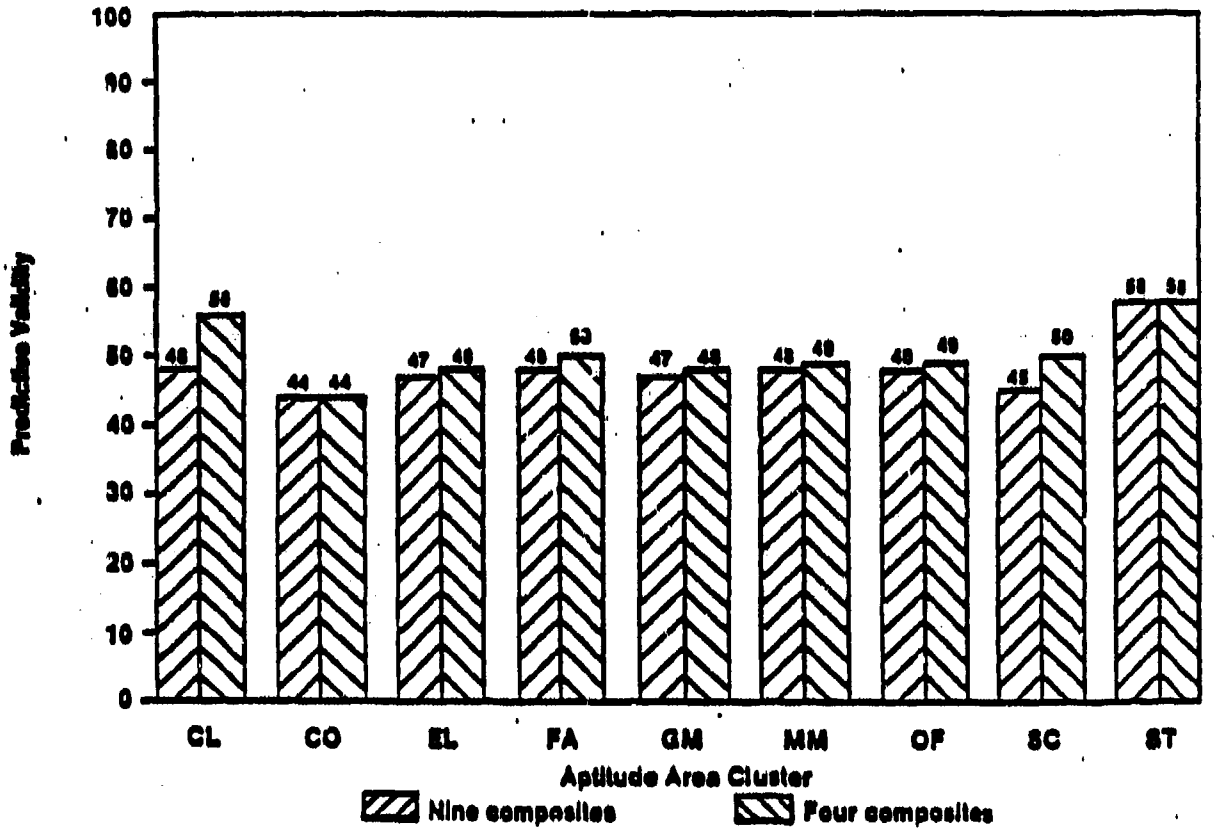


Figure I.6. Predictive validities systems for nine and four Aptitude Area composites.

A second finding 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 I.6 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 composites. Figure I.7 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% and of the SC composite by 11%.

	Current ASVAB Composite		Proposed Composite	
	Subtests	r	Subtests	r
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 I.7. A comparison of current and alternative Aptitude Area composites.

On the basis of these data, the Army decided to implement the proposed alternative composites for CL and SC, effective 1 October 1984.

A fuller discussion of the research entailed in the development and validation of the AA composites can be found in McLaughlin et al. (1984).

In Conclusion

By the end of FY84 the initial development and first pilot testing of all major predictor and criterion variables had been completed. This included the development of the first versions of the hands-on job samples and the computer-administered perceptual and psychomotor tests.

In addition, the formal field tests of the criterion measures were begun on the Batch A MOS. The predictor battery designated as the Pilot Trial Battery also underwent a more comprehensive pilot testing on approximately 650 soldiers.

Finally, by the end of FY84 the longitudinal research data base had been designed and the revalidation of the ASVAB using FY81/82 file data had been completed.

Section 4

FISCAL YEAR 1985

The third year of Project A was both resource and effort intensive. It was during FY85 that all predictor and criterion construction was completed, all field tests were completed, the field test results were analyzed, the final revisions (before validation) of the measures were made, and the Concurrent Validation data collection was begun.

Project Administration

During the third year's work, several changes were effective in the Governance Advisory Group. These changes are reflected in Figure I.8. There were also changes among the ARI task monitors and the consortium task leaders and other key personnel. The assignments for these positions at the end of FY85 are shown in Figure I.9.

Research Activities

A summary schedule of FY85 events is as follows:

1.	Completion of field tests for pilot trial predictor battery.	March 1985
2.	Analysis of predictor field test data.	October 1984-April 1985
3.	Revision of Pilot Trial Battery to form the Trial Battery.	May-June 1985
4.	Completion of Batch B criterion field tests.	March 1985
5.	Analysis of Batch A and Batch B field test data.	December 1984-June 1985
6.	Revision of criterion measures for use in Concurrent Validation.	May-June 1985
7.	Army proponent review of instruments used in Concurrent Validation.	May 1985
8.	Start of Concurrent Validation.	June 1985

Since the project's third year was such a crucial one and represented a culmination of a great deal of basic development work, it seems appropriate to use this FY85 report to summarize in some detail the first 3 years of Project A. Consequently, the major parts of Project A will be described, with the discussion emphasizing, but not limited to, the activities during FY85. This comprehensive presentation includes a discussion of how each phase of the research was conceptualized and initially formulated, as well as

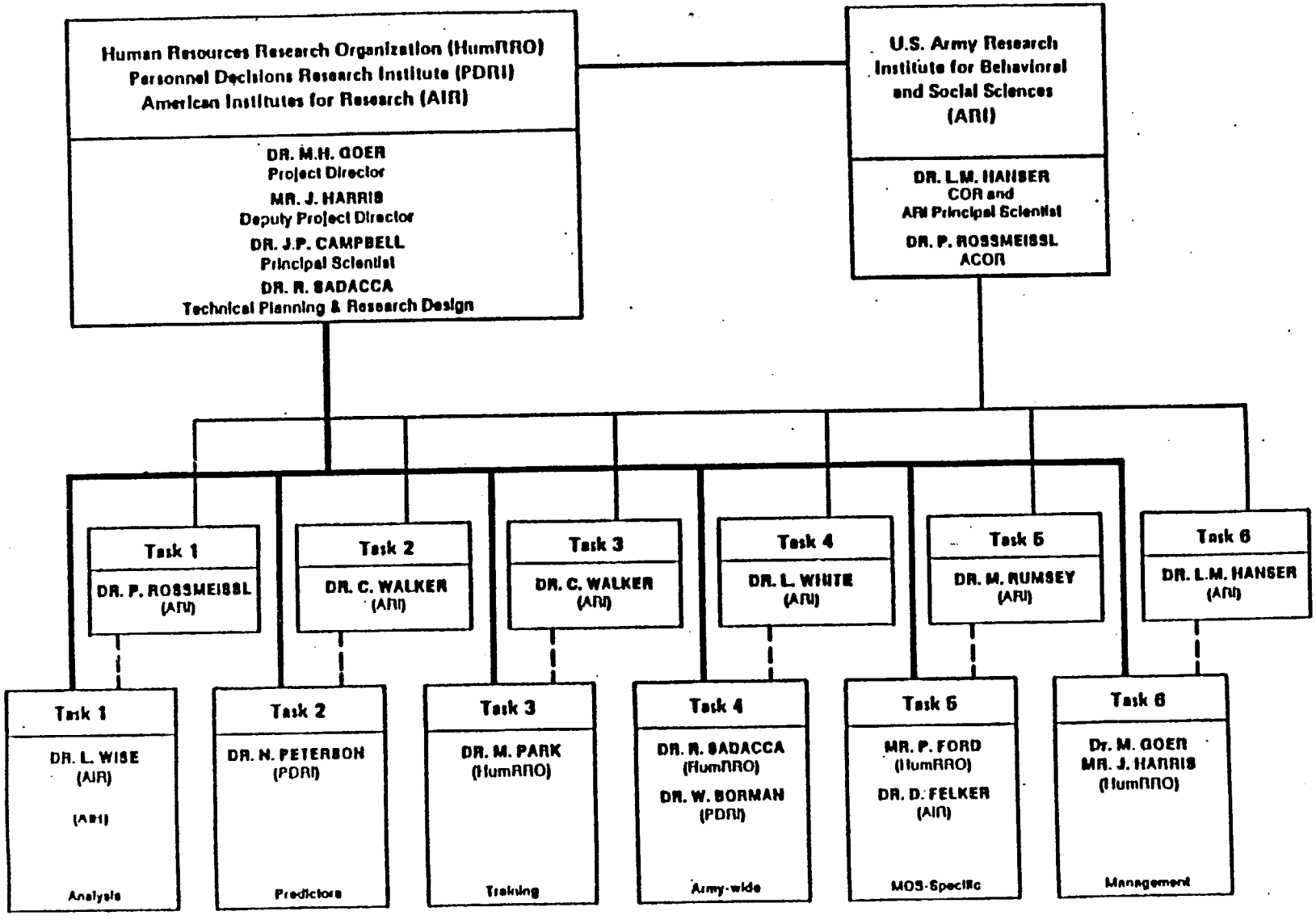


Figure I.8 Project A Management Group as of 30 September 1985.

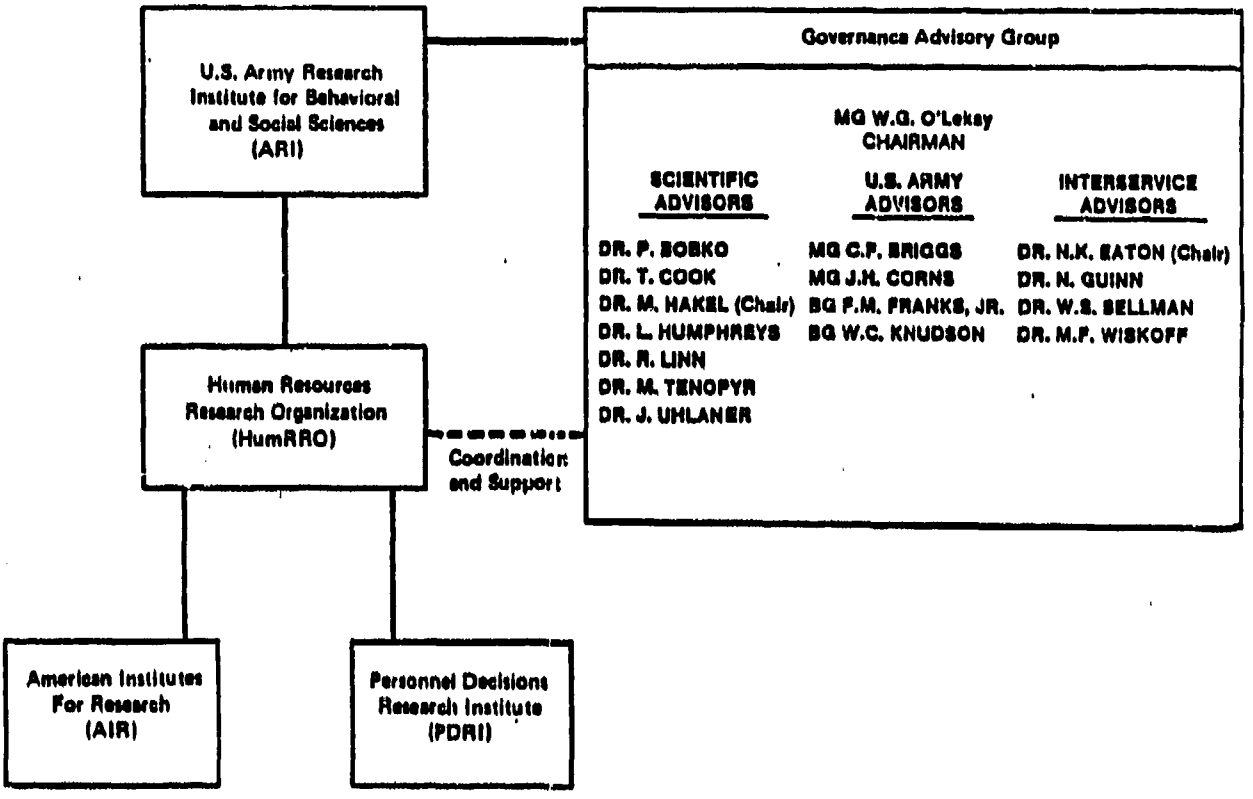


Figure I.9. Project A organization as of 30 September 1985.

a description of what was actually done. Bear with us as we go back to year one and pick up the story after the organization and staffing of the project had been agreed upon, the research plan had been completed and approved, and work had begun in earnest on each substantive task.

Organization of the Report

The remainder of this report is divided into two major sections; Part II describes the development and field test of the predictor battery, and Part III deals with the development and field test of the performance measures. By 30 May 1985 the final array of predictors and criteria to be used in the Concurrent Validation was agreed upon, and the validation collection began in June. The basic procedures employed in the Concurrent Validation are described in the last section of this report, Part IV.

This report is supplemented by an ARI Research Note (in preparation), which contains various technical papers prepared during FY85 in connection with specific aspects of the Project A research activities. These papers are listed in Appendix A of the present report.

PART II

PREDICTOR DEVELOPMENT

After discussion of the general issues involved in Project A's predictor development efforts, the specific development steps and initial pilot testing of each major predictor type will be described. After all predictors are discussed in turn, the full-scale field tests will be described and the revisions to the instruments made on the basis of the field tests will be outlined.

Section 1

INTRODUCTION TO PREDICTOR DEVELOPMENT[1]

This section describes the development, initial pilot testing, and field testing of the Trial Battery. The Trial Battery is the array of new enlisted selection/classification tests that are being evaluated in the Concurrent Validation sample. Again, the overall objective is to develop and validate tests that supplement the Armed Services Vocational Aptitude Battery (ASVAB) and broaden the domain of potential selection measures for U.S. Army first-tour enlisted personnel.

Project A has adopted a construct-oriented strategy of predictor development and endeavored to build a model of the predictor space by (a) identifying the major domains of constructs, (b) selecting measures within each domain that met a number of psychometric and pragmatic criteria, and (c) specifying those constructs that appeared to be the "best bets" for incrementing prediction of training/job performance and attrition/retention in Army jobs.

Ideally, the model would lead to the selection of a finite set of relatively independent predictor constructs that are also independent of present predictors and maximally related to the criteria of interest. If these conditions were met, then the resulting set of measures would yield valid prediction within each job, yet possess enough heterogeneity to yield valid classification of persons into different jobs.

Objective

This approach led to the delineation of a set of more concrete objectives:

1. Identify existing measures of human abilities, attributes, or characteristics that are most likely to be effective in predicting successful soldier performance and in classifying persons into MOS where they will be most successful, with special emphasis on attributes not tapped by current pre-enlistment measures.

---

[1]Part II is based primarily on ARI Technical Report 739, Development and Field Test of the Trial Battery for Project A, Norman Peterson, Editor, and a supplementary ARI Research Note (in preparation), which contains the report appendixes that present the tests used in the Pilot Trial Battery and Trial Battery administration. Authors of various portions of this report include Norman Peterson, Jody Toquam, Leaetta Hough, Janis Houston, Rodney Rosse, Jeffrey McHenry, Teresa Russell, VyVy Corpe, Matthew McGue, Bruce Barge, Marvin Dunnette, John Kamp, and Mary Ann Hanson.

2. Where appropriate, design and develop new measures or modify existing measures of these "best bet" predictors.

3. Estimate and evaluate the reliability of the new pre-enlistment measures and their vulnerability to motivational set differences, faking, variances in administrative settings, and practice effects.

4. Determine the interrelationships (or covariance) between the new pre-enlistment measures and current pre-enlistment measures.

5. Determine the degree to which the validity of new pre-enlistment measures generalizes across Military Occupational Specialties (MOS), that is, proves useful for predicting measures of successful soldier performance across quite different MOS, and, conversely, the degree to which the measures are useful for classification or the differential prediction of success across MOS.

6. Determine the extent to which new pre-enlistment measures increase the accuracy of prediction of success and the accuracy of classification into MOS over and above the levels of accuracy reached by current pre-enlistment measures.

General Research Design and Organization

To achieve these objectives, we have followed the design depicted in Figure II.1. Several things are noteworthy about the 15 subtasks in the research plan. First, five test batteries are mentioned: Preliminary Battery, Demonstration Computer Battery, Pilot Trial Battery, Trial Battery, and Experimental Battery. These appear in sequence, a schedule that allows us to improve the predictors as data are gathered and analyzed on each successive battery or set of measures. Second, a large-scale literature review and an expert judgment procedure were utilized early in the project to take maximum advantage of previous research and accumulated expert knowledge. The expert judgments were used early on to develop an initial model of both the predictor space and the criterion space, which also relied heavily on the information gained from the literature review. Third, the design includes both predictive and concurrent validation designs.

The project staff were organized into three "domain teams." One team concerned itself with temperament, biographical, and vocational interest variables and came to be called the "non-cognitive" team. Another team examined cognitive and perceptual variables and was called the "cognitive" team. The third team concentrated on psychomotor and perceptual variables and was labeled the "psychomotor" team or sometimes the "computerized" team, since all the measures developed by that team were computer-administered.

We turn now to a description of the initial research activities devoted to development of new predictors, specifically: the literature review; expert judgments; development, administration, and analysis of the Preliminary Battery; and initial development of a computer battery. As Figure II.1 shows, all of the these activities led up to a development of the Pilot Trial Battery.

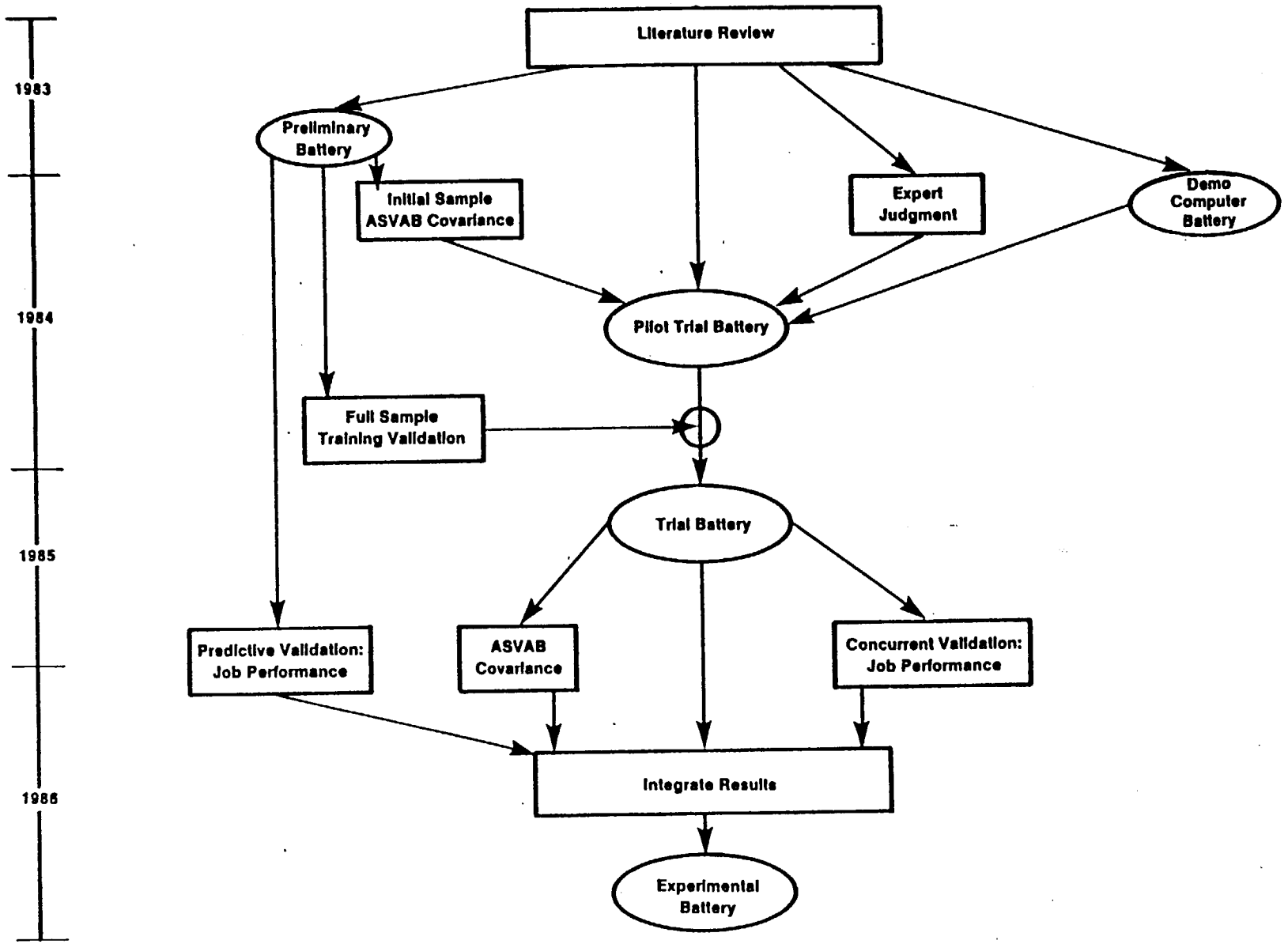


Figure II.1. Flow chart of predictor measure development activities of Project A.

Literature Review

The overriding purpose of the literature review was to gain maximum benefit from earlier research on selection/classification measures that were even remotely relevant for the jobs in the Project A job population.

Search Procedures

The search was conducted by the three research teams, each responsible for the broadly defined area of human abilities or characteristics mentioned previously. These areas, or domains, proved to be convenient for purposes of organizing and conducting literature search activities, but were not used as (nor intended to be) a final taxonomy of possible predictor measures.

The literature search was conducted in late 1982 and early 1983 (i.e., FY83). Within each of the three areas, the teams carried out essentially the same steps:

1. Compile an exhaustive list of potentially relevant reports, articles, books, or other sources.

2. Review each source and determine its relevancy for the project by examining the title and abstract (or other brief review).

3. Obtain the sources identified as relevant in the second step.

4. For relevant materials, carry out a thorough review and transfer relevant information onto summary forms specifically developed for the project.

Within Step 1, several computerized searches of relevant data bases were generated. Across all three ability areas, over 10,000 sources were identified via the computer search. Many of these sources were identified as relevant in more than one area, and were thus counted more than once.

In addition to the computerized searches, we solicited reference lists from recognized experts in each of the areas, obtained several annotated bibliographies from military research laboratories, and scanned the last several years' editions of relevant research journals as well as more general sources such as textbooks, handbooks, and appropriate chapters in the Annual Review of Psychology.

The vast majority of the references identified in Step 1 were not relevant to Project A and were eliminated in Step 2. The references identified in Step 2 were obtained and reviewed, and two forms were completed for each source: an Article Review form and a Predictor Review form (several of the latter could be completed for each source). These forms were designed to capture, in a standard format, the essential information, which varied considerably in organization and reporting style in the original documents.

The Article Review form contained seven sections: citations, abstract, list of predictors (keyed to the Predictor Review forms), description of criterion measures, description of sample(s), description of methodology,

other results, and reviewer's comments. The Predictor Review form also contained seven sections: description of predictor, reliability, norms/descriptive statistics, correlations with other predictors, correlations with criteria, adverse impact/differential validity/test fairness, and reviewer's recommendations (about the usefulness of the predictor). Each predictor was tentatively classified into an initial, working taxonomy of predictor constructs (based primarily on the taxonomy described in Peterson and Bownas, 1982).

Literature Search Results

The literature search was used in two major ways. First, three working documents were written, one for each of the three areas: cognitive/perceptual abilities, psychomotor/perceptual abilities, and non-cognitive predictors (including temperament or personality, vocational interest, and biographical data variables). These documents summarized the literature with regard to critical issues, suggested the most appropriate organization or taxonomy of the constructs in each area, and summarized the validities of the various measures for different types of job performance criteria. Second, the predictors identified in the review were subjected to further scrutiny to (a) select tests and inventories to make up the Preliminary Battery, and (b) select the "best bet" predictor constructs to be used in the "expert judgment" research activity. We turn now to a description of that screening process.

Screening of Predictors

An initial list was compiled of all predictor measures that seemed even remotely appropriate for Army selection and classification. This list was further screened by eliminating measures according to several "knockout" factors: (a) measures developed for a single research project; (b) measures designed for a narrowly specified population/occupational group (e.g., pharmacy students); (c) measures targeted toward younger age groups; (d) measures requiring special apparatus for administration; (e) measures requiring unusually long testing times; (f) measures requiring difficult or subjective scoring; and (g) measures requiring individual administration.

Knockout factor (d) was applicable only with regard to screening for the Preliminary Battery, which could not have any computerized tests or other apparatus since it was to be administered early in the project, before such testing devices could be developed. Factor (d) was <u>not</u> applied with regard to screening measures for inclusion in the expert judgment process.

The result of the application of knockout factors was a second list of candidate measures. Each of these measures was evaluated, by at least two researchers, on the 12 factors shown in Figure II.2. (A 5-point rating scale was applied to each of the 12 factors.) Discrepancies in ratings were resolved by discussion. There was not always sufficient information for a variable to allow a rating on all factors.

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 as a measure 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 studies, experimental studies, 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 II.2. Factors used to evaluate predictor measures for the Preliminary Battery.

This second list of measures, each with a set of evaluations, was input to (a) the final selection of measures for the Preliminary Battery and (b) the final selection of constructs to be included in the expert judgment process.

Expert Forecasts of Predictor Construct Validities

The procedure used in the expert judgment process was to (a) identify criterion categories, (b) identify an exhaustive range of psychological constructs that may be potentially valid predictors of those criterion categories, and (c) 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 of hundreds of subjects. That is, experienced personnel psychologists are effective "validity generalizers" for cognitive tests, although they do tend to underestimate slightly the true validity as obtained from empirical research.

Consequently, 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. Peterson and Bownas (1982) provide a complete description of the methodology which has been used successfully by Bownas and Heckman (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 criterion variables is given to "experts" in personnel selection and classification, typically personnel psychologists. These experts estimate 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 is 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. Those variances can then be factor analyzed to identify clusters of predictors within which the measures are expected to exhibit similar patterns of correlations with different performance components. Similarly, the criterion covariances can be examined to identify clusters of criteria predicted by a common set of predictors.

Such procedures helped in identifying redundancies and overlap in the predictor set. The 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 form permits easier comparison with the results of meta-analyses of empirical estimates of criterion-related validity coefficients. Third, these clusters provide a model or theory of the predictor-criterion performance space.

Method

To carry out the expert judgments, a sample of subject matter experts (SMEs) was selected, a universe of predictor variables and a universe of criterion variables were identified, and materials that would allow the experts to provide reliable estimates of criterion-related validity were prepared.

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, Human Resources Research Organization, Personnel Decisions Research Institute, and American Institutes for Research. Not all of the employees were directly involved with Project A although all of the consultants were.

Identification of Predictor Variables. The predictor variables evaluated with regard to the 12 relevant factors (see Screening of Predictors, above) were used in the expert judgment process. Variables were included if they received generally high evaluations and if they added to the comprehensiveness of coverage for a particular domain of predictor variables. The names and definitions of these variables are shown in Appendix C of ARI Technical Report 739 noted previously.

Materials describing each of the 53 variables were prepared. 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 which 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.

Identification of Criterion Variables. Several types of criterion variables were identified. The first type was a set of specific job task categories. Short of enumerating all job tasks in the nearly 240 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.

The procedure used was based on more general job descriptions of a representative sample of 111 jobs that had been previously clustered by job experts as part of the MOS sample selection described in the introduction to this volume. 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 activities (e.g., transcribe, annotate, sort, index, file, retrieve) lead 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 categories were generated. Most of these categories applied to several jobs, and most of the jobs were characterized by activities from several categories.

The second type of criterion variable was a set that described performance in initial Army training. Two sources of information were used to identify appropriate training performance variables: archival records of soldiers' performance in training, and interviews with trainers. This information was obtained for eight MOS: Radio/Teletype Operator, MANPADS Crewman, Light Vehicle/Power Generator Mechanic, Motor Transport Operator, Food Service Specialist, M60 and M1 Armor Crewman, 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.

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.

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 information oriented to behavior.

After the interviews were conducted and the archives examined, information from both sources was pooled and categorized. Since the task or MOS-specific performance variance was already covered elsewhere, four variables were used to represent training performance. Their names and definitions are shown in Appendix C in ARI Technical Report 739.

The final type of criterion variable was a set of general performance categories developed as part of Task 4 work. Nine behavioral dimensions were named and defined. In the final step, six more criterion variables were added. 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 that did not emerge elsewhere. The last four are all important "outcome" criterion variables; that is, they represent outcomes of individual behavior that have negative or positive value to the Army (e.g., disciplinary actions), but the outcomes could occur because of a variety of individual behaviors.

In all, then, 72 possible criterion constructs were identified and defined for use in the expert judgment task.

Instructions and Procedures. Detailed instructions were provided for each judge along with the materials describing the predictor and criterion

variables. First, each judge was provided with information about the concepts of "true validity," criterion-related validity corrected for such artifacts as range restriction and reliability, and unaffected by variation in sample sizes. Judges were asked to make estimates of the level of true validity on a 9-point scale. A rating of "1" meant a true validity in the range of .00 to .10; "2," .11 to .20; and so forth, to "9," .81 to .90.

Second, descriptions of the 53 predictor variables were placed into three 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 the rating task.

Third, each judge estimated the validity of each predictor for each criterion. The order of the predictor groups (A, B, C) was counterbalanced across judges, so 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).

Ratings were made on separate Judgment Record Sheets. Before making any judgments about a predictor, the expert was to read the descriptive information and review the examples of items measuring 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. The validities of the first predictor variable were to be estimated for all 72 criteria before the judge moved on to the next predictor.

All judges completed the task during the first week of October 1983.

Results

A number of analyses were carried out: reliability of the judgments, means and standard deviations of the estimated validities within each predictor/criterion cell and for various marginal values, and factor analyses of the predictors (based on their validity profiles across the criteria) and the criteria (based on their validity profiles across the predictors).

The estimated validities were highly reliable when averaged across raters. The reliability of the mean estimated cell validities was .96. The factor analyses were based on these cell means. The most pertinent analysis for purposes of this report concerns the factor analysis of the predictors.

Factor solutions with 2 through 24 factors were calculated; eigen values diminished below 1.0 after 9 or 10 factors. No more than eight factors were interpretable, so the eight-factor solution was selected as most reasonable. The eight interpretable factors were named: I, Cognitive Abilities; II, Visualization/Spatial; III, Information Processing; IV, Mechanical; V, Psychomotor; VI, Social Skills; VII, Vigor; VIII, Motivation/Stability.

These eight factors appeared to be composed of 21 clusters, based on the profile of loadings of each predictor variable across factors. This

hierarchical structure of the predictor variables is shown in Figure II.3. Inspection of the profile clarifies the meanings of both 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 non-cognitive 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.

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 measuring 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 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.

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, non-cognitive 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 that refer to general activity level. The first, Athletic Abilities/Energy, includes two constructs which point toward a

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	N. 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	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental Information Processing	INFORMATION PROCESSING
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests 51. Artistic Interests (Negative)	M. 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 Interests	
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 46. Non-delinquency 53. Conventional Interests	N. Traditional Values/Conventionality/Non-delinquency	
44. Locus of Control 47. Work Orientation	O. Work Orientation/Locus of Control	MOTIVATION/ STABILITY
38. Cooperativeness 45. Emotional Stability	P. Cooperation/Emotional Stability	

Figure II.3. Hierarchical map of predictor space.

physical perspective while the second, Dominance/Self-Esteem, points toward 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 toward the individual vis-a-vis his or 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.

The expert judgment task thus resulted in a hierarchical model of predictor space that served as a guide for the development of new, pre-enlistment measures (the Pilot Trial Battery) for Army enlisted ranks. (Wing, Peterson, & Hoffman, 1984, provide a detailed presentation of the expert judgment process and results.) However, this model was not the only information that guided the development of the Pilot Trial Battery, and we turn now to the other major source of guidance, our experience in the development of the Preliminary Battery.

Development and Administration of the Preliminary Battery

The Preliminary Battery (PB) was a set of proven "off-the-shelf" measures intended to overlap very little with the Army's current pre-enlistment predictors. The collection of data on a number of predictors that represent types of predictors not currently in use by the Army would allow an early determination of the extent to which such predictors contributed unique variance. Also, the collection of predictor data (from soldiers in training) early in the project allowed an assessment of predictive validity much earlier than if we waited until the trial Battery was developed (see Figure II.1). Some of the Preliminary Battery measures were also included in the pilot tests of the Trial Battery as marker variables.

Selection of Preliminary Battery Measures

As described earlier, the literature review identified a large set of predictor measures, each with ratings by the researchers on 12 psychometric and substantive evaluation factors (see Figure II.2). These ratings were used to select a smaller set of measures as serious candidates for inclusion in the Preliminary Battery. Two major practical constraints came into play: (a) No apparatus or individualized testing methods could be used because the time available to prepare for battery administration was relatively short, and because the battery would be administered to a large number of soldiers (several thousand) over a 9-month period by relatively unsophisticated test administrators; and (b) only 4 hours were available for testing.

The research staff made an initial selection of "off-the-shelf" measures, but there were still too many measures for the time available. The preliminary list was presented to a joint meeting of the ARI and consortium research staffs, and the available information about each measure was presented and discussed. A final set of measures was selected at this meeting, subject to review by several external (to Project A) consultants who had been retained for their expertise in various predictor domains. Subsequently,

these experts reviewed the selected measures and made several "fine-tuning" suggestions.

The Preliminary Battery included the following:

- Eight perceptual-cognitive measures
  - Five from the Educational Testing Service (ETS) French Kit (Ekstrom, French, & Harman, 1976)
  - Two from the Employee Aptitude Survey (EAS) (Ruch & Ruch, 1980)
  - One from the Flanagan Industrial Tests (FIT) (Flanagan, 1965)
- Eighteen 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 (DPQ) (Tellegen, 1982)
  - One from the California Psychological Inventory (CPI) (Gough, 1975)
  - The Rotter I/E scale (Rotter, 1966)
  - Validity scales from both the DPQ and the Personality Research Form (PRF) (Jackson, 1967)
- Owens' Biographical Questionnaire (BQ) (Owens and Schoenfeldt, 1979)

  The BQ could be scored for either 11 scales for males or 14 for females based on Owens' 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, unlike some of Owens' scales. Items tapping religious or socio-economic status were deleted from Owens' instrument for this use, and items tapping physical fitness and vocational-technical course work were added.

Appendix D in ARI Technical Report 739 shows all the scale names and number of items for the Preliminary Battery.

In addition to the Preliminary Battery, scores were available for the Armed Services Vocational Aptitude Battery, which all soldiers take prior to entry into service. ASVAB's 10 subtests are named below, with the test acronym and number of items in parentheses:

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 are considered to be power tests except for NO and CS, which are speeded. Prior research (in Kass, Mitchell, Grafton, & Wing, 1983) has shown the reliability of the subtests to be within expectable limits for cognitive tests of this length (i.e., .78 - .92).

Sample and Procedure

The Preliminary Battery was administered to soldiers entering Advanced Individual Training (AIT) for four MOS: 05C, Radio Teletype Operator (MOS code was later changed to 31C); 19 E/K, Tank Crewman; 63B, Vehicle and Generator Mechanic; and 71L, Administrative Specialist. Almost all soldiers entering AIT for these MOS during the period 1 October 1983 to 30 June 1984 completed the Preliminary Battery. We are here concerned only with the sample of soldiers who completed the battery between 1 October 1983 and 1 December 1983, approximately 2,200 soldiers.

The battery was administered at five training posts by civilian or military staff already employed on site. Task 2 staff traveled to these sites to deliver battery administration manuals and to train the persons who would administer the battery. Before its implementation, the Preliminary Battery was administered to a sample of 40 soldiers at Fort Leonard Wood to test the instructions, timing, and other administration procedures. The results of this tryout were used to adjust the procedures, prepare the manual, and identify topics to be emphasized during administrator training.

Analyses

An initial set of analyses was performed to inform the development of the Pilot Trial Battery and we summarize those findings here. They are more completely reported in Hough, Dunnette, Wing, Houston, and Peterson (1984).

Three types of analyses were done. First, the psychometric characteristics of each scale were explored. These analyses included descriptive statistics, item analyses (including numbers of items attempted in the time allowed), internal consistency reliability estimates, and, for the temperament inventory, percentage of subjects failing the scales intended to detect random or improbable response patterns. Second, the covariances of the scales within the various conceptual domains (i.e., cognitive, temperament, biographical data, and vocational interest) and across these domains were investigated. Third, the covariances of the Preliminary Battery scales with ASVAB measures were investigated to identify any PB constructs that showed excessive redundancy with ASVAB constructs.

The psychometric analyses showed some problems with the cognitive tests. The time limits appeared too stringent for some tests, and one test, Hidden Figures, appeared to be much too difficult for the population being

tested. The lesson learned was that the Pilot Trial Battery measures should be more accurately targeted (in terms of difficulty of items and time limits) toward the population of persons seeking entry into the U.S. Army.

No serious problems were unearthed with regard to the temperament, biodata, and interest scales. Item-total correlations were acceptably high and in accordance with prior findings and score distributions were not excessively skewed or different from expectation. About 8% of the respondents failed the scale that screened for inattentive or random responding on the temperament inventory, a figure that is in accord with findings on other selection research.

Covariance analyses showed that vocational interest scales were relatively distinct from the biographical and temperament scales, but the latter two types of scales showed considerable covariance. Five factors were identified from the 40 non-cognitive scales, two that were primarily vocational interests and three that were combinations of biographical data and temperament scales. These findings led us to consider, for the Pilot Trial Battery, combining biographical and temperament item types to measure the constructs in these two areas. The five non-cognitive factors showed relative independence from the cognitive PB tests, with the median absolute correlations of the scales within each of the five factors with each of the eight PB cognitive tests ranging from .01 to .21. This confirmed our expectations of little or no overlap between the cognitive and non-cognitive constructs.

Correlations and factor analysis of the 10 ASVAB subtests and the eight PB cognitive tests confirmed prior analyses of the ASVAB (Kass et al., 1983) and the relative independence of the PB tests. Although some of the ASVAB-PB test correlations were fairly high (the highest was .57), most were less than .30 (49 of the 80 correlations were .30 or less, 65 were .40 or less). The factor analysis (principal factors extraction, varimax rotation) of the 18 tests showed all eight PB cognitive tests loading highest on that factor. The non-cognitive scales overlapped very little with the four ASVAP factors identified in the factor analysis of the ASVAB subtests and PB cognitive tests. Median correlations of non-cognitive scales with the ASVAB factors, computed within the five non-cognitive factors, ranged from .03 to .32, but 14 of the 20 median correlations were .10 or less.

The experience in training battery administrators and monitoring the administration over the 9-month period provided useful information for collecting data later with the Pilot Trial Battery and Trial Battery.

Initial Computer-Administered Battery Development

Because computerized testing was a new area of test development, the initial phase is given special attention here. The measures are described in more detail in a later subsection. There were four phases of activities: (a) information gathering about past and current research in the area of perceptual/psychomotor measurement and computerized methods of testing such abilities; (b) construction of a demonstration computer battery; (c) selection of commercially available microprocessors and peripheral devices, writing of software for testing several abilities using this hardware, and tryout

of this hardware and software; (d) continued development of software, and the design and construction of a custom-made peripheral device, which is now called a response pedestal.

Compared to the paper-and-pencil measurement of cognitive abilities and the major non-cognitive variables, computerized measurement of psychomotor and perceptual abilities was in a relatively primitive state. Much work had been done in World War II using electro-mechanical apparatus, but relatively little work had occurred since then. Microprocessor technology held out the promise of improving measurement in this area, but the work was (and still is) in its early stages.

Phase 1: Information Gathering

While almost no literature was available on computer-administered (especially microprocessor-driven) testing of psychomotor/perceptual abilities for selection/classification purposes, there was considerable literature available on the taxonomy or structure of such abilities, based primarily on work done in World War II or shortly thereafter. Also, work from this era showed that testing such abilities with electro-mechanical apparatus did produce useful levels of validity for such jobs as aircraft pilot, but that such apparatus experienced reliability problems.

To obtain the most current information, in the spring of 1983 we visited four military laboratories engaged in relevant research. The four sites visited were the Air Force Human Resources Laboratory, Brooks Air Force Base; the Naval Aerospace Medical Research Laboratory, Pensacola Naval Station; and the Army Research Institute Field Units at Fort Knox, Kentucky, and Fort Rucker, Alabama. During these site visits we gathered much information, but focused primarily on the answers to five questions:

1. What computerized measures are actually in use?

   Over 60 different measures were found across the four sites. A sizable number of these were specialized simulators that were not relevant for Project A (e.g., a helicopter simulator weighing several tons that is permanently mounted in an air-conditioned building). However, there were many measures in the perceptual, cognitive, and psychomotor areas that were relevant.

2. What computers were selected for use?
3. What computer languages are being used?

   Three different microprocessors (Apple, Terak, and PDP 11) and three different computer languages (PASCAL, BASIC, and FORTRAN) appeared to account for most of the activity. However, there appeared to be relatively little in common among the four sites.

4. How reliable are these computerized measures?
5. What criterion-related validity evidence exists for these measures so far?

Data were being collected at all four sites to address the reliability and criterion-related validity questions, but very little documented information was available. This was not surprising in light of the fact that most of the measures had been developed only very recently.

Despite the lack of evidence on reliability and validity, we did learn some valuable lessons. First, large-scale testing could be carried out on microprocessor equipment (AFHRL was doing so). Second, a variety of software and hardware could produce satisfactory results. Third, it would be highly desirable to have the testing devices or apparatus be as compact and simple in design as possible to minimize "down time" and make transportation feasible. Fourth, it would be highly desirable to develop our software and hardware devices to be as completely self-administering (i.e., little or no input required from test monitors) as possible and as impervious as possible to prior experience with typewriting and playing video games.

Phase 2: Demonstration Battery

After these site visits, a short demonstration battery was programmed on the Osborne 1, a portable microprocessor. This short battery was self-administering, recorded time-to-answer and the answer made, and contained five tests: simple reaction time, choice reaction time, perceptual speed and accuracy (comparing two alphanumeric phrases for similarity), verbal comprehension, and a self-rating form (indicating which of two adjectives "best" describes the examinee, on a relative 7-point scale). We also experimented with the programming of several types of visual tracking tests, but did not include these in the self-administered demonstration battery.

No data were collected, but experience in developing and using the battery convinced us that BASIC did not allow enough power and control of timing to be useful for our purposes. The basic methods for controlling stimulus presentation and response acquisition through a keyboard were thoroughly explored. Techniques for developing a self-administering battery of tests were tried out.

The second activity during this phase was consultation with three experts at the University of Illinois about perceptual/psychomotor abilities and their measurement.[2] The major points were:

- The results obtained in World War II using electro-mechanical, psychomotor testing apparatus probably do generalize to the present era in terms of the structure of abilities and the usefulness of such abilities for predicting job performance in jobs like aircraft pilot.

---

[2]Charles Hulin, John Adams, and Phillip Ackerman.

- The taxonomy of psychomotor skills and abilities probably should be viewed in a hierarchical fashion, and perhaps Project A's development efforts would be best focused on two or three relatively high-level abilities such as gross motor coordination, multilimb constant processing tasks, and fine manipulative dexterity.

- Rate of learning or practice effects are viewed as a major concern. If later test performance was more valid than early test performance, or if early test performance was not valid at all and later test performance was, then it was unlikely that psychomotor testing would be practically feasible in the operational military selection environment. There were, however, no empirically based answers to these questions, and it was acknowledged that research is necessary to answer them.

Phase 3: Selection and Purchase of Microprocessors and Development and Try-out of Software

On the basis of information from the first two phases, we defined the desirable characteristics of a microprocessor useful for Project A. Desired characteristics, as outlined in the fall of 1983 were:

1. Reliability--The machine should be manufactured and maintained by a company that has a proven record and the machine itself should be capable of being moved from place to place without breaking down.

2. Portability--The computer must be easily moved between posts during development efforts.

3. Most Recent Generation of Machine--Progress is very rapid in this area; therefore, we should get the latest "proven" type of machine.

4. Compatibility--Although extremely difficult to achieve, a desirable goal is to use a machine that is maximally compatible with other machines.

5. Appropriate Display Size, Memory Size, Disk Drives, Graphics, and Peripheral Capabilities--We need a video display that is at least 9 inches (diagonally), but it need not be color. Since we will be developing experimental software, we need a relatively large amount of random access memory. Also we require two floppy disk drives to store needed software and to record subjects' responses. High-resolution graphics capability is desirable for some of the kinds of tests. Finally, since several of the ability measurement processes will require the use of paddles, joysticks, or other similar devices, the machine must have appropriate hardware and software to allow such peripherals.

In the end we selected the Compaq portable microprocessor with 256K RAM, two 320K disk drives, a "game board" for accepting input from peripheral devices such as joysticks, and software for FORTRAN, PASCAL, BASIC, and assembly language programming. Six of these machines were purchased in December 1983. We also purchased six commercially available, dual-axis joysticks.

We chose to prepare the bulk of the software using PASCAL as implemented by Microsoft, Inc. PASCAL software is implemented using a compiler that permits modularized software development, it is relatively easy for others to read, and it can be implemented on a variety of computers.

Some processes, mostly those that are specific to the hardware configuration, had to be written in IBM-PC assembly language. Examples include interpretation of the peripheral device inputs, reading of the real-time-clock registers, calibrated timing loops, and specialized graphics and screen manipulation routines. For each of these identified functions, a PASCAL-callable "primitive" routine with a unitary purpose was written in assembly language. Although the machine-specific code would be useless on a different type of machine, the functions were sufficiently simple and unitary in purpose so that they could be reproduced with relative ease.

The overall strategy of the software development was to take advantage of each researcher's input as directly as possible. It quickly became clear that the direct programming of every item in every test by one person (a programmer) was not going to be successful in terms of either time constraints or quality of product. To make it possible for each researcher to contribute her or his judgment and effort to the project, it was necessary to plan to, as much as possible, take the "programmer" out of the step between conception and product.

The testing software modules were designed as "command processors" which interpreted relatively simple and problem-oriented commands. These were organized in ordinary text written by the various researchers using word processors. Many of the commands were common across all tests. For instance, there were commands that permitted writing specified text to "windows" on the screen and controlling the screen attributes (brightness, background shade, etc.). A command could hold a display on the screen for a period of time measured to 1/100th-second accuracy. There were commands that caused the programs to wait for the respondent to push a particular button. Other commands caused the cursor to disappear or the screen to go blank during the construction of a complex display.

Some of the commands were specific to particular item types. These commands were selected and programmed according to the needs of a particular test type. For each item type, we decided upon the relevant stimulus properties to vary and built a command that would allow the item writer to quickly construct a set of commands for items which he or she could then inspect on the screen.

These techniques made it possible for entire tests to be constructed and experimentally manipulated by psychologists who could not program a computer.

As this software was written, we used it to administer the computerized tests to small groups of soldiers (N = 5 or fewer) at the Minneapolis Military Entrance Processing Station (MEPS). The soldiers completed the battery without assistance from the researchers, unless help was absolutely necessary, and were then questioned. The nature of the questions varied over the progress of these developmental tryouts, but mainly dealt with clarity of

instructions, difficulty of tests or test items, screen brightness problems, difficulties in using keyboard or joysticks, clarity of visual displays, and their general (favorable/unfavorable) reaction to this type of testing.

These tryouts were held from 20 January 1984 through 1 March 1984, and a total of 42 persons participated in nine separate sessions. The feedback received from the participants was extremely useful in determining the shape of the test, prior to the first pilot test of the Pilot Trial Battery.

Phase 4: Continued Software Development and Design/Construction of a Response Pedestal

By the end of Phase 3, we had developed a self-administering, computerized test battery that was implemented on a Compaq portable computer. The subjects responded on the normal keyboard for all tests except for a tracking test which required them to use a joystick, a commercially available device normally used for video games. Seven different tests had been programmed for the battery.

During the fourth phase of development, several significant events occurred. We made field observations of some combat MOS to obtain information for further development of computerized tests; the first pilot test of the computerized battery was completed; we designed and constructed a custom-made response pedestal for the computerized battery; and a formal review of progress was conducted.

The primary result of the review was the identification and priority-setting of the ability constructs for which computerized tests should be developed. A second result was a decision to go to the field to observe several combat arms MOS to target the tests more closely to those skills.

These field observations subsequently took place at several posts. In addition to observing soldiers in the field, we operated various training aids and simulators that were available during our visits. The MOS for which we were able to complete these observations were 11B (Infantryman), 13B (Cannon Crewman), 19K (Tank Crewman), 16S (MANPADS Crewman), and 05C (Radio/Teletype Operator).

The first pilot test of the Pilot Trial Battery occurred at Fort Carson during this phase. (See Section 2 for a description of the sample and procedures of that pilot test.) With regard to the computerized tests, the same procedures were used as for the MEPS tryouts in Phase 3.

A total of 20 soldiers completed the computerized battery. The information obtained at this pilot test primarily confirmed a major concern that had surfaced during the MEPS tryouts--namely, the undesirability of the computer keyboard and commercially available joysticks for acquiring responses. Feedback from subjects (and observation of their test taking) indicated that it was difficult to pick out one or two keys on the keyboard, and that rather elaborate, and therefore confusing, instructions were needed to use the keyboard in this manner. Even with such instructions, subjects frequently missed the appropriate key, or inadvertently pressed keys because they were leaving their fingers on the keys in order to retain the appropriate position

for response. Also, there was variability in the way subjects prepared for test items, and more or less random positioning of their hands added unwanted (error) variance to their scores. Similar issues arose with regard to the joysticks, but the main problems were their lack of durability and the large variance in their operating characteristics.

After consultation with ARI and other Project A researchers, Task 2 staff decided to develop a custom-made response pedestal to alleviate these problems as much as possible. Accordingly, we drew up a rough design for such a pedestal and contracted with an engineering firm to fabricate a prototype. We tried out the first prototype, suggested modifications, and had six copies produced in time for the Fort Lewis pilot test in June 1984.

Finally, we wrote additional software to test the abilities that had been chosen for inclusion in the Pilot Trial Battery and to accommodate the new response pedestal.

Identification of Pilot Trial Battery Measures

In March 1984, a formal In Progress Review (IPR) meeting was held to decide on the measures to be developed for the Pilot Trial Battery. Information from the literature review, expert judgments, initial analyses of the Preliminary Battery, and the first three phases of computer battery development was presented and discussed. Task 2 staff made recommendations for inclusions of measures and these were evaluated and revised. Figure II.4 shows the results of that deliberation process.

This set of recommendations constitutes the initial array of predictor variables for which measures would be constructed and then submitted to a series of pilot tests and field tests, with revisions being made after each phase. The specific measures, the steps in their construction, and their final form after pilot and field testing are described in later sections of Part II.

Pilot Tests and Field Tests of the Pilot Trial Battery

There were three pilot tests of the measures developed for the Pilot Trial Battery. These took place in Fort Carson in April 1984, Fort Campbell in May 1984, and Fort Lewis in June 1984. At the first two sites not all Pilot Trial Battery measures were administered, but the complete battery was administered at Fort Lewis. Sections 2, 3, 4, and 5 describe these pilot tests, resulting analyses, and revisions to measures prior to the field tests. The reports of data analyses emphasize the Fort Lewis administration since it was the first time the complete battery was administered and provided the largest pilot test sample. (The pilot tests are sometimes referred to as "tryouts" in the remainder of this report.)

There were three field tests of the Pilot Trial Battery. These occurred at Fort Knox, Fort Bragg, and the Minneapolis MEPS in Fall, 1984. These field tests, as well as the resulting revisions of the Pilot Trial Battery, are described in Section 6.

Final Priority*	Predictor Category	Pilot Trial Battery Test Names
Cognitive:		
7	Memory	(Short) Memory Test - Computer
6	Number	Number Memory Test - Computer
8	Perceptual Speed & Accuracy	Perceptual Speed & Accuracy - Computer Target Identification Test - Computer
4	Induction	Reasoning Test 1 Reasoning Test 2
5	Reaction Time	Simple Reaction Time - Computer Choice Reaction Time - Computer
3	Spatial Orientation	Orientation Test 1 Orientation Test 2 Orientation Test 3
2	Spatial Visualization/Field Independence	Shapes Test
1	Spatial Visualization	Object Rotations Test Assembling Objects Test Path Test Maze Test
Non-Cognitive, Biodata/Temperament:		
1	Adjustment	ABLE (Assessment of Background Life Experiences)
2	Dependability	
3	Achievement	
4	Physical Condition	
5	Potency	
6	Locus of Control	
7	Agreeableness/Likeability	
1	Validity Scales	
Non-Cognitive, Interests:		
1	Realistic	AVOICE (Army Vocational Interest Career Examination)
2	Investigative	
3	Conventional	
4	Social	
5	Artistic	
6	Enterprising	
Psychomotor:		
1	Multilimb Combination	Target Tracking Test 2 - Computer Target Shoot - Computer
2	Precision	Target Tracking Test 1 - Computer
3	Manual Dexterity	(None)

*Final priority arrived at via consensus of March 1984 IPR attendants.

Figure II.4. Predictor categories discussed at IPR in March 1984, linked to Pilot Trial Battery test names.

Section 2

SUMMARY OF PILOT TESTS PROCEDURES

The initial pilot testing of the predictor battery was carried out in three different samples. Not all tests were administered to each sample and revisions were made in the instruments after each data collection. The basic procedures are described below, in summary fashion, to help maintain clarity for the reader as the results are discussed in later sections. The first three administrations of the Pilot Trial Battery were at Fort Carson, Fort Campbell, and Fort Lewis.

The tables in this section list measures that have not yet been discussed in detail (i.e., the new tests designed to measure the constructs identified in Section 1). The individual new tests will be fully described as part of the discussion of the pilot test results in Sections 3-5. ABLE is the new inventory developed to include temperament and biographical items. AVOICE is an interest inventory which is a modification of the VOICE (Vocational Interest Career Examination) originally developed by the Air Force. The marker tests were the off-the-shelf instruments that had also been included in the Preliminary battery.

Pilot Test 1: Fort Carson

Sample and Procedure

On 17 April 1984, 43 soldiers at Fort Carson, Colorado, participated in the first pilot testing of the Pilot Trial Battery. The testing session ran from 0800 hours to 1700 hours, with two 15-minute breaks (one mid-morning and one mid-afternoon), and a 1-hour break for lunch.

Groups of five soldiers at a time were randomly selected to take computerized measures in a separate room while the remaining soldiers took paper-and-pencil tests (new cognitive tests and selected marker tests). When a group of five soldiers completed the computerized measures, they were individually and collectively interviewed about their reactions to the computerized tests, especially regarding clarity of instructions, face validity of tests, sensitivity of items, and general disposition toward such tests. The soldiers then returned to the paper-and-pencil testing session, and another group of five was selected to take the computer measures.

Thus, the maximum N for any single paper-and-pencil test was 38 (43 minus 5). Computerized measures were administered to a total of 20 soldiers. The new paper-and-pencil cognitive tests in the Pilot Trial Battery were each administered in two equally timed halves, to investigate the Part 1/Part 2 correlations as estimates of test reliability.

Actual test administration was completed by approximately 1545 hours. Ten soldiers were then selected to give specific, test-by-test feedback about paper-and-pencil tests in a small group session, while the remaining soldiers participated in a more general feedback and debriefing session.

Tests Administered

Table II.1 contains a list of all the tests administered at Fort Carson, in the order in which they were administered, with the time limit and the number of items for each test.

Pilot Test 2: Fort Campbell

Sample and Procedure

The second pilot testing session was conducted at Fort Campbell, Kentucky, on 16 May 1984. Fifty-seven soldiers attended the 8-hour session, and all 57 completed paper-and-pencil tests. No computerized measures were administered at this pilot session. Once again, the 10 new cognitive tests were administered in two equally timed halves, to investigate Part 1/Part 2 correlations. Because we were still experimenting with time limits on the new cognitive tests, soldiers were asked to mark which item they were on when time was called for each of these tests, and then continue to work on that part of the test until they finished. Finishing times were recorded for all the tests (Parts 1 and 2 separately, where appropriate).

Test administration was completed at approximately 1600 hours, and the group was divided. Ten individuals were selected to provide specific feedback concerning the new non-cognitive measures, and the remaining individuals provided feedback on the new cognitive measures.

Tests Administered

Table II.2 lists all the tests and inventories administered at Pilot Test 2 along with the time limit and number of items for each. There were 10 new cognitive tests with 5 cognitive marker tests, and 2 new non-cognitive inventories, with 1 non-cognitive marker inventory. No computerized measures were administered.

Pilot Test 3: Fort Lewis

For the third pilot testing session, approximately 24 soldiers per day for 5 days (11-15 June 1984) were available for testing at Fort Lewis, Washington. A total of 118 soldiers participated. Their mean age and time in the Army were 22.8 and 2.5 years, respectively. There were 97 men and 22 women, and 66 whites, 30 blacks, and 14 Hispanics. They were distributed over a wide range of MOS. Test sessions ran from 0800 hours to 1700 hours with short breaks in the morning and afternoon, and a 1-hour lunch break. The entire Pilot Trial Battery, including new cognitive and non-cognitive measures, was administered to all soldiers.

Once again, the new paper-and-pencil cognitive tests were administered in two equally timed halves to investigate Part 1/Part 2 correlations as estimates of test reliability. Individuals were not allowed any extra time to work on each test beyond the time limits, but finishing times were recorded for individuals completing tests before time was called.

Table II.1

Tests of Pilot Trial Battery Administered at Fort Carson (17 April 1984)

Test	Time Limit (Mins.)	No. of Items	Type of Test
Paper-and-Pencil Tests			
1. Path Test	9	35	New, Cognitive
2. Reasoning Test 1	14	30	New, Cognitive
3. EAS Test 1 - Verbal Comprehension	5	30	Marker, Cognitive
4. Orientation Test 1	8	20	New, Cognitive
5. Shapes Test	16	54	New, Cognitive
6. EAS Test 2 - Numerical Ability	10	75	Marker, Cognitive
7. Object Rotation Test	7	60	New, Cognitive
8. ETS Choosing a Path	8	16	Marker, Cognitive
9. Orientation Test 2	8	20	New, Cognitive
10. Reasoning Test 2	11	32	New, Cognitive
11. Orientation Test 3	12	20	New, Cognitive
12. Assembling Objects Test	16	30	New, Cognitive
13. Maze Test	9	24	New, Cognitive
14. Mental Rotations Test	10	20	Marker, Cognitive
15. ETS Hidden Figures	14	16	Marker, Cognitive
16. ETS Map Planning	6	40	Marker, Cognitive
17. ETS Figure Classification	8	14	Marker, Cognitive
18. EAS Test 5 - Space Visualization	5	50	Marker, Cognitive
19. FIT Assembly	10	20	Marker, Cognitive
Computer Measures[a]			
1. Simple Reaction Time	None	15	New, Perceptual/ Psychomotor
2. Choice Reaction Time	None	15	New, Perceptual/ Psychomotor
3. Perceptual Speed & Accuracy	None	80	New, Perceptual/ Psychomotor
4. Tracing Test	None	26	New, Perceptual/ Psychomotor
5. Short Memory Test	None	50	New, Perceptual/ Psychomotor
6. Hidden Figures Test	None	32	New, Perceptual/ Psychomotor
7. Target Shoot	None	20	New, Perceptual/ Psychomotor

[a] All computer measures were administered using a Compaq portable microprocessor with a standard keyboard plus a commercially available dual-axis joystick.

Table II.2

Pilot Tests Administered at Fort Campbell (16 May 1984)

Paper-and-Pencil Tests	Time Limit (Mins.)	No. of Items	Type of Test
1. Path Test	9	44	New, Cognitive
2. Reasoning Test 1	14	30	New, Cognitive
3. EAS Test 1 - Verbal Comprehension	5	30	Marker, Cognitive
4. Orientation Test 1	9	30	New, Cognitive
5. Shapes Test	16	54	New, Cognitive
6. Object Rotation Test 2	9	90	New, Cognitive
7. Reasoning Test 2	11	32	New, Cognitive
8. Orientation Test 2	8	20	New, Cognitive
9. ABLE (Assessment of Background and Life Experiences)	None	291	New, Non-Cognitive
10. Orientation Test 3	12	20	New, Cognitive
11. Assembling Objects Test	16	40	New, Cognitive
12. Maze Test	8	24	New, Cognitive
13. AVOICE (Army Vocational Interest Career Examination)	None	306	New, Non-Cognitive
14. ETS Hidden Figures	14	16	Marker, Cognitive
15. ETS Map Planning	6	40	Marker, Cognitive
16. ETS Figure Classification	8	14	Marker, Cognitive
17. FIT Assembly	10	20	Marker, Cognitive
18. POI (Personal Opinion Inventory)	None	121	Marker, Non-Cognitive

After each soldier completed the computer-administered battery, he or she was asked about general reactions to the computerized battery, the clarity and completeness of the instructions, the perceived difficulty of the tests, and the ease of using the response apparatus.

Tests Administered

The tests administered at Pilot Test 3, in Fort Lewis, are listed in Table II.3 with the time limit and number of items in each test.

Table II.3

Pilot Tests Administered at Fort Lewis (11-15 June 1984)

Administration Group	Test	Time Limit (Mins.)	No. of Items	Type of Test
	Paper-and-Pencil Tests			
	Path Test	8	44	New, Cognitive
	Reasoning Test 1	12	30	New, Cognitive
C1	Orientation Test 1	10	30	New, Cognitive
	Shapes Test	16	54	New, Cognitive
	Object Rotation Test	8	90	New, Cognitive
	Reasoning Test 2	10	32	Marker, Cognitive
	Maze Test	6	24	New, Cognitive
	SRA Word Grouping	5	30	Marker, Cognitive
	Orientation Test 2	10	24	New, Cognitive
C2	Orientation Test 3	12	20	New, Cognitive
	Assembling Objects Test	16	40	New, Cognitive
	ETS Map Planning	16	40	Marker, Cognitive
	Mental Rotations Test	10	20	Marker, Cognitive
	DAT Abstract Reasoning	13	25	Marker, Cognitive
NC	ABLE	None	268	New, Non-Cognitive
	AVOICE	None	306	New, Non-Cognitive
	Computerized Measures[a]			
	Simple Reaction Time	None	15	New, Perceptual/ Psychomotor
	Choice Reaction Time	None	15	New, Perceptual/ Psychomotor
	Perceptual Speed & Accuracy	None	80	New, Perceptual/ Psychomotor
	Target Tracking Test 1	None	18	New, Perceptual/ Psychomotor
	Target Tracking Test 2	None	18	New, Perceptual/ Psychomotor
	Target Identification Test	None	44	New, Perceptual/ Psychomotor
	Memory Test	None	50	New, Perceptual/ Psychomotor
	Target (Shoot) Test	None	40	New, Perceptual/ Psychomotor

[a] All computer measures were administered via a custom-made response pedestal designed specifically for this purpose. No responses were made on the computer keyboard. A Compaq microprocessor was used.

Summary of Pilot Tests

The Pilot Test Battery, initially developed in March 1984, went through three pilot testing iterations by August 1984. After each iteration, observations noted during administration were scrutinized, data analyses were conducted, and the results were carefully examined. Revisions were made in specific item content, test length, and time limits, where appropriate. Table II.4 summarizes the three Pilot Test sessions conducted during this period, with the total sample size for each, and the number and types of tests administered at each.

Table II.4

Summary of Pilot Testing Sessions for Pilot Trial Battery

Pilot Test No.	Location	Date	Total Sample Size	Number/Type of Tests Administered
1	Fort Carson	17 April 1984	43	10 New Cognitive 9 Marker Cognitive 0 New Non-Cognitive 0 Marker Non-Cognitive 7 Computerized Measures
2	Fort Campbell	16 May 1984	57	10 New Cognitive 5 Marker Cognitive 2 New Non-Cognitive 1 Marker Non-Cognitive 0 Computerized Measures
3	Fort Lewis	11-15 June 1984	118	10 New Cognitive 4 Marker Cognitive 2 New Non-Cognitive 0 Marker Non-Cognitive 8 Computerized Measures

The following sections in Part II contain discussions of each test, inventory, and measure in the Pilot Trial Battery, its evolution through the pilot testing process, and its status as of the end of August 1984.

Section 3

DEVELOPMENT OF COGNITIVE PAPER-AND-PENCIL MEASURES

This section describes the development of the paper-and-pencil cognitive predictor measures, up to the point at which they were ready for field testing as part of the Pilot Trial Battery. As described previously, cognitive ability constructs had been evaluated and prioritized according to their judged relevance and importance for predicting success in a variety of the Army MOS. These priority judgments were used to plan the development activities for cognitive paper-and-pencil tests.

Each cognitive predictor category is discussed in turn. Within each category are a definition of the target cognitive ability and an outline of the strategy followed to develop the measure(s) of the target ability. This includes identifying (a) the target population or target MOS for which the measure is hypothesized to most effectively predict success; (b) published tests that served as markers for each new measure; (c) intended level of item difficulty, and (d) type of test (i.e., speed, power, or a combination). The test itself is then described and example items are provided. Results from the first two pilot test administrations or tryouts are reported, to explain and document subsequent test revisions. Finally, psychometric test data obtained from the third pilot test, conducted at Fort Lewis, are discussed.

The last portion of this section presents a summary of the newly developed cognitive ability tests. This includes a discussion of test intercorrelations, results from a factor analysis of the intercorrelations, and results from subgroup analyses of test scores.

General Issues

Before describing the individual tests, we would like to summarize certain general issues germane to all the cognitive paper-and-pencil measures.

Target Population

The population for which these tests have been developed is the same one to which the Army supplies the ASVAB, that is, persons applying to enlist in the Army. However, that target population was, practically speaking, inaccessible during the development process. We were constrained to the use of incumbents. Enlisted soldiers represent a restricted sample of the target population in that they all have passed enlistment standards and, furthermore, almost all of the soldiers that we were able to use in our pilot tests had also passed Basic and Advanced Individual Training. Thus, they are presumably more qualified, more able, more persevering, and so forth, on the average, than are the individuals in the target population. We tried to take this into account for (a) developing tests that have a broad range of item difficulties and (b) selecting items of somewhat lower difficulty.

Power vs. Speed

Another decision to be made about each test was its placement on the power vs. speed continuum. Most psychometricians would agree that a "pure" power test is a test administered in such a way that each person is allowed enough time to attempt all items, and that a "pure" speeded test is a test administered in such a way that no one taking the test has enough time to attempt all of the items. In practice, most tests fall somewhere between the two extremes. It also is the case that a power test usually contains items that not all persons could answer correctly, even given unlimited time to complete the test, while a speeded test usually contains items that all or almost all persons could answer correctly, given enough time to attempt the items.

As a matter of practical definition, an "80% completion" rule-of-thumb was used to define a power test. That is, if a test could be completed by 80% of all those taking the test, then we considered it a "power" test.

Reliability

Several procedures are available to assess the reliability of a measure and each provides distinct information about a test. Split-half reliability estimates were obtained for each paper-and-pencil test administered at the first three pilot test sites. For each pilot test, each test was administered in two separately timed parts. Reliability estimates are obtained by correlating scores from the two parts. The Spearman-Brown correction procedure was then used to estimate the reliability for the whole test. This estimate of reliability is appropriate for either speeded or power tests.

Hoyt internal consistency reliability estimates are also reported for each test, providing the average reliability across all possible split-test halves. Ths procedure is less appropriate for speeded tests because it overestimates the reliability.

Individual Test Descriptions

We turn now to the descriptions of the individual tests, which are discussed within cognitive ability constructs. This description is given in some detail because these are new measures that are of fundamental importance for the basic goals of Project A. As mentioned above, a standard format is used to describe the development of each instrument. Readers who are not interested in the specifics of predictor content may wish to turn to the summary sections.

Construct - Spatial Visualization

Spatial visualization involves the ability to mentally manipulate components of two- or three-dimensional figures into other arrangements. The process involves restructuring the components of an object and accurately

discerning their appropriate appearance in new configurations. This construct includes several subcomponents, two of which are:

- Rotation - the ability to identify a two-dimensional figure when seen at different angular orientations within the picture plane. It also includes three-dimensional rotation or 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.

- Scanning - the ability to visually survey a complex field to find a particular configuration representing a pathway through the field.

Currently, no ASVAB measures are designed specifically to measure spatial abilities. Because of this, spatial visualization received a developmental priority rating of one (see Figure II.4). The visualization construct was divided into two parts: visualization/rotation and visualization/scanning. We developed two tests within each of these areas; these four tests are described below.

Spatial Visualization - Rotation

The two tests developed for this ability are Assembling Objects and Object Rotation. The former involves three-dimensional figures, while the latter involves two-dimensional objects.

Assembling Objects Test

Development Strategy. Predictive validity estimates provided by expert raters suggest that measures of the visualization/rotation construct would be effective predictors of success in MOS that involve mechanical operations and construction and drawing or using maps. The Assembling Objects Test was designed to yield information about the potential for success in such MOS.

Published tests identified as markers for Assembling Objects include the Employee Aptitude Survey Space Visualization (EAS-5) and the Flanagan Industrial Test (FIT) Assembly. EAS-5 requires examinees to count three-dimensional objects depicted in two-dimensional space, whereas the FIT Assembly involves mentally piecing together objects that are cut apart or disassembled. The FIT Assembly was selected as the more appropriate marker for our purposes because it has both visualization and rotation components involving mechanical or construction activities. The Assembling Objects Test was designed to assess the ability to visualize how an object will look when its parts are put together correctly. It was intended that this measure would combine power and speed components, with speed receiving greater emphasis.

Test Description. In the original form of the Assembling Objects Test, subjects were asked to complete 30 items within a 16-minute time limit. Each item presents subjects with components or parts of an object. The task is to select, from among four alternatives, the one object that depicts the components or parts put together correctly. The two item types are included in the test; examples of each are shown in Figure II.5.

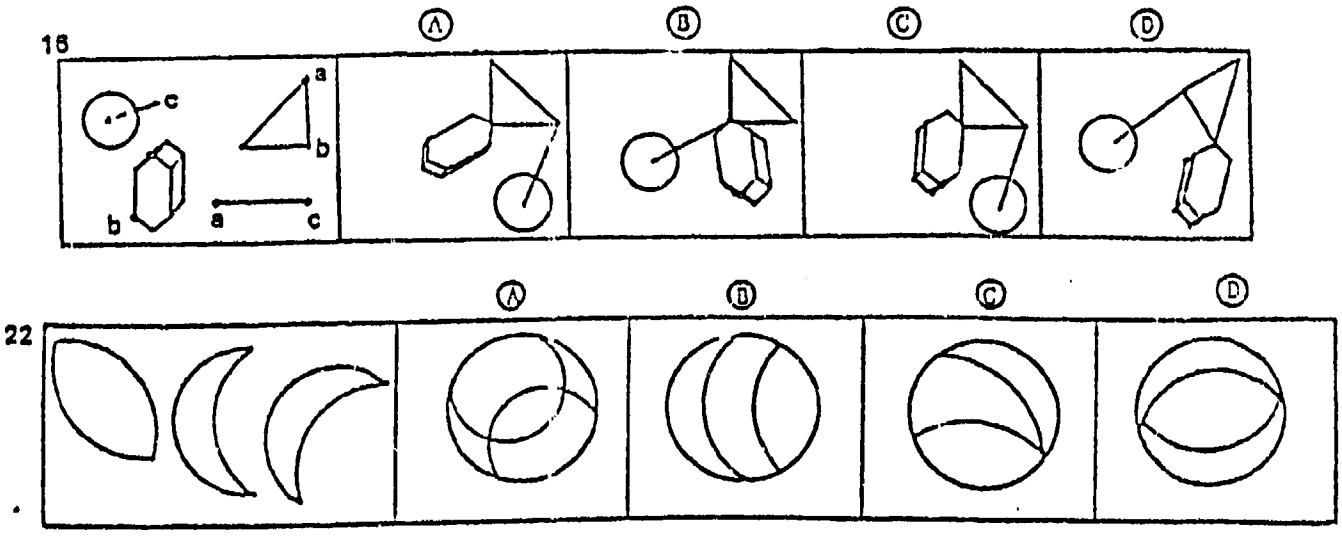


Figure II.5 Sample items from Assembling Objects Test.

Results from the first tryout, conducted at Fort Carson, indicated that the test may have suffered from ceiling effects. That is, nearly all recruits in this sample (N = 36) completed the test and their mean score was 24.2 (SD = 5.05). Further, item difficulty levels were somewhat higher than intended (mean = .80, SD = .12, median = .83).

Therefore 10 new, more difficult items, five for each item type, were constructed and added to the test to reduce the likelihood of ceiling effects. The 16-minute time limit was retained for the second tryout, at Fort Campbell. Nearly all subjects completed the test (mean = 37.3, SD = 4.75) and the mean score was 26.3 (SD = 8.34, N = 56). Item difficulty levels were lower for the revised test (mean = .68, SD = .15, median = .72). Inspection of these results indicated that the test possessed acceptable psychometric qualities, so no changes were made in preparation for the Fort Lewis pilot test.

Test Characteristics. At Fort Lewis the Assembling Objects Test contained 40 items with a 16-minute time limit. The mean number of items completed, standard deviation, and range were 37.6, 3.83, and 18 to 40, respectively. Corresponding values for number correct (or test score) were 28.1, 7.51, and 7 to 40. Item difficulties range from .31 to .92 with a mean of .70 (SD = 14.7). Item-total correlations range from .18 to .60 with a mean of .44 (SD = 9.99). Parts 1 and 2 correlate .65 with each other. Reliabilities are estimated at .79 by split-half methods (Spearman-Brown corrected), and .89 with Hoyt's estimate of reliability.

Correlations between scores on this measure and scores on other Pilot Trial Battery paper-pencil measures are reported at the end of this section. It is important, however, to note the correlations between this test and

marker tests. Both marker tests were administered in the Fort Carson tryout and the FIT Assembly was also used at Fort Campbell. Results from Fort Carson indicate that scores on Assembling Objects correlate .74 with scores on EAS-5 and .76 with scores on FIT Assembly (N = 30). Results from Fort Campbell indicate that this test correlates .64 with FIT Assembly (N = 54). This last value represents a better estimate of the relationship between Assembling Objects and its marker, FIT Assembly, because of the revisions made to Assembling Objects following the first tryout at Fort Carson.

Modifications for the Fort Knox Field Test. In preparation for the Fort Knox administration, some Assembling Objects items were redrawn to clarify the figures. The item response format was modified to approximate a format suitable for machine scoring, a change that was made in all of the tests being prepared for field test administration.

Object Rotation Test

Development Strategy. Published tests serving as markers for the Object Rotation measure include Educational Testing Service's (ETS) Card Rotations, Thurstone's Flags Test, and Shephard-Metzler Mental Rotations. Each of these measures requires the subject to compare a test object with a standard object to determine whether the two represent the same figure with one simply turned or rotated or whether the two represent different figures. The first two measures, ETS Card Rotations and Thurstone's Flags, involve visualizing two-dimensional rotation of an object, whereas the Mental Rotations test requires visualizing three-dimensional objects depicted in two-dimensional space.

Object Rotation Test items were constructed to reflect a limited range of item difficulty levels ranging from very easy to moderately easy, and designed to be easier than those in the Assembling Objects Test. The new test had more items and a shorter time limit than the Assembling Objects Test.

Test Description. The initial version contained 60 items with a 7-minute time limit. The subject's task involved examining a test object and determining whether the figure represented in each item is the same as the test object, only rotated, or is not the same as the test object (e.g., flipped over). For each test object there are five test items, each requiring a response of "same" or "not same." Sample test items are shown in Figure II.6.

Results from the Fort Carson administration indicated that this test suffered from ceiling effects. For example, item difficulty levels averaged .92 (SD = .05). Therefore, we decided to add 30 new items to the test and to increase the time limit to 9 minutes for the second tryout at Fort Campbell.

Results from the second tryout indicated that subjects, on the average, completed 87.6 (SD = 8.0) of the 90 items and obtained a mean score of 77.0 (SD = 12.1). The time limit was reduced to 8 minutes for the Fort Lewis administration to obtain a more highly speeded test.

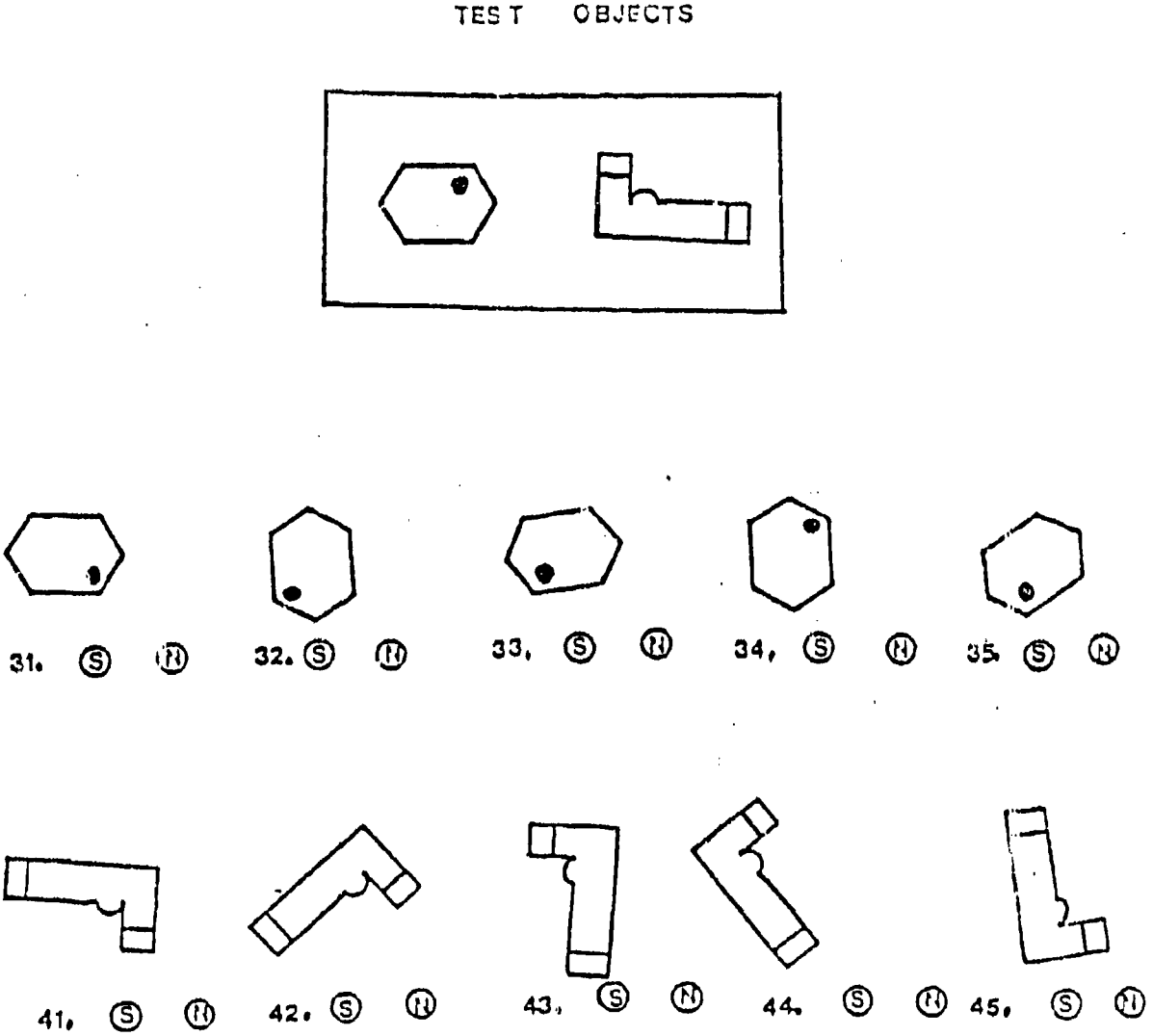


Figure II.6. Sample items from Object Rotation Test.

Test Characteristics. Detailed results from the Fort Lewis pilot test showed fairly high completion rates (mean = 84.6 and SD = 10.8), with a range of 48 to 90. Test scores, computed by the total number correct, ranged from 36 to 90 with a mean of 73.4 (SD = 15.4). Item difficulty levels range from .59 to .98 with a mean of .81 (SD = .11). Item-total correlations (item validities) average .44 (SD = .17), ranging from .09 to .79. Parts 1 and 2 correlate .73 with each other. The split-half reliability estimate, corrected for test length, is .86 while the Hoyt estimate is .96.

The marker test for Object Rotation, Mental Rotations, was administered at two of the three pilot test sites. Data collected at the Fort Carson

tryout indicate that the two measures correlate .60 (N = 30), whereas data from Fort Lewis indicate the two correlate .56 (N = 118).

Modifications for the Fort Knox Field Test. Results from the Fort Lewis pilot test indicated that all test items possessed desirable psychometric properties. However, the time limit was decreased to 7.5 minutes to make the test even more speeded and avoid a possible ceiling effect. The response format was modified to approximate a format suitable for machine scoring.

Spatial Visualization - Scanning

The second component of spatial visualization ability which was emphasized in predictor development is spatial scanning. Spatial scanning tasks require the subject to visually survey a complex field and find a pathway through it, utilizing a particular configuration. The Path Test and the Maze Test were developed to measure this component of spatial visualization.

Path Test

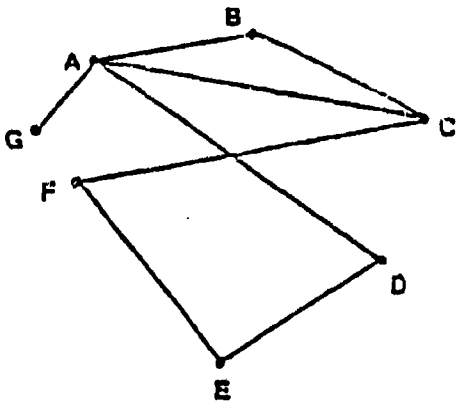
Development Strategy. Published tests serving as markers for construction of the Path Test include ETS Map Planning and ETS Choosing a Path. In thse measures, examinees are provided with a map or diagram. The task is to follow a given set of rules or directions to proceed through the pathway or to locate an object on the map.

Results from earlier research with the marker tests, ETS Map Planning and ETS Choosing a Path, indicated that both tests are highly speeded and were very difficult for the target sample (Hough et al., 1984). For example, 80% of the subjects (Army recruits) completed only 16 of the 40 items contained in the Map Planning Test, and only 5 of the 16 items in the Choosing a Path Test. Consequently, Path Test items were constructed to yield difficulty levels for the target population ranging from very easy to somewhat difficult and the test time was established to place more emphasis on speed than on power.

Test Description. The Path Test requires subjects to determine the best path or route between two points. Subjects are presented with a map of airline routes or flight paths. Figure II.7 contains a flight path with four sample items. The subject's task is to find the "best" path or the path between two points that requires the fewest number of stops. Each lettered dot is a city that counts as one stop; the beginning and ending cities (dots) do not count as stops.

In its original form, the Path Test contained 35 items with a 9-minute time limit. Subjects were asked to record the numbers of stops for each item in a corresponding blank space.

Results from the first tryout, conducted at Fort Carson, revealed that the test was too easy. That is, virtually all of the subjects completed the test and they obtained a mean score of 29.9. Item difficulty levels ranged from .48 to 1.00 with a mean of .85. To reduce the potential for ceiling effects, an additional map or flight path with 13 items was added to the test. In addition, four very easy items were deleted, resulting in 44 items on the revised test. The 9-minute limit was retained.



The route from:	Number of Stops:
1. A to F	① ② ③ ④ ⑤
2. G to E	① ② ③ ④ ⑤
3. C to D	① ② ③ ④ ⑤
4. G to F	① ② ③ ④ ⑤

Figure II.7. Sample items from Path Test.

Results from the second tryout indicate that, on the average, subjects completed 40.7 items and obtained a mean score of 32.6 (SD = 7.0). Item difficulty levels ranged from .55 to .96 with a mean of .80. To prepare for the third tryout, conducted at Fort Lewis, the test response format was revised to allow subjects to circle the number of stops (i.e., 1-5) instead of filling in a blank. In addition, the time limit was reduced from 9 minutes to 8 minutes to increase the speededness of the test.

Test Characteristics. In results from the Fort Lewis tryout of the revised Path Test, subjects, on the average, completed 35.3 of the 44 items (SD = 8.3). Test scores, computed by the total number correct, ranged from 0 to 44 with a mean of 28.3 (SD = 9.1). Item difficulty levels range from .20 to .91 with a mean of .64. Item-total correlations average .47 with a range of .25 to .69. Parts 1 and 2 correlate .70. The split-half reliability estimate, corrected for test length, is .82. The Hoyt internal consistency value is .92.

One or both marker tests were administered at all pilot test sites. Data from the first tryout indicate that the original Path Test correlates .34 with ETS Choosing a Path and -.01 with ETS Map Planning. The reader is reminded that results from Fort Carson are based on a very small sample size (N = 19) and that the Path Test was greatly modified following this tryout.

The ETS Map Planning Test was also administered at the Fort Carson and Fort Lewis tryouts. These data indicate that the Path Test and Map Planning correlate .62 (N= 54) and .48 (N = 118), respectively.

Modifications for the Fort Knox Field Test. The Path Test remained unchanged except that the response format was modified to approximate a format suitable for machine scoring.

Maze Test

Development Strategy. The Maze Test is the second measure constructed to assess spatial visualization/scanning. The development strategy mirrors that of the Path Test, with the same marker tests. The Maze Test, however, differs from the Path Test in that the task required involves finding the one pathway that allows exit from the maze, while the Path Test was designed to measure visualization/scanning ability under highly speeded conditions.

Test Description. For the first pilot test administration the Maze Test contained 24 rectangular mazes. Each included four entrance points, labeled A, B, C, and D, and three exit points indicated by an asterisk (*). The task is to determine which of the four entrances leads to a pathway through the maze and to one of the exit points. A 9-minute limit was established.

Results from the first tryout, at Fort Carson, indicate that the Maze Test suffered from ceiling effects. Subjects, on the average, completed 23.3 of the 24 items and obtained a mean score of 22.1 (SD = 2.18). To increase test score variance, the test was modified in two ways. First, an additional exit was added to each test maze. Figure II.8 contains a sample item from the original test and the same item modified for the Fort Campbell tryout. Second, the time limit was reduced from 9 to 8 minutes.

Data obtained at the second tryout, conducted at Fort Campbell, indicate that completion rates were again high (mean = 22.5). Therefore, for the third tryout the time limit for completing the 24 maze items was reduced to 6 minutes.

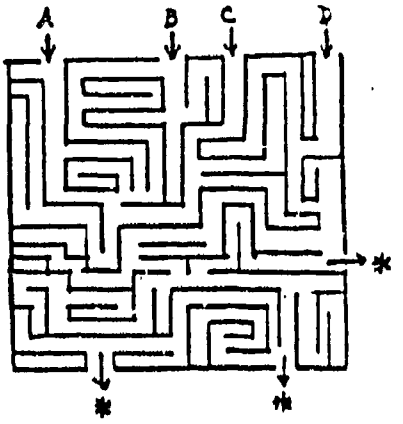
Test Characteristics. Results from the Fort Lewis tryout indicate that the reduced time produced a drop in the completion rate for the Fort Lewis sample (mean = 20.6). Test scores, computed by the total number correct, ranged from 8 to 24 with a mean of 19.3 (SD = 4.4). Item difficulty levels range from .41 to .98 with a mean of .81. Item-total correlations average .48 (SD = .22). Parts 1 and 2 correlate .64 with each other. The split-half reliability estimate corrected for test length is .78 and the Hoyt reliability estimate for this test is .88.

One or both of the marker tests, ETS Choosing a Path and ETS Map Planning, were administered at the three pilot test sites. Results from Fort Carson indicate that the Maze Test correlates .24 (N = 29) with Choosing a Path, and .36 (N = 30) with Map Planning. These values must be viewed with caution because of the small sample size and because of modifications made to the Maze Test following this tryout.

Map Planning was also administered at the Fort Campbell and Fort Lewis tryouts. Data collected at these posts indicate that Map Planning correlates .45 (N = 55) and .63 (N = 118), respectively, with the Maze Test.

Modifications for the Fort Knox Field Test. Results from the last pilot test administration showed that the Maze Test could be slightly more speeded. The percentage of subjects completing this test is higher than for

FORT CARSON



FORT CAMPBELL

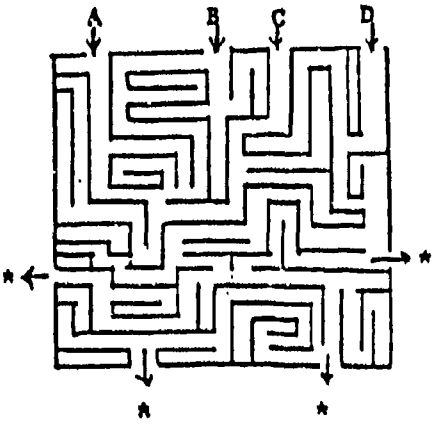


Figure II.8. Sample items from Maze Test.

the Path Test (i.e., 38% for the Maze Test, and 19% for the Path). Therefore, the time limit was reduced from 6 minutes to 5.5 minutes for the Fort Knox field test. In addition, the response format was modified to approximate that for machine scoring.

Construct - Field Independence

This construct involves the ability to find a simple form when it is hidden in a complex pattern. Given a visual percept or configuration, field independence refers to the ability to hold the percept or configuration in mind so as to disembed it from other well-defined perceptual material.

This construct received a mean validity estimate of .30 from the panel of expert judges, with the highest estimate of .37 appearing for MOS that involve detecting and identifying targets. Field independence received a priority rating of two for inclusion in the battery. One instrument, the Shapes Test, was developed to measure this construct.

Shapes Test

Development Strategy. The marker test for the Shapes Test is the ETS Hidden Figures Test, a measure included in the Preliminary Battery (Hough et al., 1984). In this test, subjects are asked to find one of five simple figures located in a more complex pattern. Initial analyses of the Preliminary Battery results indicated that for the target population, first-term enlisted soldiers, the Hidden Figures Test suffers from limited test score variance and possibly floor effects. For example, the initial data indicate that 80% of the sample completed fewer than 4 of the 16 test items.

Our strategy for constructing the Shapes Test was to use a task similar to that in the Hidden Figures Test while ensuring that the difficulty level of test items was geared more toward the Project A target population. Further, we decided to include more types of items that reflect varying difficulty levels. We wanted the test to be speeded, but not nearly so much as the ETS Hidden Figures Test.

Test Description. At the top of each test page are five simple shapes; below these shapes are six complex figures. Subjects are instructed to examine the simple shapes and then to find the one simple shape located in each complex figure (see Figure II.9).

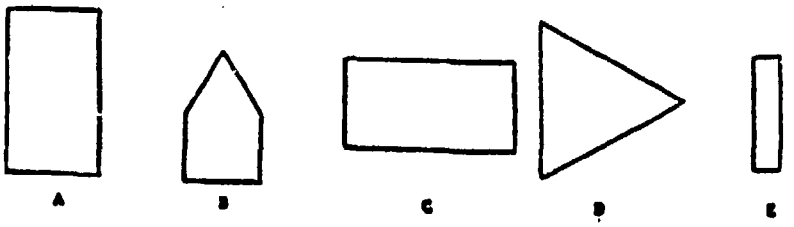
In the first tryout at Fort Carson, the Shapes Test contained 54 items with a 16-minute time limit. Results from this first tryout indicated that most subjects were able to complete the entire test and most subjects obtained very high scores (mean score = 49.3).

To prepare for the Fort Campbell tryout, nearly all test items were modified to increase item difficulty levels. Examples of item modifications are provided in Figure II.9. As is shown, by adding a few lines to each complex pattern, the test items administered at Fort Campbell were made more difficult than the items administered at the Fort Carson tryout.

Results from Fort Campbell indicate that test item modifications were successful. Subjects, on the average, completed 43.5 of the 54 items within the 16-minute time limit, and obtained a mean score of 30.9 (SD = 23.5).

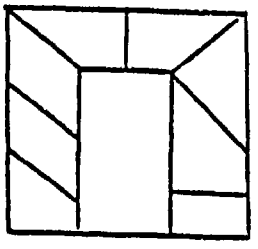
This test was modified only slightly for the Fort Lewis administration. For example, a few complex figures were revised to ensure that one and only one simple figure could be located in each complex figure.

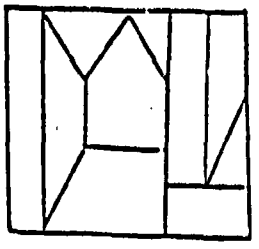
Test Characteristics. For the Fort Lewis sample the mean number completed was 42.4. The mean number correct was 29.3 (SD = 9.1) with a range of 12 to 51, indicating that the measure does not suffer from ceiling effects. Item difficulty levels range from .10 to .97 with a mean of .54. Reliability estimates indicate that Parts 1 and 2 correlate .69; the Spearman-Brown correction is .82, and the Hoyt reliability estimate is .89.



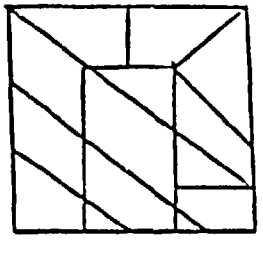
Complex Figures

Fort Carson





Fort Campbell



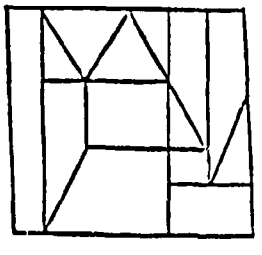


Figure II.9. Sample items from Shapes Test.

The marker test, ETS Hidden Figures Test, was administered at the first two tryouts. Results from the Fort Carson tryout indicate that the original version of the Shapes Test correlates .35 with the Hidden Figures Test (N = 29). Data from Fort Campbell indicate that the revised Shapes Test correlates .50 with its marker (N = 56).

Modifications for the Fort Knox Field Test. The Shapes Test required only minor revisions for this final tryout. For example, item-total correlations for a few items indicated that more than one shape could be located in a complex figure test item, so these figures were modified. In addition, the response format was changed to approximate that for machine scoring.

Construct - Spatial Orientation

This construct involves the ability to maintain one's bearings with respect to points on a compass and to maintain appreciation of one's location relative to landmarks in the environment.

This particular construct was not included in the list of predictor constructs evaluated by the expert panel. However, conceptualization and measurement of this ability construct first appeared during World War II, when researchers for the Army Air Forces (AAF) Aviation Psychology Program explored a variety of constructs to aid in selecting air crew personnel. Results from the AAF Program indicated that measures of spatial orientation were useful in selecting pilots and navigators (Guilford & Lacey, 1947). Also, during the second year of Project A, a number of job observations suggested that some MOS involve critical job requirements of maintaining directional orientation and establishing location, using features or landmarks in the environment. Consequently, three different measures of this construct were formulated.

Orientation Test 1

Development Strategy. Direction Orientation Form B (CP515B) developed by researchers in the AAF Aviation Psychology Program served as the marker for Orientation Test 1. The strategy for developing Orientation Test 1 involved generating items that duplicated the task in the AAF test. Each item contains six circles. The first, the standard compass or "given" circle, indicates the direction of North. This is usually rotated out of the typical or conventional position. The remaining circles are test compasses that also have directions marked on them. For this test, item construction was limited to one of seven possible directions: South, East, West, Southwest, Northwest, Southeast, and Northeast. The plan for this test was to ask subjects to complete numerous compass directional items within a short period of time. Orientation Test 1 was designed as a highly speeded test of spatial orientation.

Test Description. Each test item presented subjects with six circles. In the test's original form, the first, or Given, circle, indicated the compass direction for North. For most items, North was rotated out of its conventional position. Compass directions also appeared on the remaining five circles. The subject's task was to determine, for each circle, whether

or not the direction indicated was correctly positioned by comparing it to the direction of North in the Given circle. (See Example 1 in Figure II.10.)

When administered to the Fort Carson sample, this test contained 20 item sets requiring 100 responses (i.e., for every item, compass directions on five circles must be evaluated). Subjects were given 8 minutes to complete the test. Test scores were determined by the total number correct; the maximum possible was 100.

Results from this administration indicated that nearly all subjects completed the items within the time allotted and the mean score was 82.7 (SD = 17.8). Item difficulty levels indicate that most items were moderately easy (mean = .83). For the Fort Campbell tryouts, we attempted to create more difficult items by modifying directional information provided in the Given circle. That is, rather than indicating the direction for North, compass directions for South, East, and West were provided. These directions were also rotated out of conventional compass position. (See Example 2, Figure II.10.) Orientation Test 1, as presented at the Fort Campbell tryout, contained 30 item sets or 150 items, administered in three separately timed parts. Parts 1 and 2 included the original test items, and Part 3 the new (non-North) items. Part 3 was preceded by additional test instructions informing subjects about the change in Given circle directions. Subjects were given 3 minutes to complete each part.

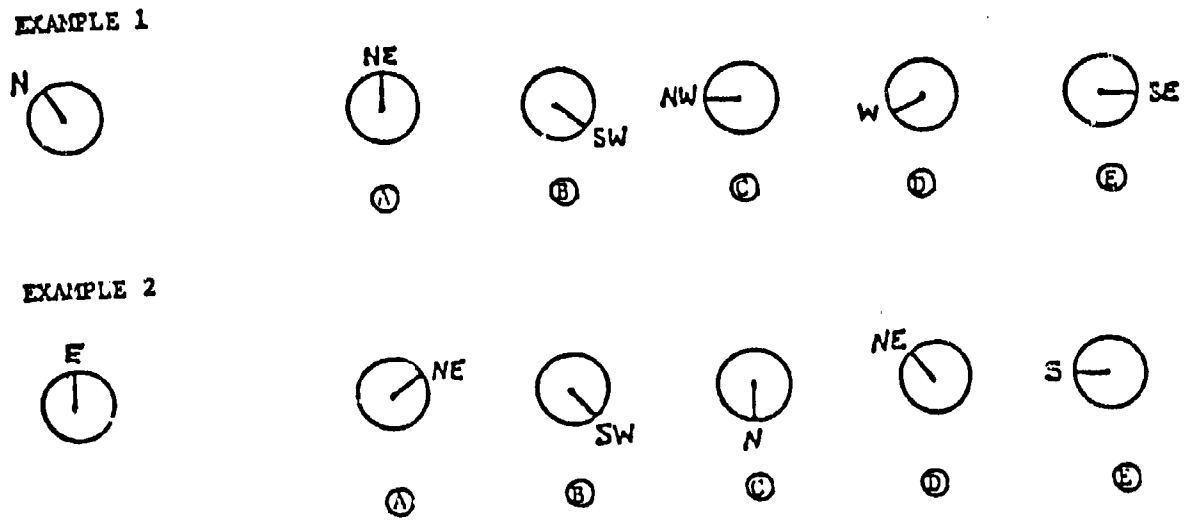


Figure II.10. Sample items from Orientation Test 1.

In this second tryout, scores on Part 3 yielded lower correlations with Parts 1 and 2 (both are equal to .44); Parts 1 and 2 correlated .87. From this information we reasoned that the new items were assessing additional information about subjects' abilities to maintain orientation. Item sets from Part 3 were then mixed with item sets from Parts 1 and 2 to create a test with 30 item sets (150 items) for the Fort Lewis tryout. Further, the time limit was increased to a total of 10 minutes. Test instructions were modified to explain that items vary throughout the test with respect to information provided in the Given circle.

Test Characteristics. At the Fort Lewis tryout completion rates for the total test indicated that, on the average, examinees attempted 25 of the 30 item sets and obtained a mean score of 117.8 (SD = 24.1). Item difficulty levels range from .21 to .97 with a mean of .79. The correlation between Parts 1 and 2 is .86. Reliability estimates are .92 for split-half Spearman-Brown corrected and .97 for Hoyt.

Modifications for the Fort Knox Field Test. Very few changes were made. Response format was modified to approximate a format scorable by machine. Orientation Test 1 contains 30 item sets (150 items) with a 10-minute time limit.

Orientation Test 2

Development Strategy. The second measure of spatial orientation was also designed to tap abilities that might predict success for MOS that involve maintaining appreciation of one's location relative to landmarks in the environment or in spite of frequent changes in direction. Orientation Test 2 is a relatively new approach to assessing spatial orientation abilities. Although no particular test served as its model, it is similar to a measure designed by Army Air Force researchers to select pilots, navigators, and bombardiers (Directional Orientation: CP5150).

The task designed for Orientation Task 2 asks subjects to mentally rotate objects and then to visualize how components or parts of those objects will appear after the object is rotated. Item difficulty levels were varied by altering the degree of rotation required to correctly complete each part of the task. Because of the complexity of the task, Orientation 2 was initially viewed as power test of spatial orientation.

Test Description. For Orientation Test 2, each item contains a picture within a circular or rectangular frame. The bottom of the frame has a circle with a dot inside it. The picture or scene is not in an upright position. The task, then, is to mentally rotate the frame so that the bottom of the frame is positioned at the bottom of the picture. After doing so, one must then determine where the dot will appear in the circle. (See Figure II.11 sample items.) For the Fort Carson tryout, this test contained 20 items with an 8-minute time limit.

Results from this administration indicate that the time limit was sufficient (mean number completed = 19.9), but that item difficulty levels were somewhat lower than desired (mean = .52). Item-total correlations were, however, impressive (mean = .48, SD = .10). The only potential problem

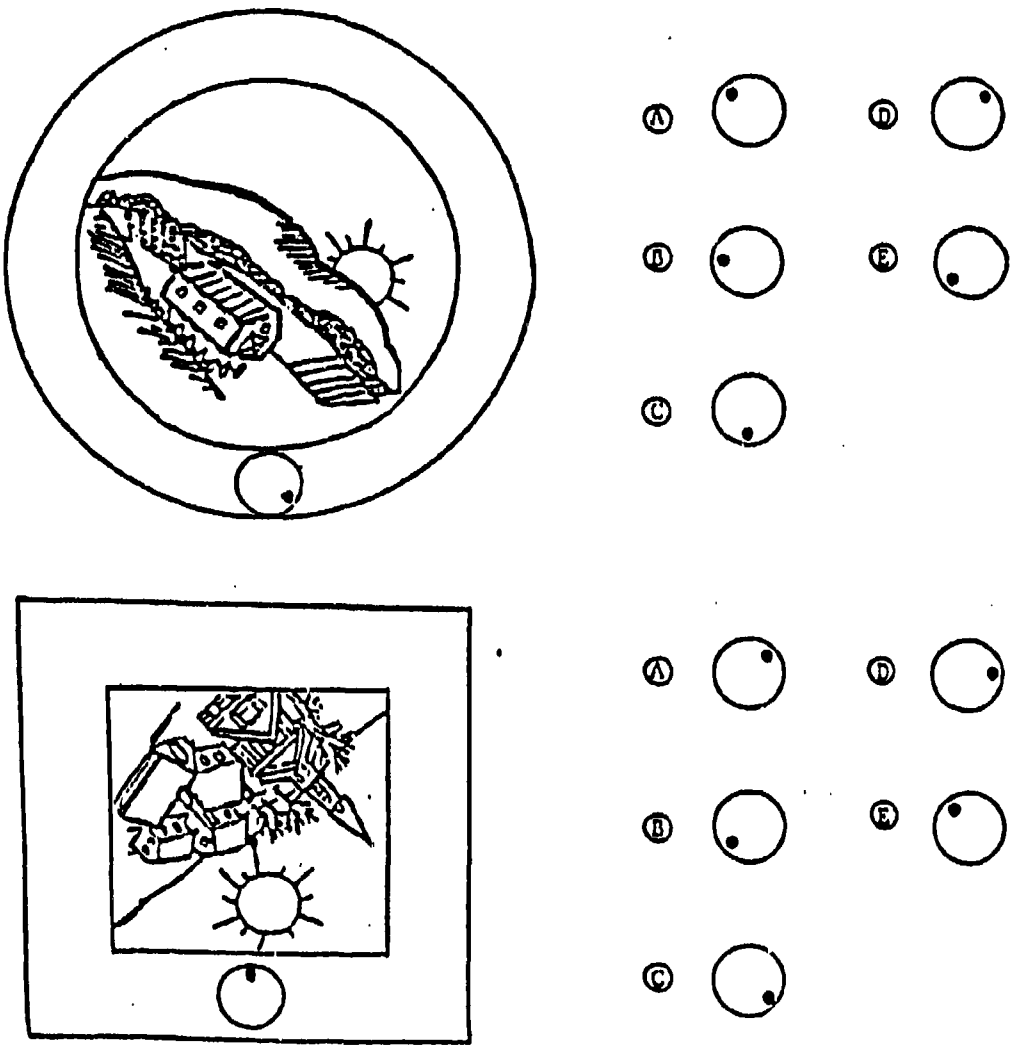


Figure II.11. Sample items from Orientation Test 2.

involved the test instructions, which some respondents found difficult. For the Fort Campbell tryout, test instructions were modified to clarify the nature of the task.

Data collected at Fort Campbell provide information similar to the Fort Carson data; however, the mean score and item difficulty levels indicated that the test was more difficult for this group than for the Fort Carson sample (mean score = 8.6; mean item difficulty = .43). Because of these lower item-difficulty levels, four new, easier items were added.

Orientation Test 2, as administered to the Fort Lewis sample, contained 24 items, and a 10-minute time limit was established to correspond to the increase in the number of items. Test scores on this measure are determined by the total number correct.

Test Characteristics. The Fort Lewis data indicate that Orientation Test 2 is a power test (mean number completed = 23.7, SD = 1.0). Subjects obtained a mean score of 11.5 (SD = 6.2). Item difficulty levels range from .19 to .71 with a mean of .48; this represents a slight increase from the Fort Campbell tryout. Item-total correlations remain high (mean of .53). Scores from Parts 1 and 2 correlate .80. Correcting this value for test length yields a split-half reliability estimate of .89. The Hoyt internal consistency value is also .89.

Modifications for the Fort Knox Field Test. This measure was virtually unchanged for the Fort Knox administration. Only the response format was modified to approximate machine scoring.

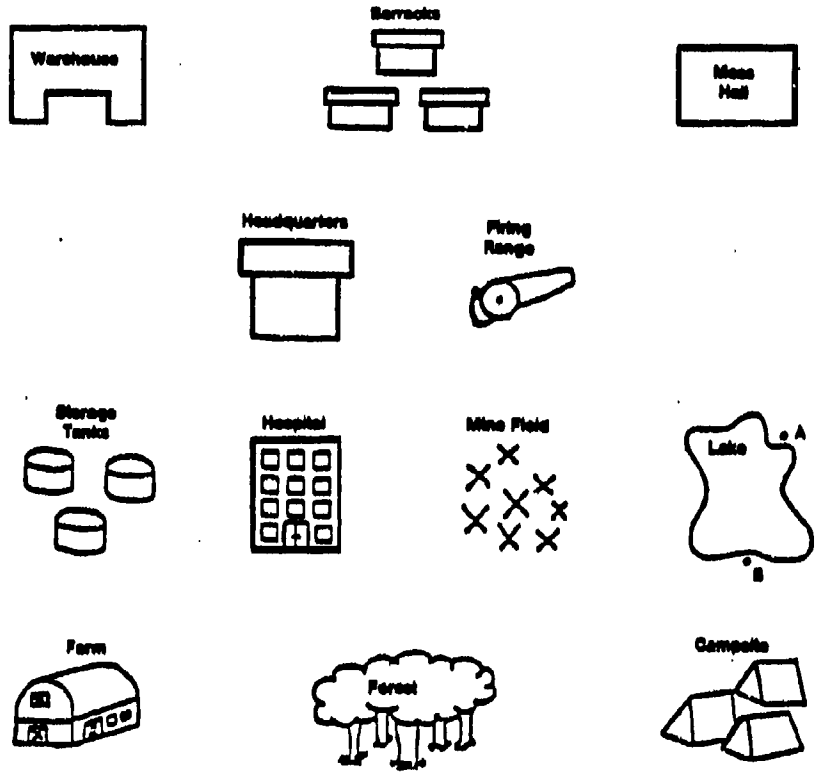
Orientation Test 3

Development Strategy. This test was also designed to measure spatial orientation and was modeled after another spatial orientation test, Compass Directions, developed by researchers in the AAF Aviation Psychology Program. The AAF measure was designed to assess the ability to reorient oneself to a particular ground pattern quickly and accurately when compass directions are shifted about. Orientation Test 3 was designed to assess the same ability using a similar test format. This test was designed to place somewhat more emphasis on speed than on power.

Test Description. In its original form, Orientation Test 3 presented subjects with a map that includes various landmarks such as a barracks, a campsite, a forest, a lake. Within each item, subjects are provided with compass directions by indicating the direction from one landmark to another, such as "the forest is North of the campsite." Subjects are also informed of their present location relative to another landmark. Given this information, the subject must determine which direction to go to reach yet another structure or landmark. Figure II.12 contains one test map and two sample items. Note that for each item, new or different compass directions are given.

For the Fort Carson tryout, the test contained two maps with 10 questions about each map, for a total of 20 items. Subjects were given 12 minutes to complete the test. Results from this first tryout revealed very few problems with the test (e.g., test instructions were clear, the time was sufficient, no floor nor ceiling effects appeared). Thus, this measure remained unchanged for the Fort Campbell pilot test.

Results from the second tryout yielded similar information (e.g., no ceiling nor floor effects, completion rates were acceptable). These data, however, indicated that for a few items, two responses could be correct because of a lack of precision in drawing the two maps. Accordingly, landmarks on each map were repositioned to ensure that one and only one



1. The forest is due west of the barracks. You are at headquarters. Which direction must you travel in order to reach the firing range?

   1. N 2. NE 3. E 4. SE 5. S 6. SW 7. W 8. NW

2. The firing range is southwest of the hospital. You are at the farm. Which direction must you travel in order to reach the campsite?

   1. N 2. NE 3. E 4. SE 5. S 6. SW 7. W 8. NW

Figure II.12. Sample items from Orientation Test 3.

correct answer existed for each item. When administered to the Fort Lewis sample, Orientation Test 3 contained 20 test items with a 12-minute time limit. Test scores are determined by the total number correct.

Test Characteristics. On the average, subjects at Fort Lewis completed 18.0 items (SD = 2.7). The mean score of 8.7 (SD = 5.8) indicates that subjects correctly answered about one-half of the items attempted. Item difficulty levels range from .24 to .63 with a mean of .43. Item-total correlations range from .48 to .72 with a mean of .59 (SD = .07). Part 1 and Part 2 correlate .79. The split-half reliability estimate corrected for test length is .88, while the Hoyt internal consistency estimate is .90.

Data from Fort Carson indicate that Orientation Test 3 correlates .66 with Orientation Test 1 (N = 29) and .42 with Orientation 2 (N = 31). Values for these same measures administered at Fort Campbell are .72 and .54 (N = 56). Data from Fort Lewis indicate that these measures correlate .68 and .65 (N = 118).

Modifications for the Fort Knox Field Test. This test was virtually unchanged for the Fort Knox field test. The only exception involves changes to approximate a machine-scorable response format.

Construct - Induction/Figural Reasoning

This construct involves the ability to generate hypotheses about principles governing relationships among several objects.

Example measures of induction include the Employee Aptitude Survey Numerical Reasoning (EAS-6), Educational Testing Service (ETS) Figure Classification, Differential Aptitude Test (DAT) Abstract Reasoning, Science Research Associates (SRA) Word Grouping, and Raven's Progressive Matrices. These paper-and-pencil measures present subjects with a series of objects such as figures, numbers, or words. To complete the task, subjects must first determine the rule governing the relationship among the objects and then apply the rule to identify the next object in the series. Two different measures of the construct were developed for Project A.

Reasoning Test 1

Development Strategy. The plan for developing Reasoning Test 1 was to construct a test that was similar to the task appearing in EAS-6, Numerical Reasoning, but with one major difference: Items would be composed of figures rather than numbers. Test items were constructed to represent varying degrees of difficulty ranging from very easy to very difficult. Following item development activities, time limits were established to allow sufficient time for subjects to complete all or nearly all items. Thus, Reasoning Test 1 was designed as a power measure of induction.

Test Description. Reasoning 1 test items present subjects with a series of four figures; the task is to identify from among five possible answers the one figure that should appear next in the series. In the original test, subjects were asked to complete 30 items in 14 minutes. Sample items are provided in Figure II.13.

Results from the first tryout, conducted at Fort Carson, indicate subjects, on the average, completed 29.5 items and obtained a mean score of 20.7 (SD = 3.5). Inspection of difficulty levels indicated that items were unevenly distributed between the two test parts, so items were reordered to ensure that easy and difficult items were equally distributed throughout both test parts. Only minor modifications were made to test items; for example, one particularly difficult item was redrawn to reduce the difficulty level.

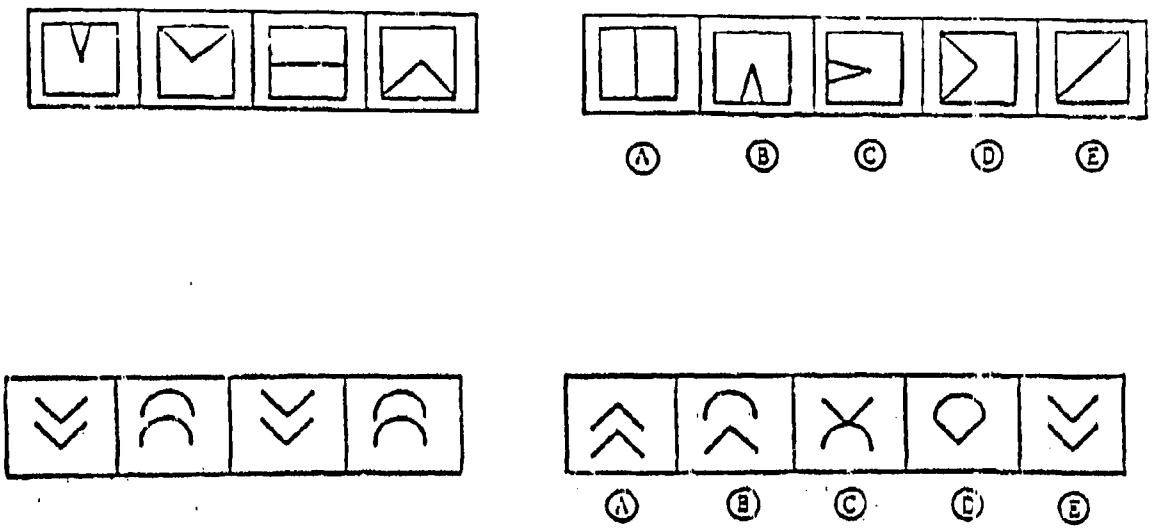


Figure II.13. Sample items from Reasoning Test 1.

Data collected at Fort Campbell indicate that again nearly all subjects completed the test. Further, test administrators reported that those who completed the test finished early. As a consequence, the 14-minute time limit was reduced to 12 minutes. Also, two items were revised because distractors yielded higher item-total correlations than the correct response.

Test Characteristics. Fort Lewis subjects, on the average, completed 29.4 items with about 84% of the subjects completing the entire test. Test scores, computed from the total number correct, ranged from 4 to 29 with a mean of 19.6. Item difficulty levels range from .26 to .92 with a mean of .65. Part 1 and Part 2 correlate .64. The split-half reliability estimate corrected for test length is .78, while the Hoyt value is .86.

Two other measures of induction, SRA Word Grouping and DAT Abstract Reasoning, were administered at the Fort Lewis tryout. Results indicate that Reasoning Test 1 correlates .47 with Word Grouping and .74 with Abstract Reasoning. These data are compatible with our understanding of the two marker measures of induction. The former contains a verbal component while the latter measures induction via figural reasoning.

Modifications for the Fort Knox Field Test. Test instructions were revised slightly for the Fort Knox field test, and the response format was modified to approximate that used for machine scoring.

Reasoning Test 2

Development Strategy. This measure also was designed to assess induction using items that require figural reasoning.

Published tests serving as markers for Reasoning 2 include EAS-6, Numerical Reasoning, and ETS Figure Classification. The original development strategy was to develop Reasoning Test 2 fairly similarly to the ETS test. Initial analyses conducted on ETS Figure Classification data (N = 1,863) indicated that this test was too highly speeded for the target population (Hough et al., 1984). For example, 80% of recruits taking the Figure Classification test finished fewer than half of the 112 items. Further, although item difficulty levels varied greatly, the mean value indicated most items are moderately easy (mean = 73, SD = 22).

Therefore, although the ETS Figure Classification test served as the marker in early test development planning for Reasoning 2, the new measure differed in several ways. First, ETS Figure Classification requires subjects to perform two tasks: to identify similarities and differences among groups of figures and then to classify test figures into those groups. Items in Reasoning Test 2 were designed to involve only the first task. Second, test items were constructed to reflect a wide range of difficulty levels, with the average item falling in the moderately difficult range. Finally, because the items would be more difficult overall, Test 2 could contain fewer items. The test was thus designed as a power measure of figural reasoning, with a broad range of item difficulties.

Test Description. Test items present five figures. Subjects are asked to determine which four figures are similar in some way, thereby identifying the one figure that differs from the others. (See Figure II.14.) This test, when first administered, contained 32 items with an 11-minute time limit.

Results from the Fort Carson tryout indicated that nearly all subjects completed the entire test. Item difficulty levels were somewhat higher than expected, ranging from .05 to 1.0 with a mean of .71 (SD = .29). Because eight of the test items yielded item difficulty levels of .97 or above, these items were either modified or replaced to increase item difficulties. Moreover, inspection of item difficulties indicated that Part I contained a greater proportion of the easier items, so items were redistributed throughout the test.

For the Fort Campbell tryout, Reasoning Test 2 again contained 32 items with an 11-minute time limit. Analysis of the data indicated desirable psychometric qualities. For example, nearly all subjects completed the test. Test scores ranged from 9 to 26 with a mean of 19.1 (SD = 3.5) and difficulty levels decreased. Although the part-part correlation increased from the first tryout, it still remained low (i.e., Fort Campbell $\underline{r}$ = .40 versus Fort Carson $\underline{r}$ = .32).

A few changes were made in the test prior to the third tryout. For example, four items contained a distractor that was selected more often and that yielded a higher item-total correlation than the correct response; these items were revised. Test administrators at Fort Campbell noted that the time limit could be reduced without significantly altering test completion rates, so the limit was reduced to 10 minutes for the next administration.

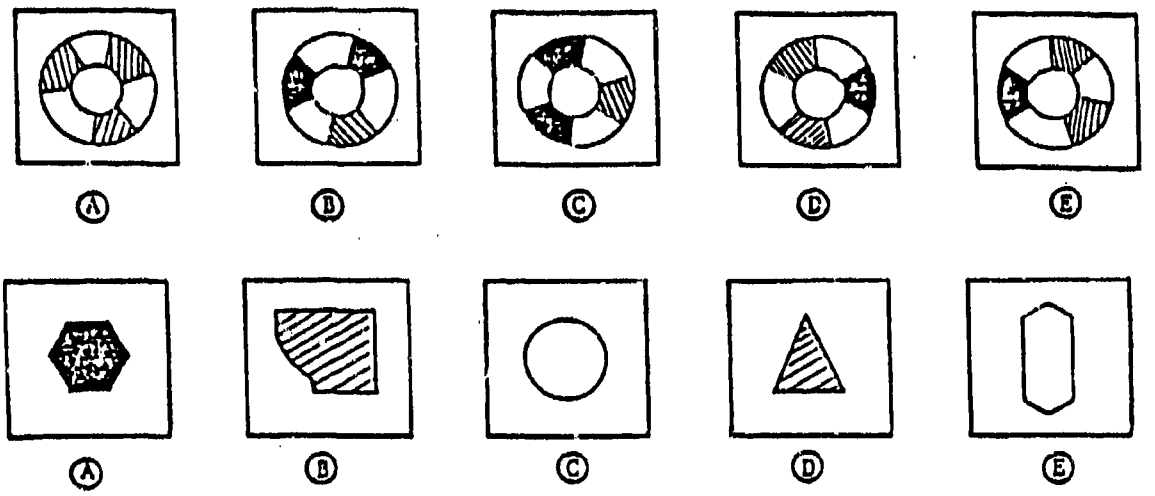


Figure II.14. Sample Items from Reasoning Test 2.

Test Characteristics. In the third tryout 70% completed the entire test (however, 84% completed the separately timed first half and 79% completed the second half). Scores ranged from 11 to 28 with a mean of 21.8 (SD = 3.4). Item difficulties range from .17 to 1.0 with a mean of .64. Parts 1 and 2 correlate .46. The split-half reliability estimate corrected for test length is .63 while the Hoyt value is .61. These values suggest that this is a more heterogeneous test of figural reasoning than is Reasoning Test 1.

The marker test, ETS Figure Classification, was administered at the first two tryouts. Correlations between Reasoning Test 2 and its marker are .35 (N = 30 at Fort Carson) and .23 (N = 56 at Fort Campbell). These low correlations are not too surprising, given the task requirement differences and the power versus speed component differences between these two measures. Two other measures of induction, SRA Word Grouping and DAT Abstract Reasoning, were administered at the third tryout. These data indicate that Reasoning Test 2 correlates .48 with Word Grouping and .66 with Abstract Reasoning (N = 118). Once again, these differences in correlations are as expected, since Word Grouping contains a verbal component whereas Abstract Reasoning, like Reasoning Test 2, assesses induction using figural items.

Modifications for the Fort Knox Field Test. The response format was modified to approximate that used for machine scoring. Reasoning Test 2 contained 32 items with a 10-minute time limit for the Fort Knox field test.

Summary of Pilot Test Results for Cognitive Paper-and-Pencil Measures

In this section, we summarize the data available as of August 1984 for the 10 cognitive paper-and-pencil measures. This includes test score information, intercorrelations among the 10 measures, and results from factor analyses. The bulk of the data reported here was obtained from the Fort Lewis tryout. Table II.5 summarizes the Fort Lewis data. All data are based on a sample size of 118, with the exception of the Path Test, which is based on a sample size of 116.

Table II.5

Cognitive Paper-and-Pencil Measures: Summary of Fort Lewis Pilot Test Results

Measure	No. of Items	Mean Score	SD	Mean Item-Difficulty Levels	Split-Half[a] r_{xx}
SPATIAL VISUALIZATION					
Rotation					
Assembling Objects	40	28.14	7.51	.70	.79
Object Rotation	90	73.36	15.40	.82	.86
Scanning					
Path	44	28.28	9.08	.64	.82
Mazes	24	19.30	4.35	.80	.78
FIELD INDEPENDENCE					
Shapes	54	29.28	9.14	.54	.82
SPATIAL ORIENTATION					
Orientation 1	150	117.86	24.16	.79	.92
Orientation 2	24	11.53	6.20	.48	.89
Orientation 3	20	8.71	5.78	.44	.88
REASONING					
Reasoning 1	30	19.64	5.75	.66	.78
Reasoning 2	32	21.82	3.38	.64	.63

[a] All reliability estimates (split halves with Part 1-Part 2 separately timed) have been corrected with the Spearman-Brown procedures.

Table II.6 contains the intercorrelation matrix for the 10 cognitive ability measures. One of the most obvious features of this matrix is the high level of correlation across all measures. The correlations across all test pairs range from .40 to .68. These data suggest that the test measures overlap in the abilities assessed.

This finding is not altogether surprising. For example, 4 of the 10 measures were designed to measure spatial abilities such as visualization, rotation, and scanning. The Shapes Test, designed to measure field independence, also includes visualization components. The three tests constructed to measure spatial orientation involve visualization and rotation tasks. The final two measures, Reasoning Test 1 and Reasoning Test 2, also require visualization at some level to identify the principal governing relationships among figures and to determine the similarities and differences among figures. Thus, across all measures, abilities needed to complete the required tasks overlap to some degree. This overlap is demonstrated in the intercorrelation matrix.

To provide a better understanding of the underlying structure, the intercorrelation matrix was factor analyzed. Several solutions were computed, ranging from two to five factors. The rotated orthogonal solution for four factors appeared most meaningful. Results from this solution appear in Table II.7. As shown in the table, to interpret results from the four-factor solution, we first identified all factor loadings of .35 or higher. Next, we examined the factor loading pattern for each measure and then identified measures with similar patterns to form test clusters. Five test clusters or groups, labeled A through E, are identified in Table II.7. These clusters represent a first attempt to identify the underlying structure of the cognitive measures included in the Pilot Trial Battery. Each test cluster is described below.

Group A - Assembling Objects and Shapes Test. Recall that the Shapes Test requires the subject to locate or disembed simple forms from more complex patterns, while the Assembling Objects Test requires the subject to visualize how an object will appear when its components are put together. Both measures require subjects to visualize objects or forms in new or different configurations. Further, these measures contain both power and speed components, with each falling more toward the speed end of the continuum.

Group B - Object Rotation, Path, and Maze Tests. Object Rotation involves two-dimensional rotation of objects or forms while the Path and Maze tests involve visually scanning a map or diagram to identify the best pathway or the one pathway that leads to an exit. These measures are all highly speeded; that is, subjects are required to perform the tasks at a fairly rapid rate. Further, the tasks involved in each of these measures appear less complex or easier than those involved in the Assembling Objects or Shapes tests.

Group C - Orientation Test 1 and Orientation Test 3. Orientation Test 1 requires the examinee to compare compass directions provided on a test circle and a given circle, while Orientation Test 3 involves using a map, compass directions, and present location to determine which direction to go to reach

Table II.6

Intercorrelations Among the 10 Cognitive Paper-and-Pencil Measures: Pilot Test Data[a]

Measure	1. Assembling Objects	2. Object Rotation	3. Path	4. Maze	5. Shapes	6. Orientation 1	7. Orientation 2	8. Orientation 3	9. Reasoning 1	10. Reasoning 2
1. Assembling Objects	--									
2. Object Rotation	.33	--								
3. Path	.52	.45	--							
4. Maze	.59	.57	.60	--						
5. Shapes	.61	.50	.51	.56	--					
6. Orientation 1	.62	.52	.54	.52	.56	--				
7. Orientation 2	.60	.45	.48	.51	.47	.53	--			
8. Orientation 3	.62	.50	.40	.47	.60	.68	.65	--		
9. Reasoning 1	.62	.52	.60	.58	.59	.59	.56	.60	--	
10. Reasoning 2	.53	.50	.48	.52	.51	.54	.53	.53	.63	--

[a]All correlations are computed from a sample size of 118 except those involving the Path Test, which are based on sample size of 116.

Table II.7

Rotated Orthogonal Factor Solution for Four Factors on Cognitive Paper-and-Pencil Measures: Pilot Test Data[a]

	I	II	III	IV	h^{2b}
Shapes	.47	.49			.568
Assembling Objects	.47	.48			.621
Object Rotation	.50	.37			.473
Path	.55		.40		.541
Mazes	.76				.727
Orientation 1	.39	.57			.617
Orientation 3		.79		.35	.827
Orientation 2		.35		.74 ← D	.684
Reasoning 1	.39	.35	.67		.778
Reasoning 2	.37	.36	.44		.521

A: Shapes, Assembling Objects
B: Object Rotation, Path, Mazes
C: Orientation 1, Orientation 3
D: Orientation 2

[a] Factor loadings of .35 or higher are shown.

[b] h^2 = Proportion of total test score variance in common with other tests, or common variance.

a landmark on the map. Both measures require examinees to quickly and accurately orient themselves with respect to directions on a compass and landmarks in the environment, despite shifts or changes in the directions. Both are highly speeded measures of spatial orientation.

Group D - Orientation Test 2. This measure involves mentally rotating a frame so that it corresponds to or matches up with the picture inside, and then visualizing how components on the frame (a circle with a dot) will appear after it has been rotated. This appears to be a very complex spatial measure that requires several abilities such as visualization, rotation, and orientation. In addition to the task complexity differences, this measure may also differ from other spatial measures on the power-speed continuum. Unlike the other spatial measures included in the Pilot Trial Battery, Orientation Test 2 is a power rather than a speed test.

Group E - Reasoning Test 1 and Reasoning Test 2. Reasoning Test 1 requires one to identify the principle governing the relationship or pattern among several figures, while Reasoning Test 2 involves identifying similarities among several figures to isolate the one figure that differs from the others. As noted above, these measures appear to involve visualization abilities. The reasoning task involved in each, however, distinguishes these measures from the other tests included in the Pilot Trial Battery.

Results from analyses of the Fort Lewis data provide a preliminary structure for the cognitive paper-and-pencil tests designed for the Pilot Trial Battery. Correlations among the measures indicate that all measures require spatial visualization abilities at some level. The measures may, however, be distinguished by the type of task, task complexity, and speed and power component differences.

In this section we have focused on the cognitive paper-and-pencil measures. Other cognitive measures in the Pilot Trial Battery were administered via computer and are described in the following section. Correlations among the cognitive paper-and-pencil tests and the cognitive computer-administered tests are also reported in that section. Administration and results of the field tests are reported in Section 6.

Section 4

DEVELOPMENT OF COMPUTER-ADMINISTERED TESTS

In this section the development steps and pilot test results for the computer-administered measures are described. Before discussing the tests themselves, we will briefly describe a critical piece of equipment designed especially for pilot administrations of the computerized tests in the Pilot Trial Battery.

The microprocessor selected for use, the COMPAQ, contains a standard keyboard, and in early tryouts of the computer battery subjects were asked to make their responses on this keyboard. These preliminary trials suggested that the use of a keyboard may provide an unfair advantage to subjects with typing or data entry experience, and that the standard keyboard did not provide adequate experimental control during the testing process. Consequently, a separate response pedestal was designed and built and was ready for use in the final pilot test at Fort Lewis.

This response pedestal is depicted in Figure II.15. Note that it contains two joysticks (one for left-handed subjects and one for right-handed subjects), two sliding resistors, a dial for entering demographic data such as age and social security number, two red buttons, three response buttons--blue, yellow, and white--and four green "home" buttons. (One of the "home" buttons is not visible in the diagram; it is located on the side of the pedestal.)

The "home" buttons play a key role in capturing subjects' reaction time scores. They control the onset of each test item or trial when reaction time is being measured. To begin a trial, the subject must place his or her hands on the four green buttons. After the stimulus appears on the screen and the subject has determined the correct response, he or she must remove his or her preferred hand from the "home" buttons and press the correct response button.

The procedure involving the "home" buttons serves two purposes. First, control is added over the location of the subjects' hands while the stimulus item is presented. In this way, hand movement distance is the same for all subjects and variation in reaction time due to position of subjects' hands is reduced to nearly zero.

Second, procedures involving these buttons are designed to assess two theoretically important components of reaction time measures--decision time and movement time. Decision time includes the period between stimulus onset and the point at which the subject removes his or her hands to make a response; this interval reflects the time required to process the information to determine the correct response. Movement time involves the period between removing one's hands from the "home" buttons and striking a response key. The "home" buttons on the response pedestal, then, are designed to investigate the two theoretically independent components of reaction time. Results from an investigation of these measures appear throughout this section.

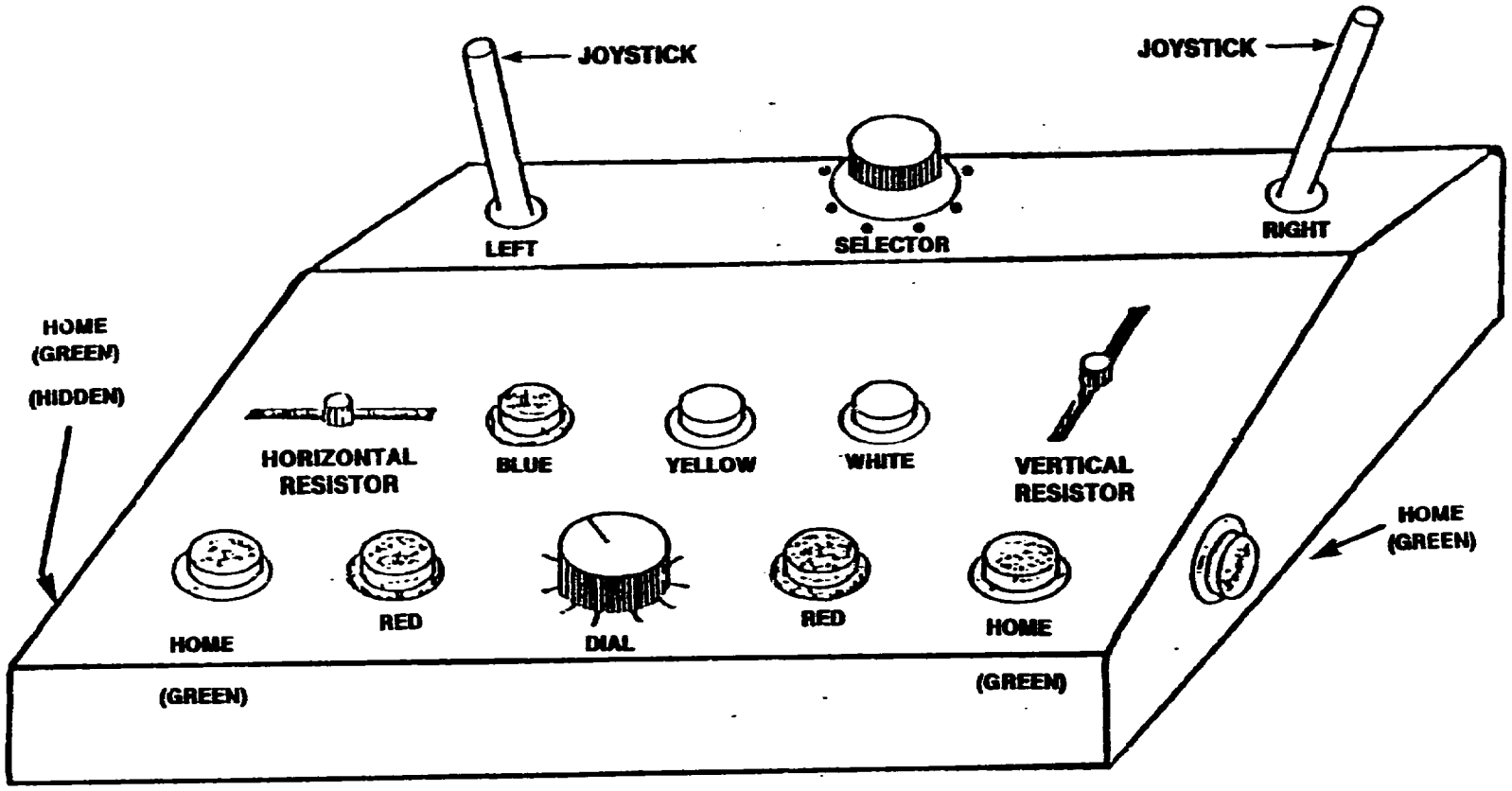


Figure II.15. Response pedestal for computerized tests.

On the following pages, we describe the development and pilot testing of computer-administered tests designed to measure three cognitive ability constructs and two psychomotor ability constructs. Tests were also developed to measure two additional constructs but were not pilot tested.

Construct - Reaction Time (Processing Efficiency)

This construct involves speed of reaction to stimuli--that is, the speed with which a person perceives the stimulus independent of any time taken by the motor response component of the classic reaction time measures. According to our definition of this construct, which is an indicator of processing efficiency, it includes both simple and choice reaction time.

Simple Reaction Time: Reaction Time Test 1

The basic paradigm for this task stems from Jensen's research involving the relationship between reaction time and mental ability (Jensen, 1982).

At the computer console, the subject is instructed to place his or her hands on the green "home" buttons. On the computer screen, a small box appears. After a delay period (ranging from 1.5 to 3.0 seconds) the word "yellow" appears in the box. The subject must remove the preferred hand from the "home" buttons to strike the yellow key. The subject must then return both hands to the ready position to receive the next item.

This test contains 15 items. Although it is self-paced, subjects are given 10 seconds to respond before the computer "time-outs"[1] and prepares to present the next item.

Test Characteristics. Table II.8 contains data on the test characteristics from the Fort Lewis pilot test. Variables in the upper part of the table provide descriptive information about test performance. Note that, on the average, subjects read the test instructions in 2.5 minutes, although this time ranges from about half a minute to 5 minutes. Further, subjects completed the test in 1.2 minutes; this ranged from .8 minute to over 5 minutes. Total test time, then, ranged from 1.6 to 7.1 minutes with a mean of 3.7 minutes. Very few subjects timed-out or provided invalid responses; the maximum number of time-outs for any subject was three, the maximum number of invalid responses was one. Finally, percent-correct values indicate nearly all subjects understood the task and performed it correctly.

---

[1]Time-outs occur if a subject fails to respond within a specified period of time. Invalid responses occur when the subject strikes the wrong key. In both cases, the item disappears from the computer screen and, after the subject gets in the ready position, the next item appears on the screen.

Table II.8

Reaction Time Test 1 (Simple Reaction Time): Fort Lewis Pilot Test

Descriptive Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	2.51	.81	.63 - 5.01
Time to Complete Test (minutes)	1.22	.62	.79 - 5.19
Total Test Time (minutes)	3.72	.99	1.59 - 7.10
Time-Outs (number per person)	.05	.31	0 - 3
Invalid Responses (number per person)	.07	.26	0 - 1
Percent Correct	99%	3%	80 - 100%

Test Scores[a]	Mean	SD	Range	Rxx[b]
Decision Time (10 items)	30.50	10.15	17.90 - 109.78	.91
Trimmed[c] Decision Time (8 items)	29.25	8.10	18.75 - 82.00	.92
SD - Decision	7.85	12.05	.92 - 118.26	.77
Movement Time (10 items)	27.35	8.98	15.50 - 91.33	.75
Trimmed Movement Time (8 items)	26.01	7.26	15.50 - 55.86	.94
SD - Movement	6.68	12.77	.75 - 121.07	.20
Total Time (10 items)	57.84	15.78	37.90 - 149.56	.90
Trimmed Total Time (10 items)	55.92	13.86	37.75 - 124.71	.94
SD - Total	11.79	16.80	1.58 - 125.85	.66

[a] All values reported are in hundredths of a second.

[b] Rxx = odd-even correlations, corrected to full test length using the Spearman-Brown formula.

[c] Trimmed scores are based on response to items 6-15, excluding the highest and lowest scores.

Test Scores. To identify variables of interest, we reviewed the literature in this area. (See Keyes, 1985.) This review indicated that the reaction time is often calculated for decision time, movement time, and total time. In addition, intra-individual variation measures (the standard deviation of total reaction time scores) calculated for each subject appear to provide useful information. Considering problems related to practice effects, only the last 10 responses were included in the mean reaction scores. Further, because subtle events may produce extreme reaction scores for a single item, trimmed scores, which include responses to items 6 through 15 with the highest and lowest reaction time values removed, were used for decision, movement, and total time.

Mean values for all the above measures were calculated. They appear in Table II.8 along with reliability estimates for each measure, computed using an odd-even method with a Spearman-Brown correction.

The relationships among these measures of reaction time were examined by computing all pairwise correlations. These results indicate that a low to moderate relationship exists between movement time and decision time (r = .32 for 10 items). Movement time appears to be providing kinds of information similar to total time (r = .77 for 10 items). Decision time, however, provides additional information (r = .50 for 10 items).

Correlations calculated with paper-and-pencil cognitive measures indicate that decision time, total standard deviation, and percentage correct are virtually unrelated to scores on these paper-and-pencil measures. Total reaction time, however, correlates highest with the Maze Test (-.39), the Path Test (-.23), and Orientation Test 1 (-.23). The detailed information on intercorrelations between the computer-administered tests and the cognitive paper-and-pencil tests is provided in the final portion of Section 4.

Finally, scores on these measures were correlated with video experience. Prior to completing the computer tests, subjects had been asked to rate, on a 5-point scale, their degree of experience[2] with video game playing. Mean decision trimmed and mean total trimmed times correlate near zero with this variable. Total standard deviation correlates .19 and percent correct correlates -.20 with this measure.

Modifications for Fort Knox Field Test. This test remained the same for the Fort Knox field test.

Choice Reaction Time: Reaction Time Test 2

Reaction time for two response alternatives is obtained by presenting the term BLUE or WHITE on the screen. The subject is instructed when one of these appears, to move his or her preferred hand from the "home" keys to strike the key that corresponds with the term appearing on the screen (BLUE or WHITE).

---

[2]A rating of 1 indicated no experience with video games; 5 indicated much experience.

This measure contains 15 items, with 7 requiring responses on the WHITE key and 8 requiring responses on the BLUE key. Although the test is self-paced, the computer is programmed to allow 9 seconds for a response before going on to the next item.

Test Characteristics. Table II.9 provides data describing this test as given at Fort Lewis. Note that subjects were reading the instructions more quickly than they were for Simple Reaction Time and were also finishing the test more quickly.

Information about whether the same or different hands were used to respond to all items is not reported in this table. Data on hand use indicate that 23% of the subjects (N = 26) consistently used the same hand. The remainder (77%, N = 86) switched from hand to hand at least once to respond.

Test Scores. Mean values along with standard deviations, ranges, and reliability estimates are provided in Table II.9. Note that for this measure, only the two reaction time scores provide reliable information. (These reliability estimates were calculated using an odd-even procedure with a Spearman-Brown correction.)

Another measure involves the difference between mean Choice Reaction Time scores and Simple Reaction Time scores. This value is intended to capture a speed of processing component. Note that reliability estimates suggest these values are internally consistent.

Modification for Fort Knox Field Test. No changes were made to this test for the Fort Knox field test.

Construct - Short-Term Memory

This construct is defined as the rate at which one observes, searches, and recalls information contained in short-term memory.

Memory Search Test

The marker used for this test is a short-term memory search task introduced by S. Sternberg (1966, 1969). In this test, the subject is presented with a set of one to five familiar items (e.g., letters); these are withdrawn and then the subject is presented with a probe item. The subject is to indicate, as rapidly and as accurately as possible, whether or not the probe was contained in the original set of items, now held in short-term memory. Generally, mean reaction time is regressed against the number of objects in the item or stimulus set. The slope of this function can be interpreted as the average increase in reaction time with an increase of one object in the memory set, or the rate at which one can access information in short-term memory.

The measure developed for computer-administered testing is very similar to that designed by Sternberg. At the computer console, the subject is instructed to place his or her hands on the green home buttons. The first

Table II.9

Reaction Time Test 2 (Choice Reaction Time): Fort Lewis Pilot Test

Descriptive Characteristics	Mean	SD	Range	
Time to Read Instructions (minutes)	1.01	.36	.45 - 2.37	
Time to Complete Test (minutes)	.95	.13	.80 - 1.59	
Total Test Time (minutes)	1.95	.40	1.37 - 3.20	
Time-Outs (number per person)	0	0	0 - 1	
Invalid Responses (number per person)	.17	.10	0 - 1	
Test Scores	**Mean**	**SD**	**Range**	**Rxx[a]**
Mean Decision Time[b]	36.78	7.76	18.75 - 78.29	.94
Mean Total Time[b]	65.98	10.38	37.75 - 117.29	.91
SD - Total Time[b]	8.92	3.75	1.09 - 60.07	.10
Percent Correct	99%	3%	90 - 100%	-.16
Choice RT Minus Simple RT	**Mean**	**SD**	**Range**	**Rxx[a]**
Decision Time[b]	7.68	8.79	-43.70 - 33.99	.86
Total Time[b]	10.37	11.15	-44.92 - 38.71	.79

[a] Rxx = odd-even correlations corrected with the Spearman-Brown formula.

[b] Values reported are in hundredths of a second. Statistics are based on analysis of all 15 items of the test.

stimulus set then appears on the screen. A stimulus ontains one, two, three, four, or five objects (letters). Following a .5-second or 1.0-second display period, the stimulus set disappears and, after a delay, the probe item appears. Presentation of the probe item is delayed by either 2.5 seconds or 3.0 seconds. When the probe appears, the subject must decide whether or not it appeared in the stimulus set. If the item was present in the stimulus set, the subject removes his or her hands from the home buttons and strikes the white key. If the probe item was not present, the subject strikes the blue key.

Parameters of interest include, first, stimulus set length, or number of letters in the stimulus set. Values for this parameter range from one to five. The second parameter, observation period and probe delay period, includes two levels. The first is described as long observation and short probe delay; time periods are 1.0 second and 2.5 seconds, respectively. The second level, short observation and long probe delay, includes periods of .5 second and 3.0 seconds, respectively. The third parameter, probe status, indicates that the probe is either in the stimulus set or not in in the stimulus set.

Test Characteristics. Table II.10 provides descriptive information for the Memory Search Test. Subjects were allowed very few time-outs (mean = .17, SD = .80) and provided about five invalid responses (range 0 - 28). Overall, total percentage correct is 90. However, the range of percent-correct values, 44 to 100, indicates that at least one subject was performing at a lower-than-chance level.

Test Scores. Table II.10 provides information for the total time score, which was computed and then plotted against item length, defined as the number of letters in the stimulus set. These plots indicated that decision and total time produce very similar profiles, whereas movement time results in a nearly flat profile. Since decision time and total time yield similar information and movement time appears to serve as a constant, we could have used either decision or total reaction time to compute scores on this measure. We elected to use total reaction time.

Subjects receive scores on the following measures:

- Slope and Intercept - These values are obtained by regressing mean total reaction time (correct responses only) against item length. In terms of processing efficiency, the slope represents the average increase in reaction time with an increase of one object in the stimulus set; the lower the value, the faster the access. The intercept represents all other processes not involved in memory search, such as encoding the probe, determining whether or not a match has been found, and executing the response.

- Percent Correct - This value is used to screen subjects completing the test. For example, in Table II.10 we indicated that one subject correctly answered 44% of the items. Computing the above scores for this subject (i.e., slope and intercept) would be meaningless. Percent-correct scores are used to identify subjects performing at very low levels, thereby precluding computation of the above scores.

- Grand Mean - This value is calculated by first computing the mean reaction time (correct responses only) for each level of stimulus set length (i.e., one to five). The mean of these means is then computed.

Table II.10 contains the mean, standard deviation, range, and reliability estimates for each of the scores. Note that all values except the slope yield fairly high internal consistency.

Table II.10

Memory Search Test: Fort Lewis Pilot Test

Test Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	3.06	.76	1.64 - 5.81
Time to Complete Test (minutes)	9.00	.54	8.37 - 11.71
Total Test Time (minutes)	12.07	1.06	10.43 - 17.52
Time-Outs (number per person)	.17	.80	0 - 8
Invalid Responses (number per person)	4.86	4.72	0 - 28

Test Scores[a]	Mean	SD	Range	Rxx[b]
Slope[c]	7.19	6.14	-12.70 - 41.53	.54
Mean Total Time[c]	97.53	30.38	44.91 - 230.97	.84
SD - Total Time[c]	119.05	29.84	67.71 - 262.35	.88
Percent Correct	89%	10%	44 - 100%	.95

[a]See text for explanation of these measures.

[b]Rxx = odd-even correlation corrected with the Spearman-Brown formula.

[c]Values reported are in hundredths of a second. Statistics are based on an analysis of items answered correctly. (There were 50 items on the test.)

Modifications for Fort Knox Field Test. Results from an analysis of variance conducted for the three parameters were used to modify this test for the Fort Knox field test. Total reaction time served as the dependent variable for this measure. These data indicated that the two levels of observation period and probe delay yielded no significant differences in reaction time ($F = .27$; $p<.60$). For stimulus set length, levels one to five, mean reaction time scores differed significantly ($F = 84.35$; $p<.001$). This information confirms results reported in the literature, which suggest that reaction time increases as stimulus set length increases. Finally, for probe status, in or not in, mean reaction time scores also differed significantly ($F = 74.24$; $p<.001$). These values indicate that subjects, on the average, require more time to determine that a probe is not in the set than to determine that the probe is contained in the set. Results also indicated a significant interaction between stimulus length and probe status ($F = 7.46$; $p<.001$).

This information was used to modify the Memory Search Test. For example, stimulus set length had yielded significant mean reaction time score differences for the five levels. Mean reaction time for levels two and four, however, differed little from levels three and five, respectively. Thus, items containing stimulus sets with two and four letters were deleted from the test file.

Although observation period and probe delay parameters produced non-significant results, we concluded that different values for probe delay may provide additional information about processing and memory. For example, in literature in this area, researchers suggest that subjects begin with a visual memory of the stimulus objects. After a very brief period, .5 second, the visual memory begins to decay. To retain a memory of the object set, subjects shift to an acoustic memory; that is, subjects rehearse the sounds of the object set and recall its contents acoustically (Thorson, Hochhaus, & Stanners, 1976). Therefore, we changed the two probe delay periods to .5 second and 2.5 seconds. These periods are designed to assess the two hypothesized types of short-term memory--visual and acoustic.

Finally, half of the items included in the test were modified to include unusual or unfamiliar objects--symbols, rather than letters. In part, the rationale for using letters or digits involves using overlearned stimuli so that novelty of the stimulus does not impact on processing the material. We elected, however, to include a measure of processing and recalling novel or unusual material, primarily because Army recruits do encounter and are required to recall stimuli that are novel to them, especially during their initial training. Thus, one-half of the test items ask subjects to observe and recall unfamiliar symbols rather than letters.

The test then, as modified, contains 48 items--one half consisting of letters and the other half of symbols. Within each item type, three levels of stimulus length are included. That is, for items with letter stimulus sets, there are eight items with a single letter, eight with three, and eight with five letters. The same is done for items containing symbols. Within each of the stimulus length sets, four items include a 5-second probe delay and four contain a 2.5-second probe delay period. Across all items ($N = 48$), probe status is equally mixed between "in" and "not in" the stimulus set. With the test so constructed, the effects of stimulus type, stimulus set length, probe delay period, and probe status can be examined.

Construct - Perceptual Speed and Accuracy

Perceptual speed and accuracy involves the ability to perceive visual information quickly and accurately and to perform simple processing tasks with the stimulus (e.g., make comparisons). This requires the ability to make rapid scanning movements without being distracted by irrelevant visual stimuli, and measures memory, working speed, and sometimes eye-hand coordination.

Perceptual Speed and Accuracy Test

Measures used as markers for the development of the computerized Perceptual Speed and Accuracy (PS&A) Test included such tests as the Employee Aptitude Survey Visual Speed and Accuracy (EAS-4), and the ASVAB Coding Speed and Tables and Graphs tests. The EAS-4 involves the ability to quickly and accurately compare numbers and determine whether they are the same or different, whereas the ASVAB Coding Speed Test measures memory, eye-hand coordination, and working speed. The Tables and Graphs Test requires the ability to obtain information quickly and accurately from material presented in tabular form.

The computer-administered Perceptual Speed and Accuracy Test requires the ability to make a rapid comparision of two visual stimuli presented simultaneously and determine whether they are the same or different. Five different types of stimuli are presented: alpha, numeric, symbolic, mixed, and word. Within the alpha, numeric, symbolic, and mixed stimuli, the character length of the stimulus is varied. Four different levels of stimulus length or "digit" are present: two-digit, five-digit, seven-digit, and nine-digit. Four items are included in each Type by Digit cell. For example, four items are two-digit alphas (e.g., XA). In its original form this test had:

16 two-digit items
16 five-digit items
16 seven-digit items
16 nine-digit items
16 word items

80 total items

Same and different responses were balanced in every cell except one (the four two-digit numeric items were inadvertently constructed to require all "same" responses). Some example items are shown below:

1.	96293	96298	(Numeric five-digit)
2.	+/$\overline{0+}$<>2	+/$\overline{0+}$<>2	(Symbolic seven-digit)
3.	James Braun	James Brown	(Words)

Reaction times were expected to increase with the number of digits included in the stimulus. The rationale for including various types of stimuli was simply that soldiers often encounter various types of stimuli in military positions.

The subject is instructed to hold the home keys down to begin each item, release the home keys upon deciding whether the stimuli are the same or different, and depress a white button if the stimuli are the same or a blue button if the stimuli are different. The measures obtained are: response hand, percent correct, total reaction time, decision time, movement time, time for instructions, and total test time.

Test Characteristics. The computerized Perceptual Speed and Accuracy Test was administered to 112 individuals at Fort Lewis. Some of the overall test characteristics are shown in Table II.11.

Table II.11

Overall Characteristics of Perceptual Speed and Accuracy Test: Fort Lewis Pilot Test

	Mean	SD	Range
Time Spent on Instructions (minutes)	2.36	.59	1.37 - 4.30
Time Spent on Test Portion (minutes)	7.82	1.04	5.82 - 12.41
Total Testing Time (minutes)	10.18	1.37	7.45 - 14.88
Time-Outs (number per person)	9.57	6.17	0 - 35
Invalid Responses (number per person)	.94	1.20	0 - 6

Two two-way analyses of variance were performed on reaction times for correct responses. The first was a Type (4 levels) by Digit (4 levels) ANOVA of total reaction times. The results showed significant main effects for Type [F (3,333) = 11.99, p<.001], Digits [F (3,333) = 871.46, p<.001], and their interaction [F (9,999) = 44.14, p<.001]. The second ANOVA was on movement times. Pure movement time should be a constant when response hands are balanced. The results suggested that subjects were still making their decision about the stimuli after releasing the home keys. That is, the movement time ANOVA for Type by Digits yielded a significant main effect for Digits [F (3,333) = 19.94, p<.001]. The interaction of Digits and Type was also significant [F (9,999) = 7.22, p<.001].

The implications of these results are:

- Scores should be formed on total reaction times (for correct responses) instead of decision times because subjects appear to continue making a decision after releasing the home keys.

- Means should be computed separately for each set of times with a particular digit level (i.e., two, five, seven, and nine). Number of digits had a greater effect on mean reaction time than did type. Since only correct response reaction times are being used, subjects could raise their scores on a pooled reaction time by simply not responding to the nine-digit items. Thus, the mean reaction times to correct responses for each digit level should be equally weighted. The grand mean of the mean reaction times for each digit level was computed.

- The nine-digit symbolic items were probably too easy. Mean reaction times for the nine-digit symbolic items were substantially less than those for the other nine-digit items. Further inspection of the items showed that some of the items were probably being processed in "chunks" (e.g., <<++++*//).

- Total reaction times for correct responses could be regressed on digit and intercepts and slopes computed for individuals by means of repeated measures regression (i.e., the trend appeared to be linear).

Test Scores. As a whole, the scores on the computer-administerd Perceptual Speed and Accuracy Test were quite reliable (see Table II.12). Reliability coefficients ranged from .85 for the intercept of the regression of total reaction time on digits to .97 for the grand mean of the mean reaction times for the four non-word categories and the word category.

Interrelationships Among PS&A Scores. Ideally, efficient performance on the PS&A Test would produce a low intercept, a low slope, and high accuracy, combined with a fast grand mean reaction time score. Data analyzed from the Fort Lewis testing (N = 112) suggest that this relationship may occur infrequently. As shown in Table II.13, the relationship of the slope with the intercept is negative. That is, low intercepts tend to correspond with steep slopes. However, it is possible that individuals who obtained low intercepts simply had more "room" to increase their reaction times within the 7-second time limit, thus increasing their slope scores. Since high intercept values were related to slower grand mean reaction times, as well as less accurate performance, and more "time-outs" occurred on the nine-digit items, it is likely that the 7-second time limit produced a ceiling effect.

The high positive correlation between the slope and accuracy suggests that performing accurately is related to a substantial increase in reaction time as the stimuli increase in length. Steeper slopes also correspond with slower grand mean reaction times. These slower reaction times were also related to higher accuracy.

Table II.12

Scores From Perceptual Speed and Accuracy Test: Fort Lewis Pilot Test

Score[a]	Mean	SD	Range	Rxx[b]
Grand Mean of Mean Reaction Times for Non-Word Items	279.99	57.97	85.67 - 386.49	.97
Mean Reaction Time for Word Items	351.74	68.39	198.64 - 518.64	.91
Grand Mean of Mean Reaction Times for Word and Non-Word Items	294.22	57.13	109.34 - 412.75	.97
Intercept	89.37	36.48	12.99 - 210.34	.85
Slope	33.14	9.78	-.75 - 52.11	.89
Percent Correct	86.9%	8.0%	56.3 - 100%	

[a] Reaction Time values are in hundredths of a second and are based on analysis of items answered correctly. (There were 80 items on the test.)

[b] Split halves (odd-even) reliability estimates, Spearman-Brown corrected.

Table II.13

Intercorrelations Among Perceptual Speed and Accuracy Test Scores: Fort Lewis Pilot Test

Score	Intercept	Slope	Percent Correct
Slope	-.27[a]		
Percent Correct	-.26[b]	.64[a]	
Grand Mean[c]	.36[b]	.79[a]	.45[a]

[a] $p \leq .001$

[b] $p \leq .003$

[c] Grand mean reaction time in this section refers to:

$$\text{Grand Mean} = \frac{\bar{X}_2\text{-digits} + \bar{X}_5\text{-digits} + \bar{X}_7\text{-digits} + \bar{X}_9\text{-digits} + \bar{X}\ \text{words}}{5}$$

Modifications for Fort Knox Field Test. Several changes were made to this test. A reduction in the number of items was desirable in order to cut down the testing time, and the reliability of the test scores indicated that the test length could be considerably reduced without causing the reliabilities to fall below acceptable levels (see Table II.12).

Item deletion was accomplished in two ways. First, all the seven-digit items were deleted (16 items). Such deletions should have little effect on the test scores, since the relationship between number of digits and reaction time is linear, and the items containing two, five, and nine digits should provide sufficient data points.

Second, four items were deleted from each of the remaining three digit categories (two, five, and nine) and from the "word" items. Thus, 16 more items were deleted. The following factors were considered in selecting items for deletion:

- Item intercorrelations within stimulus type and digit size were examined. In many cases, one item did not correlate highly with the others. Items that produced the lowest intercorrelations were deleted. Use of this criterion resulted in 13 item deletions.

- When item interrelations did not differ substantially, accuracy rates and variances were reviewed. These factors did not indicate any clear candidates for deletion.

- When all the above were approximately equal, the decision to delete an item was based on its correct response (i.e., "same" or "different"). The item which would have caused an imbalance between the responses (if retained) was deleted. This was, in effect, a random selection.

Several other changes were made, either to correct perceived shortcomings or to otherwise improve the test. The symbolic, nine-digit items were modified to make them more difficult. (As previously noted, these items had originally been developed in such a way that the symbols were in "chunks," thus making the items, in effect, much shorter than the intended nine digits; these group symbols were broken up.) Five items were changed so that the correct response was "different" rather than "same" in order to balance type of correct response within digit level. Finally, the time allowed to make a response to an item was increased from 7 seconds to 9 seconds in order to give subjects sufficient time to respond, especially for the more difficult items.

The revised test, then, contained 48 items; 36 were divided into 12 Type (alpha, numeric, symbolic, mixed) by Number of Digits (two, five, nine) cells, and 12 were "word" items.

Target Identification Test

The Target Identification Test is a measure of the perceptual speed and accuracy construct. The objects perceived are meaningful figures, however, rather than being made up of numbers, letters, or symbols. In this test,

each item shows a target object near the top of the screen and three labeled stimuli in a row near the bottom of the screen. Examples are shown in Figure II.16. The subject is to identify which of three stimuli represents the same object as the target and to press, as quickly as possible, the button (blue, yellow, or white) that corresponds to that object.

The objects shown are based on military vehicles and aircraft as shown on the standard set of flashcards used to train soldiers to recognize equipment presently being used by various nations. Twenty-two drawings of objects were prepared.

Several parameters were varied in the stimulus presentation. In addition to type of object, a second parameter was the position of the correct response--on the left, in the middle, or on the right side of the screen. A third was the orientation of the target object--whether the object is "facing" in the same direction as the stimuli or in the opposite direction.

A fourth parameter was the angle of rotation (from horizontal) of the target object. Seven different angular rotations were used for the Fort Lewis administration: 0°, 20°, 25°, 30°, 35°, 40°, and 45°. The fifth parameter was the size of the target object. Ten different levels of size reduction were used in the Fort Lewis administration: 40%, 50%, 55%, 60%, 65%, 75%, 80%, 85%, 90%, and 100%. Fifty percent reduction means that the target object was half the size of the stimulus objects at the bottom of the screen; 100% is full size.

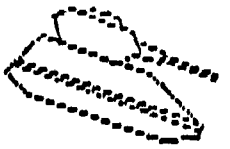
There was no intention of creating a test that had items tapping each cell of a crossed design for these five parameters. Instead, we viewed this tryout of the test as an opportunity to explore a number of different factors that could conceivably affect test performance. A total of 44 items were included on the test.

Test Characteristics. Table II.14 shows data from the Fort Lewis pilot test of the Target Identification Test. The lower part of the table shows data from the two measures of concern: total reaction time and percent correct. The test was conceived as a speeded test, in the sense that each item could be answered correctly if the subject took sufficient time to study the items and, therefore, the reaction time measure was intended to show the most variance. The data show that these intentions were achieved.

Score Variables. As noted above, the primary scores for this test were total reaction time (includes both decision and movement times) for correct responses, and the percent of responses that were correct. Total reaction time was used rather than decision time because it seems to be more ecologically valid (i.e., the Army is interested in how quickly a soldier can perceive, decide, and take some action and not just in the decision time). Also, analyses of variance showed similar results for the two measures.

Modifications for the Fort Knox Field Test. The revised test consisted of 48 items instead of 44. Two parameters of the test were left unchanged--position of the object that "matched" the target and direction in which the target object faced--even though analyses of the Fort Lewis data indicated that opposite-facing targets appeared to be more difficult and the middle position of the correct object was slightly "easier."

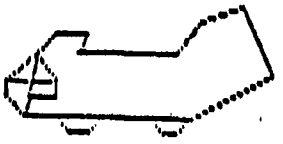
EXAMPLE 1.



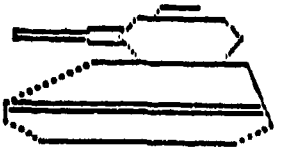
TARGET



BLUE



YELLOW

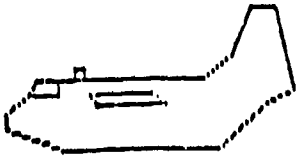


WHITE

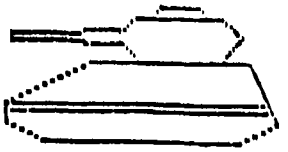
EXAMPLE 2.



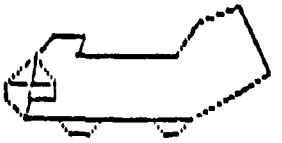
TARGET



BLUE



YELLOW



WHITE

Figure II.16. Graphic Displays of example items from the computer-administered Target Identification Test.

Table II.14

Target Identification Test: Fort Lewis Pilot Test

Descriptive Characteristics	Mean	SD	Range	
Time to Read Instructions (minutes)	2.01	1.04	1.10 - 9.21	
Time to Complete Test (minutes)	3.61	0.45	2.96 - 5.46	
Total Test Time (minutes)	5.62	1.23	4.12 - 12.81	
Time-Outs (per person)	.06	.28	0 - 2	
Invalid Responses (per person)	3.20	3.61	0 - 29	
Test Scores	**Mean**	**SD**	**Range**	**R_{xx}[a]**
Total Reaction Time[b]	218.51	68.75	113.10 - 492.95	.97
Percent Correct	92.6%	8.3%	34.1 - 100%	.78

[a] Reliability estimates computed using odd-even procedure with Spearman-Brown correction.

[b] In hundredths of a second.

Three parameters were changed. The objects to be matched with the target were made to be all from one type (helicopters or aircraft or tanks, etc.) or from two types, rather than from one, two, or three. This was done because analyses showed the "three-type" items to be extremely easy. Rotation angles were reduced from seven levels to just two, 0° and 45°, since analyses showed that angular rotations near 0° had very little effect on reaction time.

Finally, the size parameter was radically changed. The target object was either 50% of the stimulus objects, or was made to "move." The "moving" items were made to appear initially on the screen as a very small dot, completely indistinguishable, and then to quickly and successively disappear and reappear, slightly enlarged in size and slightly to the left (or right, depending on the side of the screen where the target initially appeared) of the prior appearance. Thus, the subject had to observe the moving and enlarging target until certain of matching it to one of the stimulus objects. These "moving" items were thought to represent greater ecological or content validity, but still to be a part of the figural perception construct.

Construct - Psychomotor Precision

This construct reflects the ability to make muscular movements necessary to adjust or position a machine control mechanism. This ability applies to both anticipatory movements (i.e., where the subject must respond to a stimulus condition that is continuously changing in an unpredictable manner) and controlled movements (i.e., where the subject must respond to a stimulus condition that is changing in a predictable fashion, or making only a relatively few discrete, unpredictable changes). Psychomotor precision thus encompasses two of the ability constructs identified by Fleishman and his associates, control precision and rate control (Fleishman, 1967).

Performance on tracking tasks is very likely related to psychomotor precision. Since tracking tasks are an important part of many Army MOS, development of psychomotor precision tests was made a high priority. The Fort Lewis computer battery included two measures of this ability.

Target Tracking Test 1

Target Tracking Test 1 was designed to measure subjects' ability to make fine, highly controlled movements to adjust a machine control mechanism in response to a stimulus whose speed and direction of movement are perfectly predictable. Fleishman labeled this ability control precision. The Rotary Pursuit Test (Melton, 1947) served as a model for Target Tracking Test 1.

For each trial of this pursuit tracking test, subjects are shown a path consisting entirely of vertical and horizontal line segments. At the beginning of the path is a target box, and centered in the box are crosshairs. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject's task is to keep the crosshairs centered within the target at all times. The subject uses a joystick, controlled with one hand, to control movement of the crosshairs.

Several item parameters vary from trial to trial. These include the speed of the crosshairs, the maximum speed of the target, the difference between crosshairs and target speeds, the total length of the path, the number of line segments comprising the path, and the average amount of time the target spends traveling along each segment. Obviously, these parameters are not all independent; for example, crosshairs speed and maximum target speed determine the difference between crosshairs and target speeds.

For the Fort Lewis battery, subjects were given 18 test trials. Three of the 18 paths were duplicates (the paths for trials 15-17 were identical to the paths for trials 1, 2, and 7). Ignoring these duplicates, the test was constructed so that the trials at the beginning of the test were easier than trials at the end of the test.

Test Characteristics. Table II.15 presents data for Target Tracking Test 1 based on the Fort Lewis pilot test. The 18 trials of the test required 9.07 minutes to complete. Since all subjects received the same set of paths, there was virtually no variability.

Table II.15

Target Tracking Test 1: Fort Lewis Pilot Test

Descriptive Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	1.20	.43	.33 - 3.09
Time to Complete Test (minutes)	9.07	.02	9.05 - 9.12
Total Test Time (minutes)	10.27	.43	9.42 - 12.17

Test scores	Mean	SD	Range	R_{xx}[a]
Distance[b]	1.44	.45	.95 - 3.40	.97

[a] Spearman-Brown corrected split-half reliability for odd-even trials.

[b] Square root of the average within-trial distance (horizontal pixels) from the center of the target to the center of the crosshairs, averaged across all 18 trials (or items) on the test.

Test Scores. Two classes of measures were investigated: (a) tracking accuracy and (b) improvement in tracking performance, based on the three duplicate paths included in the test. Two tracking accuracy measures were investigated, time on target and distance from the center of crosshairs to the center of the target. Kelley (1969) demonstrated that distance is a more reliable measure of tracking performance than time on target. Therefore, the test program computes the distance[3] from the crosshairs to the center of the target several times each second, and then averages these distances to derive an overall accuracy score for that trial.

Subsequently, when the distribution of subjects' scores on each trial was examined, it was found that the distribution was highly positively skewed. Consequently, the trial score was transformed by taking the square root of the average distance. As a result, the distribution of subjects' scores on each trial was more nearly normal. These trial scores were then averaged to determine an overall tracking accuracy score for each subject.

[3]The Compaq video screen is divided into 200 pixels vertically and 640 pixels horizontally, with each vertical pixel equivalent to 3 horizontal pixels. All distance measures were computed in horizontal pixel units.

Prior to the Fort Lewis test, it was expected that subjects' tracking proficiency would improve considerably over the course of the test. That was one of the reasons that initial test trials were designed to be easier than final test trials. However, analyses of the Fort Lewis data revealed that subjects' performance on trials 1, 2, and 7 actually differed little from their performance on trials 15-17. Therefore, it was decided that no measure of improvement in tracking performance would be computed.

The internal consistency reliability of the accuracy score was computed by comparing mean accuracy scores for odd and even trials. The Spearman-Brown corrected reliability was .97.

Four one-way analyses of variances were executed to determine how tracking accuracy was affected by average segment length, average time required for the target to travel a segment, maximum crosshairs speed, and difference between maximum crosshairs speed and target speed. All four item parameters were significantly related to accuracy score, with crosshairs speed accounting for the most variance and difference between target and crosshairs speed the least. All four parameters were highly intercorrelated.

Modifications for the Fort Knox Field Test. Several changes were made in the paths comprising this test for the Fort Knox field test. First, all paths were modified so that each would run for the same amount of time (approximately .36 minute). The primary reason for this change was that the program computes distance between the crosshairs and target a set number of times each second. If all paths run the same amount of time, then the accuracy measure for each trial will be based on the same number of distance assessments.

Second, three item parameters were identified to direct the format of test trials: maximum crosshairs speed, difference between maximum crosshairs speed and target speed, and number of path segments. Given these parameters and the constraint that all trials run a fixed amount of time, the values of all other item parameters (e.g., target speed, total length of the path) can be determined. Three levels were identified for each of the three parameters. These were completely crossed to create a 27-item test. Items were then randomly ordered. These procedures for item development should alleviate previous problems interpreting test results in light of correlated item parameters.

Third, in spite of these changes, which added 50% more trials to the test, testing time was actually reduced slightly (25 seconds less, it was estimated), because of the standardization of trial time.

Target Shoot Test

The Target Shoot Test was modeled after several compensatory and pursuit tracking tests used by the AAF in the Aviation Psychology Program (e.g., the Rate Control Test). The distinguishing feature of these tests is that the target stimulus moves in a continuously changing and unpredictable speed and direction. Thus, the subject must attempt to anticipate these changes and respond accordingly.

For the Target Shoot Test, a target box and a crosshairs appear in different locations on the computer screen. The target moves about the screen in an unpredictable manner, frequently changing speed and direction. The subject controls movement of the crosshairs via a joystick. The subject's task is to move the crosshairs into the center of the target, and when this had been accomplished, to press a button on the response pedestal to "fire" at the target. The subject's score on a trial is the distance from the center of the crosshairs to the center of the target at the time the subject fires. The test consists of 40 trials.

Several item parameters were varied from trial to trial. These parameters included the maximum speed of the crosshairs, the average speed of the target, the difference between crosshairs and target speeds, the number of changes in target speed (if any), the number of line segments comprising the path of each target, and the average amount of time required for the target to travel each segment. These parameters are not all independent, of course. Moreover, the nature of the test creates a problem in characterizing some trials since a trial terminates as soon as the subject fires at the target. Thus, one subject may see only a fraction of the line segments, target speeds, etc. that another subject sees.

Test Characteristics. Table II.16 presents data based on the Fort Lewis pilot test.

Test Scores. Three measures were obtained for each trial. Two were measures of firing accuracy: (a) the distance from the center of the crosshairs to the center of the target at the time of firing, and (b) whether the subject "hit" or "missed" the target. The two were very highly correlated, though the former provides quite a bit more information about firing accuracy than the latter. Therefore, distance was retained as the accuracy measure; distances were averaged across trials to obtain an overall accuracy score. The third measure was a speed measure which represented the time from trial onset until the subject fired at the target.

Split-half reliability across odd-even trials was computed for the two accuracy measures.

Changes for Fort Knox. The test was not modified for the Fort Knox administration.

Construct - Multilimb Coordination

The multilimb coordination construct reflects 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 hand and one foot). The ability does not apply to tasks in which trunk movement must be integrated with limb movements. It is most common in tasks where the body is at rest (e.g., seated or standing) while two or more limbs are in motion.

In the past, measures of multilimb coordination have shown quite high validity for predicting job and training performance, especially for pilots (Melton, 1947).

Table II.16

Target Shoot Test: Fort Lewis Pilot Test

Descriptive Characteristics	Mean	SD	Range
Time to Read Instructions (minutes)	1.58	.61	.51 - 5.10
Time to Complete Test (minutes)	2.22	.23	1.81 - 3.29
Total Test Time (minutes)	3.80	.68	2.71 - 7.58
No. of Trials Without Firing[a]	2.77	3.97	0 - 40

Test Scores	Mean	SD	Range	R_{xx}[b]
Distance[c]	2.83	.52	1.93 - 7.03	.93
Percent of Hits[a]	58	13	0 - 83	.78

[a]One subject failed to fire at any targets. Excluding this subject, mean, SD, and range for number of trials without firing were 2.43, 1.78, and 0-8, respectively; mean, SD, and range for percent of hits were 59, 12, and 13-83, respectively.

[b]Spearman-Brown corrected split-half reliability for odd-even trials.

[c]Square root of the distance (horizontal pixels) from the center of the target to the center of the crosshairs at the time of firing, averaged across all trials in which the subject fired at the target. (There were a total of 40 trials or items on the test.)

Target Tracking Test 2

Target Tracking Test 2 is modeled after a test of multilimb coordination developed by the AAF, the Two-Hand Cooridnation Test, which required subjects to perform a pursuit tracking task. Horizontal and vertical movements of the target-follower were controlled by two handles. Validity estimates of this test for predicting AAF pilot training success were mostly in the .30s.

Target Tracking Test 2 is very similar to the Two-Hand Coordination Test. For each trial subjects are shown a path consisting entirely of vertical and horizontal lines. At the beginning of the path is a target box, and centered in the box are crosshairs. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject manipulates two sliding resistors to control movement of the crosshairs. One

resistor controls movement in the horizontal plane, the other in the vertical plane. The subject's task is to keep the crosshairs centered within the target at all times.

This test and Target Tracking Test 1 are virtually identical except for the nature of the required control manipulation. For Target Tracking Test 1 crosshairs movement is controlled via a joystick, while for Target Tracking Test 2 crosshairs movement is controlled via the two sliding resistors. For the Fort Lewis battery, the same 18 paths were used in both tests, and the value of the crosshairs and target speed parameters was the same in both sets of trials. The only other difference between the two tests was that subjects were permitted three practice trials for Target Tracking Test 2.

Test Characteristics. The descriptive data are shown in Table II.17.

Table II.17

Target Shoot Test: Fort Lewis Pilot Test.

Descriptive Characteristics	Mean	SD	Range	
Time to Read Instructions (minutes)	3.58	.68	2.39 - 6.38	
Time to Complete Test (minutes)	9.09	.02	9.03 - 9.13	
Total Test Time (minutes)	12.67	.68	11.50 - 15.48	
Test Scores	**Mean**	**SD**	**Range**	**R_{xx}[a]**
Distance[b]	2.02	.64	0 - 4.01	.97

[a]Spearman-Brown corrected split-half reliability for odd-even trials.

[b]Square root of the distance (horizontal pixels) from the center of the target to the center of the crosshairs, averaged across all 18 trials (or items) on the test.

Test Scores. The same score was used for this test as for Tracking Test 1; that is, the square root of the average within-trial distance from the center of the crosshairs to the center of the target, averaged across all trials.

Four one-way analyses of variance were executed to determine the effects of average segment length, average time required for the target to travel a segment, maximum crosshairs speed, and difference between maximum crosshairs speed and target speed on tracking accuracy. All four item parameters were significantly related to accuracy score, with crosshairs speed accounting for the most variance and average segment length for the least. It should be noted again that all four parameters were highly intercorrelated.

Modifications for the Fort Knox Field Test. Changes in Target Tracking Test 2 for Fort Knox mirrored those made for Target Tracking Test 1. Test trials were changed completely, and the number of items was increased from 18 to 27. However, the items are not the same as those presented for Target Tracking Test 1. This was expected to reduce the correlations between these tests to some extent.

Construct - Number Operations

This construct involves the ability to perform, quickly and accurately, simple arithmetic operations such as addition, subtraction, multiplication, and division.

The current ASVAB includes a numerical operations test containing 50 very simple arithmetic problems with a 3-minute time limit. Because of low item difficulty and the speeded nature of the test, correlations with other ASVAB subtests indicate that Numerical Operations is most strongly related to Coding Speed--a measure of perceptual speed and accuracy. The present military-wide selection and classification battery, then, measures very basic number operations abilities which appear very similar to perceptual speed and accuracy abilities.

The test designed to assess number operations abilities was not completed prior to the Fort Lewis pilot test. Therefore, no data were available to evaluate this measure prior to field testing.

Number Memory Test

This test was modeled after a number memory test developed by Dr. Raymond Christal at AFHRL. The basic difference between the AFHRL test and the Number Memory Test concerns pacing of the number items. The former uses machine-paced presentation, while the latter is self-paced. Both, however, require subjects to perform simple number operations such as addition, subtraction, multiplication, and division and both involve a memory task.

In the Number Memory Test, subjects are presented with a single number on the computer screen. After studying the number, the subject is instructed to push a button to receive the next part of the problem. When the subject presses the button, the first part of the problem disappears and another number, along with an operation term such as Add 9 or Subtract 6 then appears. Once the subject has combined the first number with the second, he or she must press a button to receive the third part of the problem. Again, the second part of the problem disappears when the subject presses the

button. This procedure continues until a solution to the problem is presented. The subject must then indicate whether the solution presented is true or false. An example number operation item appears below.

Item Set	8 +6 -3 x2 -4	
Probe	Is 16 the correct answer?	
Response	T White	F Blue

Test items vary with respect to number of parts--four, six, or eight--contained in the single item. Items also vary according to the delay between item part presentation or interstimulus delay period. One-half of the items include a brief delay (.5 second) while the other half contain a lengthier delay (2.5 seconds). The test contained 27 items.

This test is not a "pure" measure of number operations, since it also is designed to bring short-term memory into play.

As noted, the test was not administered at Fort Lewis. Analyses planned for data from the Fort Knox field test administration included the impact of item length (four, six, or eight) and interstimulus delay (.5 second or 2.5 seconds) on reaction time and percent correct, as well as comparisons of mean reaction time scores for item parts requiring addition, subtraction, multiplication, and division. These data will be used to identify the measures for scoring subject responses.

Construct - Movement Judgment

Movement judgment is the ability to judge the relative speed and direction of one or more moving objects in order to determine where those objects will be at a given point in time and/or when those objects might intersect.

Movement judgment was not one of the constructs identified and targeted for test development by the literature review or expert judgments. However, a suggestion by Dr. Lloyd Humphreys, one of Project A's scientific advisors, and job observations we conducted at Fort Stewart, Fort Bragg, Fort Bliss, Fort Sill, and Fort Knox led us to conclude that movement judgment is potentially important for job performance in a number of combat MOS.

Cannon Shoot Test

As part of its Aviation Psychology Program, the AAF became interested in motion, distance, and orientation judgment and instituted development of a battery of motion picture and photograph tests (Gibson, 1947). One of these

tests was the Estimate of Relative Velocities Test, a paper-and-pencil measure. Each trial consisted of four frames. In each frame, two airplanes were shown flying along the same path in the same direction. In each subsequent frame, the trailing plane edged nearer the lead plane. The subject's task was to indicate on the final frame where the planes would intersect. Validities of this test for predicting pilot training success averaged approximately .18 (Gibson, 1947). The present test was designed to test the construct that seems to underlie the Estimate of Relative Velocities Test.

The Cannon Shoot Test measures subjects' ability to fire at a moving target in such a way that the shell hits the target when the target crosses the cannon's line of fire. At the beginning of each trial, a stationary cannon appears on the video screen. The starting position of this cannon varies from trial to trial. The cannon is "capable" of firing a shell, which travels at a constant speed on each trial. Shortly after the cannon appears, a circular target moves onto the screen. This target moves in a constant direction at a constant rate of speed throughout the trial, though the speed and direction vary from trial to trial. The subject's task is to push a response button to fire the shell so that the shell intersects the target when the target crosses the shell's line of fire.

Three parameters determine the nature of each test trial. The first is the angle of the target movement relative to the position of the cannon; 12 different angles were used. The second is the distance from the cannon to the impact point (i.e., the point at which the shell crosses the cannon's line of fire); four different distance values were used. The third parameter was the distance from the impact point to the fire point (i.e., the point at which the subject must fire the shell in order to hit the center of the target); there were also four values for this distance parameter. The last two parameters determine the speed of the target--that is, given a fixed shell speed, impact point, and fire point, the speed of the target is established.

If a completely crossed design had been used, it would have necessitated a minimum of 192 trials (i.e., 12x4x4=192). Instead, a Latin square design was employed, so the current version of the test includes only 48 trials. Three measures are assessed on each trial: (a) whether the shell hits or misses the target; (b) the distance from the shell to the center of the target at the time the target crosses the impact point; and (c) the distance from the center of the target to the fire point at the time the shell is fired. The Fort Knox field test data will be analyzed to determine which of these three measures is most reliable. This test was not administered at Fort Lewis.

Summary of Pilot Test Results for Computer-Administered Tests

Table II.18 shows the means, standard deviations, and split-half reliabilities for 24 scores computed from eight computer tests administered at the Fort Lewis pilot test, and Table II.19 shows the intercorrelations between computer test scores. Table II.20 shows the correlations between computer-administered test scores and cognitive paper-and-pencil test scores.

Table II.18

Means, Standard Deviations, and Split-Half Reliability Coefficients for 24 Computer Measure Scores Based on Fort Lewis Pilot Test Data (N = 112)

	Mean	SD	Split-Half[a] Type of Test
SIMPLE REACTION TIME (10 Items)			
Mean Decision Time (hs)[b]	29.25	8.10	.92
Mean Total Reaction Time (hs)	55.92	13.86	.94
Trimmed Standard Deviation (hs)	11.79	16.80	.66
Percent Correct	99	3	-.01
CHOICE REACTION TIME (15 Items)			
Mean Decision Time (hs)	36.78	7.75	.94
Mean Total Reaction Time (hs)	65.98	10.39	.91
Standard Deviation (hs)	8.92	3.75	.10
Percent Correct	99	3	-.16
DIFFERENCE IN SIMPLE & CHOICE REACTION TIME			
Decision Time (hs)	7.68	8.79	.86
Total Time (hs)	10.37	11.15	.79
SHORT-TERM MEMORY (50 Items)			
Intercept (hs)	97.53	30.28	.84
Slope (hs)	7.19	6.14	.54
Percent Correct	90	10	.95
Grand Mean (hs)	119.05	29.84	.88
PERCEPTUAL SPEED & ACCURACY (80 Items)			
Intercept (hs)	89.37	36.48	.85
Slope (hs)	33.14	9.78	.89
Percent Correct	87	8	.81
Grand Mean (hs)	294.22	57.13	.97
TARGET IDENTIFICATION (44 Items)			
Mean Total Time (hs)	218.51	68.75	.97
Percent Correct	93	8	.78
TARGET TRACKING 1 (18 Items)			
Mean Distance (m $\sqrt{\text{m pixels}}$)[c]	1.44	.45	.97
TARGET TRACKING 2 (18 Items)			
Mean Distance (m $\sqrt{\text{m pixels}}$)	2.01	.64	.97
TARGET SHOOT (40 Items)			
Mean Total Distance (m $\sqrt{\text{m pixels}}$)	2.83	.52	.93
Percent "Hits"	58	13	.78

[a] Odd-even item correlation corrected to full test length with the Spearman-Brown formula.

[b] hs = hundredths of seconds.

[c] m $\sqrt{\text{m pixels}}$ = mean of the square root of the mean distance from target, computed across all trials.

Table II.19

Intercorrelations of Dependent Measures Developed From Computer-Administered Tests: Fort Lewis Pilot Test

	Reaction Time						Perpetual Speed and Accuracy			Short-Term Memory				Tracking		Target Shoot	Target Identification	
	Simple - Decision Time	Simple - Total Time	Simple - Total SD	Choice - Decision Time	Choice - Total Time	CRT-SRT Total	PSA Slope	PSA Intercept	PSA Grand Mean	Memory Slope	Intercept	% Correct	Grand Mean	Tracking 1 1-Mean	Tracking 2 2-Mean	Target Shoot Mean Distance	Target ID Mean RT	% Correct
Reaction Time (RT)																		
Simple - Decision Time	**	.85	.71	.36	.37	-.66	-.01	.30	.16	.03	.30	-.11	.29	.20	-.09	-.01	.17	.11
Simple - Total Time		**	.67	.36	.57	-.65	-.04	.45	.22	-.04	.32	-.20	.30	.31	.11	.04	.31	-.12
Simple - Total SD			**	.05	.10	-.69	-.01	.14	.07	-.11	.29	-.14	.22	.08	-.15	-.01	.10	.05
Choice - Decision Time				**	.78	.31	.05	.37	.2[illegible]	.05	.29	-.06	.33	.15	.14	.13	.29	-.08
Choice - Total Time					**	.25	.04	.53	.36	-.02	.41	-.11	.40	.39	.33	.14	.45	-.07
CRT-SRT (Total)						**	.09	-.04	.08	.03	.00	.14	.02	.00	.17	.08	.06	.08
Perceptual Speed & Accuracy																		
Slope							**	-.27	.79	.06	.09	.29	.13	.06	.07	.05	.30	.27
Intercept								**	.35	.04	.44	-.43	.48	.36	.31	.19	.45	-.23
Grand Mean RT									**	.08	.37	.01	.43	.29	.27	.18	.58	.13
Short-Term Memory																		
Slope										**	-.31	-.13	.29	-.06	.02	.03	.13	-.07
Intercept											**	-.33	.82	.25	.17	.10	.45	.09
Percent Correct												**	-.41	-.14	-.24	-.25	-.15	.51
Grand Mean RT													**	.21	.19	.12	.54	.05
Tracking																		
Test 1 - Mean Distance														**	.76	.32	.46	-.10
Test 2 - Mean Distance															**	.47	.44	-.11
Target Shoot																		
Mean Distance																**	.16	-.10
Target Identification																		
Mean RT																	**	.21
Percent Correct																		**

Table II.20

Intercorrelations of Cognitive Paper-and-Pencil Tests and Computer-Administered Tests: Fort Lewis Pilot Test

Computer Tests	Cognitive Paper-and-Pencil Tests									
	Assembling Objects	Object Rotation	Path	Maze	Shapes	Orientation 1	Orientation 2	Orientation 3	Reasoning 3	Reasoning 2
Reaction Time (RT)										
Simple - Decision Time	-01	-03	-10	-23	-10	-05	06	01	-06	04
Simple - Total Time	-10	-15	-23	-39	-21	-23	-09	-17	-20	-14
Simple - Total SD	-01	-01	-10	-13	-07	-05	00	-03	-01	-01
Simple - Percent Correct	01	-07	17	02	04	07	-02	00	08	10
Choice - Decision Time	-09	-12	-17	-28	-21	-18	-17	-15	-12	-15
Choice - Total Time	-22	-27	-23	-47	-32	-36	-26	-29	-25	-25
Choice - Total SD	-20	-12	00	-05	-22	-17	-17	-15	-07	-07
Choice - Percent Correct	07	-10	-01	-05	-05	-08	-10	-05	08	-07
Perceptual Speed & Accuracy										
Slope	16	-12	11	01	-03	09	19	11	14	27
Intercept	-44	-40	-46	-57	-37	-50	-42	-44	-48	-43
Percent Correct	30	09	26	16	17	31	21	25	20	31
Grand Mean RT	-11	-35	-17	-33	-24	-22	-09	-17	-16	-01
Short-Term Memory										
Slope	04	03	-03	13	-02	-10	-04	-04	04	06
Intercept	-22	-30	-24	-40	-17	-26	-08	14	-23	-22
Percent Correct	29	17	46	34	17	31	25	29	32	28
Grand Mean RT	-20	-29	-26	-33	-18	-32	-11	-16	-21	-19
Tracking										
Test 1 - Mean Distance	-27	-45	-39	-52	-41	-39	-29	-39	-38	-30
Test 2 - Mean Distance	-32	-46	-43	-50	-36	-45	-38	-44	-35	-33
Target Shoot										
Mean Distance	-13	-14	-20	-23	-22	-21	-22	-18	-17	-10
Percent Correct	25	27	20	40	30	28	27	27	21	18
Target Identification										
Mean RT	-30	-46	-31	-50	-39	-48	-32	-43	-42	-32
Percent Correct	27	17	24	29	17	11	16	11	26	19

NOTE: Decimals have been omitted.

One concern we had prior to the Fort Lewis pilot test was the extent to which computer measure scores would be affected by differences between testing stations (a testing station is one Compaq computer and the associated response pedestal; six such testing stations were used at Fort Lewis). Recall that we mentioned in Section 1 that differences across testing apparatus and unreliability of testing apparatus had been a problem in World War II psychomotor testing and thereafter. The recent advent of microprocessor technology was viewed as alleviating such problems, at least to some degree.

We ran some analyses of variance to provide an initial look at the extent of this problem with our testing stations. Thirteen one-way ANOVAs were run with testing stations as levels and computer test scores as the dependent variables. We ran separate ANOVAs for white males and non-white males in order to avoid confounding the results with possible subgroup differences. Also, only five testing stations were used since one station did not have enough subjects assigned to it.

Of the 26 ANOVAs, only 1 reached significance at .05 level, about what would be expected by chance. These results were heartening. One reason for these results was the use of calibration software, which adjusted for the idiosyncratic differences of each response pedestal, ensuring a more standardized test administration across testing stations.

The results of the Fort Lewis pilot test of the computer-administered measures in the Pilot Trial Battery were extremely useful. The results showed very high promise for these measures. In addition, the soldiers liked the test battery. Virtually every soldier expressed a preference for the computerized tests compared to the paper-and-pencil tests. We thought there were several resons for this: novelty; the game-like nature of several tests; and the fact that the battery was, in large part, self-paced, allowing each soldier to thoroughly understand the instructions and to work through the battery at his or her own speed.

Field testing of the computer-administered measures is described in Section 6.

Section 5

DEVELOPMENT OF NON-COGNITIVE MEASURES

This section describes the non-cognitive measures developed for the Pilot Trial Battery. All are paper-and-pencil measures and the inventories are intended to assess constructs in the temperament, interests, and life history (biodata) domains.

The discussion of these scales is organized around the two inventories that were developed for Project A. The ABLE (Assessment of Background and Life Experiences) contains items that assess the important constructs of the temperament and life history (biodata) domains. The AVOICE (Army Vocational Interest Career Examination) measures relevant constructs pertaining to vocational interests.

Recall from Part II, Section 1, that the non-cognitive domain of selection information was defined and specified by a three-part strategy. First, a comprehensive literature review was used to generate an exhaustive list of potential non-cognitive indicators of relevant individual differences. On the basis of professional staff judgments, the list was reduced to a non-redundant list of variables. The existing validity evidence was then summarized around this list of variables. A brief summary of those results is presented in Tables II.21, II.22, and II.23.

The second part of the strategy was to include the temperament and biographical variables in the expert judgment forecasts of validity coefficients described in Section 1. The predicted profiles of validity coefficients for each temperament and biographical variable were then intercorrelated and clustered to generate a kind of higher order construct.

The third part of the strategy consisted of examining the empirical covariation matrix generated by the temperament and biographical measures that were included in the Preliminary Battery (i.e., the "off-the-shelf" measures administered to the samples of new recruits in the four MOS in the Preliminary Battery sample). Factor analyses of these data provided additional guidance for how best to define the non-cognitive criterion space in Project A.

All three sources of information were discussed at some length in a series of meetings attended by the relevant project staff and members of the Scientific Advisory Group. The result of these deliberations was an array of constructs that were judged to be the best potential sources of valid selection/classification information of a non-cognitive nature. The linkages among the initial variable array, the constructs chosen for measurement, the variables proposed to reflect them, and the forecasted predictor/criterion correlations are shown in Figure II.17 (Hough, 1984).

Table II.21

Summary of Criterion-Related Validities for Interest Inventories

Criterion	Median r
Training	.28
Job Proficiency	.27
Job Involvement	.30

Table II.22

Summary of Criterion-Related Validities for Biographical Inventories

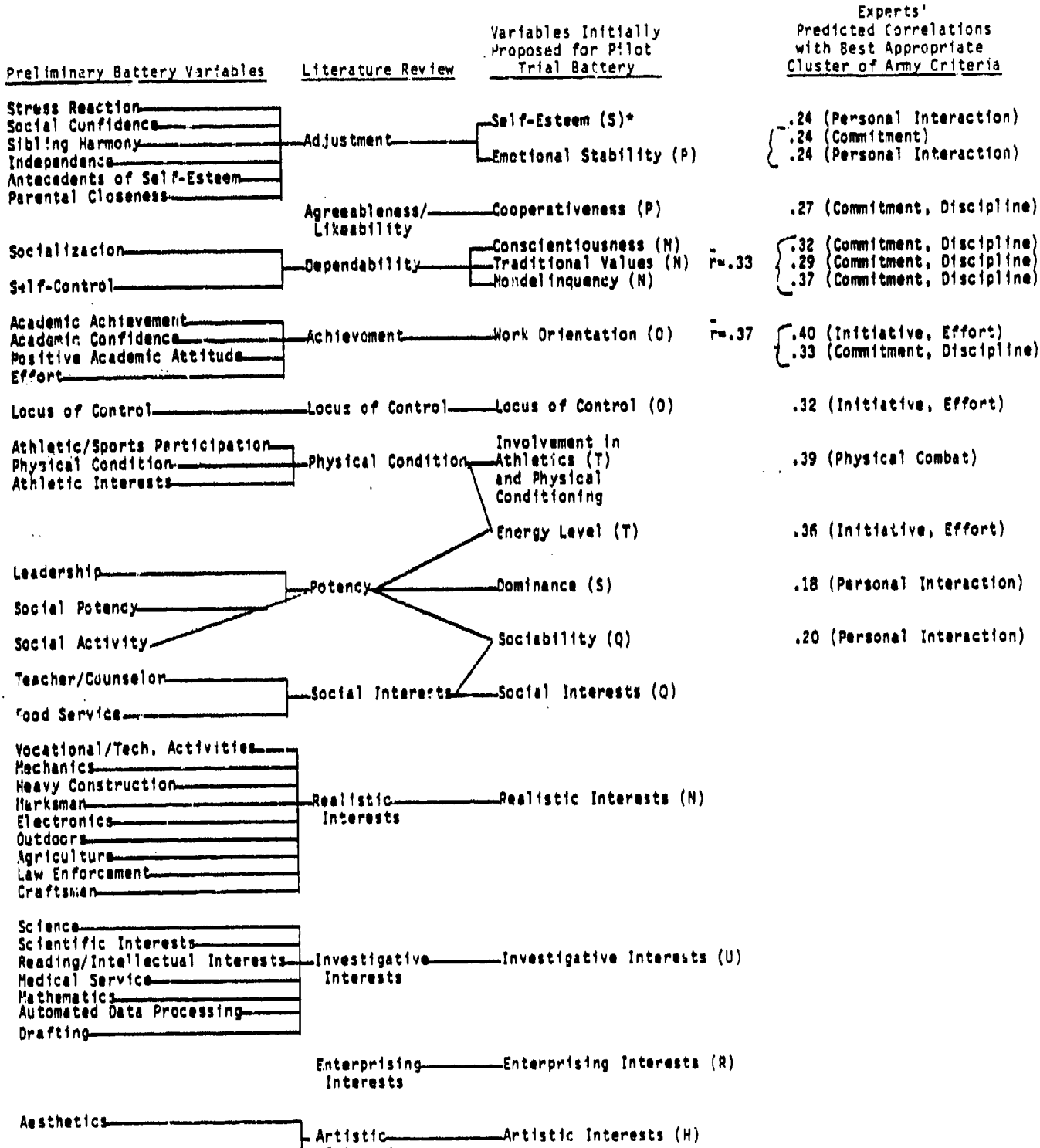
Criterion	Median r
Training	.24
Job Proficiency	.32
Job Involvement	.29
Unfavorable Military Dischange	.27
Substance Abuse	.26
Delinquency	.20

Table II.23

Summary of Criterion-Related Validities of Temperament Constructs[a]

Temperament Construct	Median r's						
	Educational Criteria	Training Criteria	Job Proficiency Criteria	Job Involvement/ Withdrawal Criteria	Unfavorable Military Discharge Criteria	Delinquency Criteria	Substance Abuse Criteria
Potency	.06	.13	.07	.04	--	-.26	.09
Adjustment	.14	.19	.11	.17	-.43	-.42	-.14
Agreeableness/Likeability	.03	.08	.03	-.02	--	-.31	-.03
Dependability	.13	.12	.11	-.09	--	-.43	-.42
Intellectance	.17	.19	.01	.14	--	-.24	.20
Affiliation	-.03	--	-.02	.09	--	--	-.07
Achievement	.30	.33	.24	--	--	-.35	.26
Masculinity	-.16	.09	.10	.03	--	-.02	-.18
Locus of Control	.32	.29	.25	--	--	--	--

[a] From Hough, 1985.



Note: From Hough (1984).

Letters in parentheses indicate predictor cluster.

Figure II.17. Linkages between literature review, expert judgments, and Preliminary and Trial Battery on Non-Cognitive Measures.

Description of ABLE Constructs/Scales

Following the identification of the construct array, item writing groups were created and items were written, revised, edited, and arranged into specific temperament and biographical scales that were intended to be valid measures of the chosen constructs. After this initial phase of item writing, revision, and scale creation, 11 substantive scales and four response bias scales were produced. Table II.24 lists the seven constructs initially chosen for measurement via the ABLE, the 11 scales subsequently developed to represent them, and four validity scales developed by Project A.

Table II.24

Temperament/Biodata Scales (by Construct) Developed for Pilot Trial Battery: ABLE - Assessment of Background and Life Experiences

Construct	Scale
Adjustment	Emotional Stability
Dependability	Nondelinquency Traditional Values Conscientiousness
Achievement	Work Orientation Self-Esteem
Physical Condition	Physical Condition
Leadership (Potency)	Dominance Energy Level
Locus of Control	Internal Control
Agreeableness/Likeability	Cooperativeness
Response Validity Scales	Non-Random Response Unlikely Virtues (Social Desirability) Poor Impression Self-Knowledge

We now discuss, in turn, each construct and the scales developed to measure that construct. The description of the number of items on each scale refers to the Fort Campbell pilot test version.

Adjustment

Adjustment is defined as the amount of emotional stability and stress tolerance that one possesses. The well-adjusted person is generally calm, displays an even mood, and is not overly distraught by stressful situations. He or she thinks clearly and maintains composure and rationality in situations of actual or perceived stress. The poorly adjusted person is nervous, moody, and easily irritated, tends to worry a lot, and "goes to pieces" in times of stress.

The scale included under the Adjustment construct is called Emotional Stability. Emotional Stability is a 31-item scale that contains items such as "Have you ever felt sick to your stomach when you thought about something you had to do?" and "Do you handle pressure better than most other people?" The scale is designed to assess a person's characteristic affect and ability to react to stress.

Dependability

The Dependability construct refers to a person's characteristic degree of conscientiousness. The dependable person is disciplined, well-organized, planful, respectful of laws and regulations, honest, trustworthy, wholesome, and accepting of authority. Such a person prefers order and thinks before acting. The less dependable person is unreliable, acts on the spur of the moment, and is rebellious and contemptuous of laws and regulations. Three ARLE scales fall under the Dependability construct--Nondelinquency, Traditional Values, and Conscientiousness.

Nondelinquency is a 24-item scale that assesses how often a person has violated rules, laws, or social norms. It includes items such as "How often have you gotten into fights?", "Before joining the Army, how hard did you think learning to take orders would be?", and "How many times were you suspended or expelled from high school?"

Traditional Values, a 19-item scale, contains items such as "Are you more strict about right and wrong than most people your age?" and "People should have greater respect for authority. Do you agree?" These items assess how conventional or strict a person's value system is, and how much flexibility he/she has in this value system.

Conscientiousness, the third Dependability scale, includes 24 items. This scale assesses the respondent's degree of dependability, as well as the tendency to be organized and planful. Items include: "How often do you keep the promises you make?", "How often do you act on the spur of the moment?", and "Are you more neat and orderly than most people?"

Achievement

The Achievement construct is defined as the tendency to strive for competence in one's work. The achievement/work-oriented person works hard, sets high standards, tries to do a good job, endorses the work ethic, and concentrates or and persists in completion of the task at hand. This person

is also confident, feels success from past undertakings, and expects to succeed in the future. The person who is less achievement-oriented has little ego involvement in his or her work, feels incapable and self-doubting, does not expend undue effort, and does not feel that hard work is desirable.

Two scales fall under the Achievement construct, including a 31-item scale entitled Work Orientation. This scale addresses how long, hard, and well the respondent typically works and also how he or she feels about work. Among the scale items are these: "How hard were you willing to work for good grades in high school?" and "How important is your work to you?"

The other scale pertaining to Achievement is called Self-Esteem, a 16-item scale that measures how much a person believes in him/herself and how successful he or she expects to be in life. Items from this scale include: "Do you believe you have a lot to offer the Army?" and "Has your life so far been pretty much a failure?"

Physical Condition

The optimal way to establish physical condition is, of course, to administer physical conditioning tests. However, since such a program was not a part of the trial battery, it was decided to ask self-report questions pertaining to perceived physical fitness levels.

The Physical Condition construct refers to one's frequency and degree of participation in sports, exercise, and physical activity.

The scale developed to tap this construct includes 14 items that measure how vigorously, regularly, and well the respondent engages in physical activity. Sample items are "Prior to joining the Army, how did your physical activity (work and recreation) compare to most people your age?" and "Before joining the Army, how would you have rated your performance in physical activities?"

Leadership (Potency)

This construct is defined as the degree of impact, influence, and energy that one displays. The person high on this characteristic is appropriately forceful and persuasive, is optimistic and vital, and has the energy to get things done. The person low on this characteristic is timid about offering opinions or providing direction and is likely to be lethargic and pessimistic.

Two ABLE scales, Dominance and Energy Level, are associated with the leadership construct. Dominance is a 17-item scale that includes such items as "How confident are you when you tell others what to do?" and "How often do people turn to you when decisions have to be made?" The scale assesses the respondent's tendency to take charge or to assume a central and public role.

The Energy Level scale is designed to measure to what degree one is energetic, alert, and enthusiastic. This scale includes 27 items, such as these: "Do you get tired pretty easily?", "At what speed do you like to work?", and "Do you enjoy just about everything you do?"

Locus of Control

Locus of Control refers to one's characteristic belief in the amount of control he or she has or 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. Persons with an external locus of control believe that what happens is beyond their personal control.

The ABLE Internal Control scale is a 21-item scale that assesses both internal and external control, primarily as they pertain to reaching success on the job and in life. The following are example items: "Getting a raise or a promotion is usually a matter of luck. Do you agree?" and "Do you believe you can get most of the things you want if you work hard enough for them?"

Agreeableness/Likeability

The Agreeableness/Likeability construct is defined as the degree of pleasantness versus unpleasantness exhibited in interpersonal relations. The agreeable and likeable person is pleasant, tolerant, tactful, helpful, not defensive, and generally easy to get along with. His or her participation in a group adds cohesiveness rather than friction. The relatively disagreeable and unlikeable person is critical, fault-finding, touchy, defensive, alienated, and generally contrary.

The ABLE Cooperativeness scale is composed of 28 items intended to assess how easy it is to get along with the person making the responses. Items include: "How often do you lose your temper?", "Would most people describe you as pleasant?", and "How well do you accept criticism?"

Validity Scales

The primary purpose of these scales is to determine the validity of responses, that is, the degree to which the responses are accurate depictions of the person completing the inventory. Four validity scales are included: Non-Random Response, Unlikely Virtues, Poor Impression, and Self-Knowledge.

Non-Random Response. The response options for this scale are composed of one right answer, scored as one, and two response options that are both wrong and are both scored zero. The content asks about information that any person is virtually certain to know. Two of the eight items from the Non-Random Response scale are:

"The branch of the military that deals most with airplanes is the:

1. Military Police
2. Coast Guard
3. Air Force"

"Groups of soldiers are called:

1. Tribes

2. Troops

3. Weapons"

The intent of this scale is to detect those respondents who cannot or are not reading the questions, and are instead randomly filling in the circles on the answer sheet.

Unlikely Virtues. This scale is aimed at detecting those who respond in a socially desirable manner (i.e., "fake good"). There are 12 items, such as: "Do you sometimes wish you had more money?" or "Have you always helped people without even the slightest bit of hesitation?"

Poor Impression. This scale does not reflect psychopathology but rather an attempt to simulate psychopathology. Persons who attempt to "fake bad" receive the most deviant scores, while psychiatric patients score average or slightly higher than average. Thus, this scale is designed to detect those respondents who wish to make themselves appear emotionally unstable when in fact they are not.

The Poor Impression scale has 23 items, most of which are also scored on another substantive ABLE scale. Items include "How much resentment do you feel when you don't get your way?", "Did your high school classmates consider you easy to get along with?", and "How often do you keep the prom- ises that you make?" The response option scored as 1 is the option that indicates the least social desirability.

Self-Knowledge. This 13-item scale is intended to identify people who are more self-aware, more insightful, and more likely to have accurate perceptions about themselves. The responses of persons high on this scale may have more validity for predicting job criteria. The following are items from the Self-Knowledge scale: "Do other people know you better than you know yourself?" and "How often do you think about who you are?"

ABLE Revisions Based on Pilot Test Results

The non-cognitive inventories were pilot tested at two of the three pilot test sites. Revision of the ABLE took place in three steps. The first was editorial revision prior to pilot testing, the second was based on Fort Campbell results, and the third was based on Fort Lewis findings. The editorial changes prior to pilot testing were made by the research staff acting on suggestions from both sponsor and contractor reviews of the instruments.

The changes resulting from the first editorial review consisted of the deletion of 17 items and the revision of 158 items. The revisions largely consisted of minor changes in wording, resulting in more consistency across items in format, phrasing, and response options.

Fort Campbell Pilot Test

When the inventory was administered at Fort Campbell, the respondents raised very few criticisms or concerns about the ABLE. Several subjects did note the redundancy of the items on the Physical Condition scale, and this 14-item scale was shortened to 9 items.

Item analyses were based on the data from 52 of the 56 Fort Campbell subjects. The four excluded were two who had more than 10% missing data and two who answered fewer than seven of the eight Non-Random Response scale items "correctly." The two statistics that were examined for each ABLE item were its correlation with the total scale on which it is scored and the endorsement frequencies for all of its response options. Items that failed to correlate at least .15 in the appropriate direction with their respective scales were considered potentially weak. Items (other than validity scale items) for which one or more of the response options was endorsed by fewer than two subjects (i.e., $< 4\%$ of the sample) were also identified. Six items fell into the former category and 68 items fell into the latter, and an additional 7 items fell into both. Many of these 81 items had already been revised or deleted during the editorial process. However, all of them were examined for revision and deletion, as appropriate. As a result, 15 items were revised for the first time, 18 items were further revised, and 6 additional items were deleted.

In summary, a total of 23 items were deleted and 173 items revised on the basis of the editorial review and Fort Campbell findings. Those items that were deleted were those that did not "fit well" either conceptually or statistically, or both, with the other items in the scale and with the construct in question. If the item appeared to have a "good fit" but was not clear or did not elicit sufficient variance, it was revised rather than deleted. The ABLE, which had begun at 291 items, was now a revised 268-item inventory to be administered at Fort Lewis.

Fort Lewis Pilot Test

The ABLE was completed by 118 soldiers during the pilot testing at Fort Lewis. One of the 118 inventories was deleted from analysis because data were missing for more than 10% of the items, and another 11 inventories were deleted because fewer than seven of the eight Non-Random Response scale items were answered "correctly." Thus, the remaining sample size was 106.

Item response frequency distributions were examined to detect items with relatively little discriminatory power. There were only three items where two of the three response choices were endorsed by less than 10% of the sample (not including validity scale items). After examining the content of these three items, it was decided to leave two of them intact, and delete one. Twenty items were revised because one of the three response choices was endorsed by less than 10% of the sample.

Overall, the inventory appeared to be functioning well and only minor revisions were required. On the following pages, the psychometric data obtained during the two pilot tests are presented.

Scale Statistics and Intercorrelations

Table II.25 presents means, standard deviations, mean item-total correlations, and Hoyt internal consistency reliabilities for each ABLE scale in each of the two pilot samples. In Table II.26 the scale intercorrelations are shown, except for the Non-Random Response and Poor Impression validity scales. It is interesting to note the low correlations between the Unlikely Virtues scale, which is an indicator of social desirability, and the other scales. This finding, although based on small samples, suggests that soldiers were not responding only in a socially desirable fashion, but were instead responding honestly.

Table II.25

ABLE Scale Statistics for Fort Campbell and Fort Lewis Pilot Samples[a]

	No. Items		Mean		SD		Mean Item-Total Correlation		Hoyt Reliability	
	A	B	A	B	A	B	A	B	A	B
ADJUSTMENT										
Emotional Stability	31	30	72.0	69.0	9.1	8.6	.47	.46	.87	.87
DEPENDABILITY										
Nondelinquency	24	25	55.9	59.1	6.3	6.3	.40	.40	.80	.78
Traditional Values	19	16	43.8	32.4	4.8	4.3	.39	.41	.73	.67
Conscientiousness	24	21	58.0	50.2	5.8	5.3	.41	.41	.80	.75
ACHIEVEMENT										
Work Orientation	31	27	74.5	62.9	8.0	7.8	.42	.48	.84	.86
Self-Esteem	16	15	37.4	34.9	5.0	4.7	.54	.52	.84	.80
LEADERSHIP (POTENCY)										
Dominance	17	16	37.7	36.6	5.0	6.1	.53	.57	.78	.86
Energy Level	27	25	61.3	59.3	7.2	7.4	.46	.52	.85	.88
LOCUS OF CONTROL										
Internal Control	21	21	51.0	49.9	6.3	6.3	.46	.46	.84	.80
AGREEABLENESS/LIKEABILITY										
Cooperativeness	28	25	63.8	56.4	7.0	6.7	.39	.43	.82	.81
PHYSICAL CONDITION										
Physical Condition	14	9	43.1	31.3	9.7	7.0	.66	.73	.92	.87
ABLE Validity Scales										
Non-Random Response	8	8	--	7.6	--	.7	--	.43	--	
Unlikely Virtues	12	12	18.0	16.6	3.2	3.5	.38	.48	.37	.71
Self-Knowledge	13	13	31.4	29.8	3.7	4.0	.43	.46	.61	.71

[a] Column A indicates Fort Campbell (N = 52) and Column B Fort Lewis (N = 106).

Table II.26

ABLE Scale Intercorrelations: Fort Campbell Pilot Test

	Emotional Stability	Nondelinquency	Traditional Values	Conscientiousness	Work Orientation	Self-Esteem	Dominance	Energy Level	Internal Control	Cooperativeness	Physical Condition	Unlikely Virtues	Self-Knowledge
Emotional Stability	--	45	51	42	42	61	42	53	47	56	22	06	13
Nondelinquency	45	--	71	67	51	53	25	33	58	52	01	13	31
Traditional Values	51	71	--	58	59	56	33	54	70	56	23	19	24
Conscientiousness	42	67	58	--	79	68	44	61	53	53	20	09	40
Work Orientation	42	51	59	79	--	72	52	77	59	47	18	10	39
Self-Esteem	61	53	56	68	72	--	65	73	62	41	26	10	22
Dominance	42	25	33	44	52	65	--	62	34	08	35	-03	23
Energy Level	53	33	54	61	77	73	62	--	55	38	27	15	21
Internal Control	47	58	70	53	59	62	34	55	--	42	06	-03	27
Cooperativeness	56	52	56	53	47	41	08	38	42	--	11	16	14
Physical Condition	22	01	23	20	18	26	35	27	06	11	--	06	02
Unlikely Virtues	06	13	19	09	10	10	-03	15	-03	16	06	--	-09
Self-Knowledge	13	31	24	40	39	22	23	21	27	14	02	-09	--

NOTE: Decimals have been omitted.

In addition to the ABLE, four well-established measures of temperament had been administered to 46 Fort Campbell soldiers to serve as marker variables. They were the Socialization scale of the California Psychological Inventory, Rotter's Locus of Control scale, and the Stress Reaction scale and Social Potency scale of the Differential Personality Questionnaire. The four scales had also been used earlier as part of the Preliminary Battery temperament inventory, known as the Personal Opinion Inventory (POI).

From a total of 46 soldiers completing the instruments, the responses of 38 were used to compute correlations between ABLE scales and the markers. Results are shown in Table II.27. While these results are based on a small sample, they do indicate that the ABLE scales appear to be measuring the constructs they were intended to measure.

TABLE II.27

Correlations Between ABLE Constructs and Scales and Personal Opinion Inventory (POI) Marker Variables[a]: Fort Campbell Pilot Test

	POI Scale			
ABLE Construct	DPQ Stress Reaction	DPQ Social Potency	Rotter Locus of Control	CPI Socialization
Emotional Stability	**-.70**	.32	.30	.32
Dominance	-.24	**.67**	.18	.22
Internal Control	-.32	.26	**.67**	.60
Nondelinquency	-.34	.10	.32	**.62**

[a] "Marker" correlations are indicated by a box.

Using the Fort Lewis data (N = 106) the correlation matrices for the 10 ABLE substantive scales were factor analyzed, both with and without the social desirability variance. Principal factor analyses were used, with rotation to simple structure by varimax rotation. Both factor matrices appear in Table II.28.

The structure of the temperament and biodata domain, as measured by the ABLE during the pilot tests, could not be specified with certainty due to the relatively small pilot test sample upon which the correlational and factor analyses were run. The larger Concurrent Validation samples will provide more definitive information. The scales do, however, appear to be measuring the same content as the marker variables that were a part of the Personal Opinion Inventory (Preliminary Battery). The internal consistency reliabilities and score distribution of the ABLE scales are more than adequate.

Description of Interest (AVOICE) Constructs/Scales

The seminal work of John Holland (1966) has resulted in widespread acceptance of a six-construct, hexagonal model of interests. Our principal problem in the development and testing of an interests measure was not which constructs to measure, but rather how much emphasis should be devoted to the assessment of each.

The interest inventory used in the Preliminary Battery is called the VOICE (Vocational Interest Career Examination), and was originally developed by the U.S. Air Force. This inventory served as the starting point for the AVOICE (Army Vocational Interest Career Examination). The intent for the AVOICE was to measure all six of Holland's constructs, as well as provide sufficient coverage of the vocational areas most important in the Army. Table II.29 shows the six interest constructs assessed by the AVOICE together with their associated scales. The Basic Interest item, one of which is written for each Holland construct, describes a person with prototypic interests. The respondent indicates how well this description fits him/her.

In addition to the Holland constructs and associated scales, the AVOICE also included six scales dealing with organizational climate and environment and an expressed interests scale. Table II.30 shows these variables and asociated measures.

As used in the pilot testing, the AVOICE included 306 items. Nearly all items were scored on a five-point scale that ranged from "Like Very Much" (scored 5) to "Dislike Very Much" (scored 1). Items in the Expressed Interests scale were scored on a three-point scale in which the response options were different for each item, yet one option always reflected the most interest, one moderate interest, and one the least interest.

Each construct/category and the scales developed for it are now discussed in turn.

TABLE II.28

Varimax Rotated Principal Factor Analyses of 10 ABLE Scales: Fort Lewis Pilot Test

ABLE Scale	Five-Factor Solution With Social Desirability Variance Included				
	I	II	III	IV	V
Dominance	[.66]	.15	.16	.00	.21
Energy Level	.45	.19	.32	.22	[.79]
Self-Esteem	[.80]	.13	.22	.30	.27
Internal Control	.33	[.52]	.15	.44	.29
Traditional Value	.18	[.78]	.29	.22	.10
Nondelinquency	.09	.50	[.56]	.41	.09
Conscientiousness	.40	.34	[.61]	.14	.16
Work Orientation	.57	.25	[.63]	.15	.24
Emotional Stability	.33	.11	.02	.43	[.53]
Cooperativeness	.08	.30	.21	[.77]	.22

	Five-Factor Solution With Social Desirability Variance Partialed Out				
	I	II	III	IV	V
Dominance	[.65]	.15	.23	-.03	.18
Energy Level	.39	.18	[.82]	.13	.36
Self-Esteem	[.79]	.12	.32	.24	.19
Internal Control	.31	[.52]	.34	.40	.14
Traditional Values	.17	[.83]	.10	.17	.18
Nondelinquency	.08	[.56]	.06	.40	.42
Conscientiousness	.40	.37	.11	.11	[.56]
Work Orientation	.57	.27	.22	.13	[.62]
Emotional Stability	.30	.08	[.60]	.35	.06
Cooperativeness	.06	.31	.26	[.78]	.13

Table II.29

Holland Basic Interest Constructs, and Army Vocational Interest Career Examination (AVOICE) Scales Developed for Pilot Trial Battery

Construct	Scale
Realistic	Basic Interest Item
	Mechanics
	Heavy Construction
	Electronics
	Electronic Communication
	Drafting
	Law Enforcement
	Audiographics
	Agriculture
	Outdoors
	Marksman
	Infantry
	Armor/Cannon
	Vehicle Operator
	Adventure
Conventional	Basic Interest Item
	Office Administration
	Supply Administration
	Food Service
Social	Basic Interest Item
	Teaching/Counseling
Investigative	Basic Interest Item
	Medical Services
	Mathematics
	Science/Chemical
	Automated Data Processing
Enterprising	Basic Interest Item
	Leadership
Artistic	Basic Interest Item
	Aesthetics

Table II.30

Additional AVOICE Measures: Organizational Climate/Environment and Expressed Interest Scales

Construct	Scale
Achievement (Org. Climate/Environment)	Achievement Authority Ability Utilization
Safety (Org. Climate/Environment)	Organizational Policies and Procedures Supervision - Human Resources Supervision - Technical
Comfort (Org. Climate/Environment)	Activity Variety Compensation Security Working Conditions
Status (Org. Climate/Environment)	Advancement Recognition Social Status
Altruism (Org. Climate/Environment)	Co-Workers Moral Values Social Services
Autonomy (Org. Climate/Environment)	Responsibility Creativity Independence
Expressed Interests	Expressed Interests

Realistic Interests

This construct is defined as a 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. Realistic Interests are associated with occupations such as mechanic, engineer, and wildlife conservation officer, and negatively associated with such occupations as social worker and artist.

The Realistic construct is by far the most thoroughly assessed of the six constructs tapped by the AVOICE, reflecting that the preponderance of work in the Army is of a Realistic nature. Fourteen AVOICE scales are included, in addition to the Basic Interest item.

Conventional Interests

Conventional Interests refer to one's degree of preference for well-ordered, systematic, and practical activities and tasks. Persons with conventional interests may be characterized as conforming, not overly imaginative, efficient, and calm. Conventional Interests are associated with occupations such as accountant, clerk, and statistician, and negatively associated with occupations such as artist or author.

In addition to the Basic Interest item, three scales fall under the Conventional Interests construct, Office Administration, Supply Administration, and Food Service. They have, respectively, 16, 13, and 17 items.

Social Interests

Social Interests are defined as the amount of liking one has for social, helping, and teaching activities and tasks. Persons with social interests may be characterized as responsible, idealistic, and humanistic. Social interests are associated with occupations such as social worker, high school teacher, and speech therapist, and negatively associated with occupations such as mechanic or carpenter.

Besides the Basic Interest item, only one scale is included in the AVOICE for assessing Social Interests, the Teaching/Counseling scale. This 70-item scale includes items such as "Give on-the-job training," "Organize and lead a study group," and "Listen to people's problems and try to help them."

Investigative Interests

This construct refers to one's preference for scholarly, intellectual, and scientific activities and tasks. Persons with investigative interests enjoy analytical, ambiguous, and independent tasks, but dislike leadership and persuasive activities. Investigative Interests are associated with such occupations as astronomer, biologist, and mathematician, and negatively associated with occupations such as salesman or politician.

Along with the Basic Interest item, Medical Services, Mathematics, Science/Chemical, and Automated Data Processing are the four AVOICE scales that tap Investigative Interests. The scales differ in length, with Medical Services containing 24 items; Mathematics, 5; Science/Chemical, 11; and Automated Data Processing, 7.

Enterprising Interests

The Enterprising construct refers to one's preference for persuasive, assertive, and leadership activities and tasks. Persons with enterprising interests may be characterized as ambitious, dominant, sociable, and self-confident. Enterprising Interests are associated with such occupations as salesperson and business executive, and negatively associated with occupations such as biologist or chemist.

Besides the Basic Interest item, only one AVOICE scale assesses the respondent's Enterprising Interests. This scale, entitled Leadership, contains six items.

Artistic Interests

This final Holland construct is defined as a person's degree of liking for unstructured, expressive, and ambiguous activities and tasks. Persons with artistic interests may be characterized as intuitive, impulsive, creative, and non-conforming. Artistic Interests are associated with such occupations as writer, artist, and composer, and negatively associated with occupations such as accountant or secretary.

In addition to the Basic Interest item, the AVOICE Aesthetics scale is designed to tap Artistic Interests, and includes five items.

Organizational Climate/Environment

Six constructs that pertain to a person's preference for certain types of work environments and conditions are assessed by the AVOICE through 20 scales of two items each. These environmental constructs include Achievement, Safety, Comfort, Status, Altruism, and Autonomy. The items that assess these constructs are distributed throughout the AVOICE, and are responded to in the same manner as the interests items, that is, "Like Very Much" to "Dislike Very Much."

Because the scales contain only two items each and for ease of presentation, Figure II.18 shows the constructs, scales, and an item from each scale.

Expressed Interests

Although not a psychological construct, expressed interests were included in the AVOICE because of the extensive research showing their validity in criterion-related studies. These studies had measured expressed interests simply by asking respondents what occupation or occupational area was of most interest to them. In the AVOICE, such an open-ended question was not feasible, so instead respondents were asked how confident they were that their chosen job in the Army was the right one for them.

Construct/Scale	Example
Achievement	
Achievement	"Do work that gives a feeling of accomplishment."
Authority	"Tell others what to do on the job."
Ability Utilization	"Make full use of your abilities."
Safety	
Organizational Policy	"A job in which the rules are not equal for everyone."
Supervision - Human Resources	"Have a boss that supports the workers."
Supervision - Technical	"Learn the job on your own."
Comfort	
Activity	"Work on a job that keeps a person busy."
Variety	"Do something different most days at work."
Compensation	"Earn less than others do."
Security	"A job with steady employment."
Working Conditions	"Have a pleasant place to work."
Status	
Advancement	"Be able to be promoted quickly."
Recognition	"Receive awards or compliments on the job."
Social Status	"A job that does not stand out from others."
Altruism	
Co-workers	"A job in which other employees were hard to get to know."
Moral Values	"Have a job that would not bother a person's conscience."
Social Services	"Serve others through your work."
Autonomy	
Responsibility	"Have work decisions made by others."
Creativity	"Try out your own ideas on the job."
Independence	"Work alone."

Figure II.18. Organizational climate/environment constructs, scales within constructs, and an item from each scale.

This Expressed Interests scale contained eight items which, as mentioned, had three response options that formed a continuum of confidence in the person's occupational choice. Selected items from this scale include: "Before you went to the recruiter, how certain were you of the job you wanted in the Army?", "If you had the opportunity right now to change your job in the Army, would you?", and "Before enlisting, how long were you interested in a particular Army job?"

AVOICE Revisions and Scale Statistics Based on Pilot Tests

Revisions were made in the AVOICE on the basis of pilot administrations at Fort Campbell and Fort Lewis. Overall, the revisions made were far less substantial for the AVOICE than for the ABLE. Editorial review of the inventory, together with the verbal feedback of Fort Campbell soldiers, resulted in revision of 15 items, primarily minor wording changes. An additional five items were modified because of low item correlations with the total scale score in the Fort Campbell data. No items were deleted based on editorial review, verbal feedback, or item analyses.

In the Fort Lewis pilot test, no revisions or deletions were made to the AVOICE items. Item response frequencies were examined to detect items that had relatively little discriminatory power; that is, three or more of the five response choices received less than 10% endorsement. There were only two such items, and, upon examination of the item content, it was decided not to revise these. Thus, a total of only 20 AVOICE items were revised based on editorial review and pilot testing.

At Fort Campbell a total of 57 soldiers completed the AVOICE, with 55 providing sufficient data for analysis. For the Fort Lewis data the responses of 4 of 118 soldiers were eliminated for exceeding the missing data criterion (10%), resulting in an analysis sample size of 114. Scale statistics for this larger sample are shown in Table II.31. Reliabilities are again excellent.

AVOICE scale means and standard deviations were also calculated separately for males and females and for blacks and whites, but the sample sizes are very small and these data are best viewed as exploratory only. As would be expected on the basis of previous research, there are differences between the sexes in mean score on certain interest scales (Table II.32). Scales such as Mechanics and Heavy Construction show higher scores for males than females. On the majority of the scales, however, the differences are less pronounced. Differences between blacks and whites are quite small and those tables are not shown.

Table II.31

AVOICE Scale Statistics for Total Group: Fort Lewis Pilot Test (N = 114)

AVOICE Scale	No. Items	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
REALISTIC					
Basic Interest Item	1	3.09	1.17	--	--
Mechanics	16	53.02	13.13	.73	.94
Heavy Construction	23	72.57	15.64	.62	.92
Electronics	20	63.94	16.86	.75	.96
Electronic Communication	7	21.44	5.73	.73	.85
Drafting	7	22.62	6.11	.76	.87
Law Enforcement	16	50.82	11.33	.63	.89
Audiographics	7	24.30	5.12	.69	.81
Agriculture	5	15.24	3.62	.61	.58
Outdoors	9	33.09	6.25	.62	.80
Marksman	5	16.57	4.48	.79	.84
Infantry	10	31.04	7.26	.64	.84
Armor/Cannon	8	23.46	6.15	.67	.83
Vehicle Operator	10	30.45	7.10	.65	.84
Adventure	8	18.84	3.60	.57	.72
CONVENTIONAL					
Basic Interest Item	1	3.00	.92	--	--
Office Administration	16	45.39	12.61	.72	.94
Supply Administration	13	36.97	9.65	.71	.92
Food Service	17	43.46	10.53	.59	.89
SOCIAL					
Basic Interest Item	1	3.25	1.03	--	--
Teaching/Counseling	7	23.61	5.20	.71	.83
INVESTIGATIVE					
Basic Interest Item	1	3.09	.95	--	--
Medical Services	24	71.32	16.65	.66	.94
Mathematics	5	15.82	4.20	.75	.80
Science/Chemical	11	30.29	8.41	.68	.88
Automated Data Processing	7	24.29	5.78	.74	.86
ENTERPRISING					
Basic Interest Item	1	3.11	1.13	--	--
Leadership	6	20.71	4.41	.72	.81

(Continued)

Table II.31 (Continued)

AVOICE Scale Statistics for Total Group: Fort Lewis Pilot Test (N = 114)

AVOICE Scale	No. Items	Mean	SD	Mean Item-Total Correlation	Hoyt Reliability
ARTISTIC					
Basic Interest Item	1	2.99	1.27	--	--
Aesthetics	5	14.73	4.12	.74	.79
ORGANIZATIONAL CLIMATE/ ENVIRONMENT DIMENSIONS					
Achievement	6	21.09	2.95	--	--
Safety	6	21.64	3.20	--	--
Comfort	10	38.50	3.83	--	--
Status	6	21.37	2.97	--	--
Altruism	6	21.67	3.28	--	--
Autonomy	6	20.46	2.33	--	--
EXPRESSED INTEREST	8	15.71	3.19	.59	.66

Table II.32

AVOICE Means and Standard Deviations Separately for Males and Females: Fort Lewis Pilot Test

AVOICE Scale	Males (N = 87) Mean	Males (N = 87) SD	Females (N = 19) Mean	Females (N = 19) SD
REALISTIC				
Basic Interest Item	3.24	1.13	2.35	1.11
Mechanics	54.93	12.51	44.05	12.28
Heavy Construction	75.31	13.24	59.70	19.22
Electronics	66.38	15.95	52.45	16.23
Electronic Communication	21.48	5.73	21.25	5.72
Drafting	22.97	6.11	21.00	5.83
Law Enforcement	51.72	11.41	46.60	9.95
Audiographics	24.27	5.03	24.45	5.52
Agriculture	15.46	3.59	14.20	3.57
Outdoors	33.94	5.75	29.10	6.92
Marksman	17.35	4.05	12.90	4.56
Infantry	31.94	7.14	26.85	6.28
Armor/Cannon	24.21	5.99	19.95	5.71
Vehicle Operator	31.05	6.52	27.60	8.81
Adventure	19.39	3.28	16.32	3.91
CONVENTIONAL				
Basic Interest Item	2.97	.92	3.15	.91
Office Administration	44.91	11.93	47.60	15.19
Supply Administration	36.95	9.56	37.10	10.09
Food Service	42.54	9.89	47.80	12.23
SOCIAL				
Basic Interest Item	3.24	1.05	3.30	.95
Teaching/Counseling	23.15	5.13	25.75	4.97
INVESTIGATIVE				
Basic Interest Item	3.10	.95	3.05	.97
Medical Services	71.10	16.65	72.40	16.59
Mathematics	15.59	4.31	16.95	3.40
Science/Chemical	30.99	8.69	27.00	5.96
Automated Data Processing	24.20	5.97	24.70	4.76

(Continued)

Table II.32 (Continued)

AVOICE means and Standard Deviations Separately for Males and Females: Fort Lewis Pilot Test

AVOICE Scale	Males (N = 87) Mean	Males (N = 87) SD	Females (N = 19) Mean	Females (N = 19) SD
ENTERPRISING				
Basic Interest Item	3.14	1.14	2.95	1.02
Leadership	20.53	4.61	21.55	3.17
ARTISTIC				
Basic Interest Item	2.96	1.25	3.15	1.31
Aesthetics	14.29	4.22	16.80	2.77
ORGANIZATIONAL CLIMATE/ ENVIRONMENT DIMENSIONS				
Achievement	20.97	2.92	21.65	3.02
Safety	21.59	3.36	21.90	2.23
Comfort	38.26	3.76	39.65	3.97
Status	21.22	3.00	22.05	2.73
Altruism	21.48	3.26	22.55	3.26
Autonomy	20.45	2.22	20.55	2.78
EXPRESSED INTEREST	15.79	3.34	15.35	2.29

Summary of Pilot Test Results for Non-Cognitive Measures

The two non-cognitive inventories of the Pilot Trial Battery, the ABLE and the AVOICE, are designed to measure a total of 20 constructs plus a validity scale category. The ABLE assesses six temperament constructs and the Physical Condition construct through 11 scales, and also includes 4 validity scales. The AVOICE measures six Holland interests constructs, six Organizational Environment constructs, and Expressed Interests through 31 scales. Altogether, the 46 scales of the inventories include approximately 600 items, 291 ABLE items and 306 AVOICE items for the Fort Campbell version, and 268 ABLE items and 306 AVOICE items for the Fort Lewis version.

Evaluation and revision of the inventories took place in three steps. First, each was subjected to editorial review by project staff prior to any pilot testing. This review resulted in nearly 200 wording changes and the deletion of 17 items. The majority of these changes applied to the ABLE.

The second stage of evaluation took place after the Fort Campbell pilot testing. Feedback from the soldiers taking the inventory and data analyst of the results (e.g., item-total correlations, item response distributions) were used to refine the inventories. Twenty-three ABLE items were deleted and 173 ABLE items were revised; no AVOICE items were deleted and 20 AVOICE items were revised.

In the third stage of evaluation, after the Fort Lewis pilot testing, far fewer changes were made. One ABLE item was deleted, 20 ABLE items were revised, and no changes were made to the AVOICE. Throughout the evaluation process, it is likely that the AVOICE was less subject to revision because it uses a common response format for all items, whereas the response options for ABLE items differ by item.

The psychometric data obtained with both inventories seemed highly satisfactory; the scales were shown to be reliable and appeared to be measuring the constructs intended. Sample sizes in these administrations were fairly small (Fort Campbell N = 52 and 55 for ABLE and AVOICE, respectively; Fort Lewis N = 106 and 114, ABLE and AVOICE, respectively), but results were similar in each sample.

Field testing of the non-cognitive measures is described in Section 6.

Section 6

FIELD TESTS OF THE PILOT TRIAL BATTERY

The previous sections have described the development of the Pilot Trial Battery, the results of three pilot tests at Forts Carson, Campbell, and Lewis, and the revisions made on the basis of the pilot data.

The final step before the Concurrent Validation was a more systematic series of field tests of all the predictor measures using larger samples. Described in this section are the field test samples and procedures, and a variety of analyses of the field test data. The predictor revisions made on the basis of these analyses are described in Section 7. The outcome of the field test/revision process was the final form of the predictor battery (i.e., the Trial Battery) to be used in the Concurrent Validation.

Field tests were conducted at three sites. The sites and basic purposes of the field test at each site are described below.

Fort Knox - The full Pilot Trial Battery was administered here to evaluate the psychometric characteristics of all the measures and to analyze the covariance of the measures with each other and with the ASVAB. In addition, the measures were re-administered to part of the sample to provide data for estimating the test-retest reliability of the measures. Finally, part of the sample received practice on some of the computer measures and were then retested to obtain an estimate of the effects of practice on scores on computer measures.

Fort Bragg - The non-cognitive Pilot Trial Battery measures, Assessment of Background and Life Experiences (ABLE) and Army Vocational Interest Career Examination (AVOICE), were administered to soldiers at Fort Bragg under several experimental conditions to estimate the extent to which scores on these instruments could be altered or "faked" when persons are instructed to do so.

Minneapolis Military Entrance Processing Station (MEPS) - The non-cognitive measures were administered to a sample of recruits as they were being processed into the Army, to obtain an estimate of how persons in an applicant setting might alter their scores.

Cognitive Paper-and-Pencil and Computer-Administered Field Tests

In this subsection are described the field tests of the cognitive paper-and-pencil tests and the computer-administered tests. These data were collected at Fort Knox. No data from the Fort Bragg and Minneapolis MEPS studies were used in these analyses.

Sample and Procedure

Data collection was scheduled for 4 weeks at Fort Knox. During the first 2 weeks, 24 soldiers were scheduled each day, for a projected total sample size of 240. These soldiers were administered the entire Pilot Trial Battery. Each group assembled at 0800 hours. The testing sessions included two 15-minute breaks, and 1 hour was allowed for lunch.

Each soldier from the first 2 weeks' sample reported back for a half day of additional testing, either in the morning or the afternoon, exactly 2 weeks after her or his first session. Each individual then completed one-third of all the paper-and-pencil tests (a retest), and completed either the computer "practice" session or the entire computer battery (a retest).

In the experiment on practice effects, practice was given on five tests: Reaction Time 2 (Choice Reaction Time), Target Tracking 1, Cannon Shoot, Target Tracking 2, and Target Shoot. These tests were selected because they were thought to be the tests that would show greatest improvement with practice (all the psychomotor tests were included). There were three practice trials. For the first two practice trials, unique items (i.e., items not appearing on the full battery test) were used for Target Tracking 1, Target Tracking 2, and Cannon Shoot. For the third trial, the original item content was used.

Due to the usual exigencies of data collection in the field, on some days fewer than 24 soldiers appeared, and on other days more than 24 soldiers appeared. Consequently, the actual sample sizes were as follows:

N = 293 completed all cognitive and non-cognitive paper-and-pencil tests

N = 256 completed computer tests

N = 112-129 completed retest of paper-and-pencil tests (N varied across tests)

N = 113 completed retest of computer tests

N = 74 completed practice effects on computer tests

Table II.33 shows the race and gender makeup for Fort Knox soldiers completing at least part of the Pilot Trial Battery. The mean age of the participating soldiers was 21.9 years (SD = 3.1). The mean years in service, computed from date of entry into the Army, was 1.6 years (SD = 0.9).

Table II.33

Race and Gender of the Fort Knox Field Test Sample for the Pilot Trial Battery

Race	Frequency
White	156
Hispanic	24
Black	121
Native American	2
Total	303

Sex	Frequency
Female	57
Male	246

Descriptive Statistics

Table II.34 shows the means, standard deviations, and reliabilities of the cognitive paper-and-pencil tests in the Pilot Trial Battery. The means and standard deviations indicate that the tests are at about the desired level of difficulty, with the possible exception of Orientation Test 3 which appeared somewhat more difficult than desirable. Internal consistency reliability estimates were relatively high, with the exception of Reasoning Test 2 (.63).

The Fort Lewis split-half coefficients, which are based on separately timed halves, provide an appropriate estimate of internal consistency for speeded tests. The interval for test-retest was 2 weeks. Reasoning was again the least reliable. Table II.35 shows gain scores that were higher than initially expected on Object Rotation, Shapes, Path, and Orientation 1 tests. Inspection of the last two columns in Table II.35 indicated that much of the gain probably occurred because the soldiers attempted more items the second time they took the test. This is certainly to be expected since the subjects would be more familiar with types and instructions.

Table II.34

Means, Standard Deviations, and Reliability Estimates for the Fort Knox Field Test of the Ten Paper-and-Pencil Cognitive Tests

Test	No. of Items	Time Allotted (in minutes)	Score Mean[a]	SD[a]	Reliability Coefficients[b]		
					Split Half (N = 118)	Coefficient Alpha (N = 290)	Test-Retest (N = 97 to 126)
Assembling Objects	40	16	26.5	8.7	.79	.92	.74
Object Rotation	90	7.5	59.6	19.0	.86	.97	.75
Path	44	8	26.4	10.2	.82	.92	.64
Maze	24	5.5	17.8	4.5	.78	.89	.71
Shapes	54	16	26.4	8.9	.82	.92	.70
Orientation 1	150	10	19.6	5.2	.92	.98	.67
Orientation 2	24	10	21.6	3.6	.89	.88	.80
Orientation 3	20	12	88.7	34.8	.88	.90	.84
Reasoning 1	30	12	11.5	6.0	.78	.83	.64
Reasoning 2	32	10	7.7	5.7	.63	.65	.57

[a] N_S range from 292 to 298 for mean and SD calculations.

[b] The split-half coefficient is computed on pilot test data from Fort Lewis, where two separately timed halves were given, and is corrected to full test length. Coefficient alphas are based on the Fort Knox data and are overestimates for the speeded tests. The test-retest interval was two weeks.

Table II.35

Gains on Pilot Test Battery for Persons Taking Tests at Both Time 1 and Time 2

Test Name	No. of Items	No. of Subjects	Time 1 Mean	Time 1 SD	Time 2 Mean	Time 2 SD	Gain[a]	Items Attempted by 75% of Subjects Time 1	Items Attempted by 75% of Subjects Time 2
Assembling Objects	40	113	25.7	9.1	28.2	8.8	0.28	32	40
Object Rotation	90	125	61.2	19.6	71.3	15.9	0.57	55	69
Shapes	54	121	27.3	10.7	34.4	11.5	0.64	30	42
Maze	24	97	17.5	4.3	18.5	4.3	0.24	17	19
Path	44	126	27.4	8.4	32.5	7.8	0.62	28	36
Reasoning 1	30	117	20.4	5.0	21.2	5.5	0.15	30	30
Reasoning 2	32	121	21.2	3.8	21.9	3.5	0.17	32	32
Orientation 1	150	123	91.9	33.0	112.5	32.0	0.63	85	110
Orientation 2	24	116	11.6	6.0	12.3	6.1	0.11	24	24
Orientation 3	20	117	7.7	5.6	8.1	5.6	0.08	16	19

[a] $\text{Gain} = \dfrac{M_2 - M_1}{\sqrt{\dfrac{SD_1^2 + SD_2^2}{2}}}$

Table II.36 presents the descriptive statistics for 19 scores derived from the computer-administered measures. In general, the cognitive/perceptual tests (except for Cannon Shoot) yield two types of scores: accuracy and speed. In addition, two derived measures can be computed: the slope and intercept obtained when reaction times are regressed against a relevant parameter of test items. For Perceptual Speed and Accuracy, such a parameter was the number of stimuli being compared in an item (i.e., two, five, or nine objects). Recall that the slope represents the average increase in reaction time with an increase of one object in the stimulus set; the lower the value, the faster the comparison. The intercept represents all other processes not involved in comparing stimuli, such as encoding the stimuli and executing the response.

Reaction times on all tests were computed only for correct responses, and the development strategy was to construct items so that every item could be answered correctly, given enough time. Consequently, the speed measures (reaction time) were expected to have more variance and be more meaningful than the accuracy measures.

Analyses of Fort Knox data indicated that total reaction time and decision time were very highly correlated and, since movement time is conceptually uninteresting, we elected to use total reaction time for all tests. There are a number of ways to score reaction time, and for the various alternatives the score distributions, intercorrelations, and reliabilities were examined. There were no striking differences between methods, and untrimmed means were used for all tests except Simple and Choice Reaction Times; because of fewer items, extreme scores could affect the mean much more for these two tests than for the others. To deal with the problem of missing data, cases with more than 10% missing were eliminated.

Procedures for scoring the Cannon Shoot Test differed from those used to score the other cognitive/perceptual tests, and a reaction time score is inappropriate because the task requires the subject to ascertain the optimal time to fire to ensure a direct hit on the target. Therefore, responses were scored by computing a deviation score composed of the difference between the time the subject fired and the optimal time to fire. These scores are summed across all items and the mean deviation time is computed.

For two of the three psychomotor tests, Target Tracking 1 and 2, the distance from the center of the crosshairs to the center of the target was computed approximately 16 times per second, or almost 350 times per trial. These distances were then averaged by the computer to generate the mean distance for each trial. However, the frequency distribution of these scores proved to be highly positively skewed and they were transformed using the natural logarithm transformation. The overall test score for each subject was then the mean of the log mean distance across trials.

Scoring of the Target Shoot Test was a bit more complicated. Three overall test scores were generated for each subject: (a) the percentage of hits; (b) the mean distance from the center of the crosshairs to the center of the target at the time of firing (the distance score); and (c) the mean time elapsed from the start of the trial until firing (the time-to-fire score).

Table II.36

Characteristics of the 19 Dependent Measures for Computer-Administered Tests: Fort Knox Field Tests (N = 256)[a]

Dependent Measure	Mean	SD	Reliability: Split-Half (r_{sh})[b]	Reliability: Test-Retest (r_{tt})[b]
PERCEPTUAL				
Simple Reaction Time (SRT)				
Mean Reaction Time (RT)	56.2 hs[c]	18.8 hs	.90	.37
Choice Reaction Time (CRT)				
Mean Reaction Time (RT)	67.4 hs	10.2 hs	.89	.56
Perceptual Speed and Accuracy (PS&A)				
Percent Correct (PC)	88%	8%	.83	.59
Mean Reaction Time (RT)	325.6 hs	70.4 hs	.96	.65
Slope	42.7 hs/ch[d]	15.6 hs/ch	.88	.67
Intercept	68.0 hs	45.0 hs	.74	.55
Target Identification				
Percent Correct (PC)	90%	10%	.84	.19
Mean Reaction Time (RT)	528.7 hs	134.0 hs	.96	.67
Short-Term Memory (STM)				
Percent Correct (PC)	85%	8%	.72	.34
Mean Reaction Time (RT)	129.7 hs	23.8 hs	.94	.78
Slope	7.2 hs/ch	4.5 hs/ch	.52	.47
Intercept	108.1 hs	23.2 hs	.84	.74
Number Memory				
Percent Correct (PC)	83%	13%	.63	.53
Mean Operation Time (RT)	230.7 hs	73.9 hs	.95	.88
Cannon Shoot				
Time Error (TE)	78.6 hs	20.3 hs	.88	.66
PSYCHOMOTOR				
Target Track 1				
Mean Log Distance	3.2	.44	.97	.68
Target Shoot				
Mean Time to Fire (std) (TF)	-.01	.48	.91	.48
Mean Log Distance (std)	-.01	.41	.86	.58
Target Track 2				
Mean Log Distance	3.91	.49	.97	.68

[a] N varies slightly from test to test.
[b] N = 120 for test-retest reliabilities, but varies slightly from test to test. r_{sh} = split-half reliability; odd-even item correlation with Spearman-Brown correction. r_{tt} = test-retest reliability, 2-week interval between administrations.
[c] hs = hundredths of a second.
[d] hs/ch = hundredths of a second per character.

Percentage of hits was less desirable as a measure because it contains relatively little information compared to the distance measure. Complications arose because subjects received no distance or time-to-fire scores on trials where they failed to fire at the target before the time limit for the trial elapsed. Consequently, the distance and time-to-fire scores for each trial were standardized and the overall distance and time score was then computed by averaging these standardized scores across all trials in which the subject fired at the target.

In Table II.36 the split-half reliabilities are odd-even correlations corrected to full test length, but note that they do not suffer from the artificial inflation that speeded paper-and-pencil measures do. This is because all items are completed by every subject.

The test-retest reliabilities are lower than the split-half reliabilities and three of them are very low. However, two of them, the percent-correct scores, are not the primary score for their respective tests, and Simple Reaction Time is viewed largely as a "warm-up" test.

Special Analyses on Computer-Administered Tests

Correlations With Video Game-Playing Experience. Field test subjects were asked the question, "In the last couple years, how much have you played video games?" There were five possible alternatives, ranging from "You have never played video games" to "You have played video games almost every day." The five alternatives were given scores of 1 to 5, respectively. The mean was 2.99, SD was 1.03 (N = 256), and the test-retest reliability was .71 (N = 113).

The 19 correlations of this item with the computer test scores ranged from -.01 to +.27, with a mean of .10. A correlation of .12 is significant at alpha = .05. We interpret these findings as showing a small, but significant, relationship of video game-playing experience to the more "game-like" tests in the battery.

Effects of "Machine" or Computer Testing Station Differences. We repeated the investigation done at the pilot test at Fort Lewis of the effect of machine or computer testing station differences on computer test scores. There were six computer testing stations, and approximately 40 male soldiers had been tested at each station. (We used only males in this analysis to avoid confounding the results with possible sex differences.) We ran a multivariate analysis of variance (MANOVA) for the 19 computer test scores, with six "machine" levels. Machine differences again had no effect on test scores.

Practice Effects on Selected Computer Test Scores. Table II.37 shows the results of the analyses of variance for the five tests included in the practice effects research. These results show only one statistically significant practice effect, the Mean Log Distance score on Target Tracking Test 2. There were three statistically significant findings for time, indicating that scores did change with a second testing, whether or not practice trials intervened between the two tests. Finally, note that the Omega squared value indicates that relatively small amounts of test score variance are accounted for by the Group, Time, or Time by Group factors.

Table II.37

Effects of Practice on Selected Computer Test Scores

Test	Dependent Measure	Source of Variance	df	F	Omega Squared
Choice Reaction Time	Trimmed Mean Reaction Time	Group	1,180	9.71*	.032
		Time	1,180	25.70*	.035
		Time x Group	1,180	.73	--
Target Tracking 1	Mean Log Distance	Group	1,178	.73	--
		Time	1,178	9.26*	.005
		Time x Group	1,178	4.11	--
Target Tracking 2	Mean Log Distance	Group	1,178	.47	--
		Time	1,178	1.30	--
		Time x Group	1,178	7.79*	.005
Cannon Shoot	Time Error	Group	1,171	3.79	--
		Time	1,171	.16	--
		Time x Group	1,171	5.72	--
Target Shoot	Mean Log Distance	Group	1,171	.41	--
		Time	1,171	9.28*	.012
		Time x Group	1,171	.08	--

*Denotes significance at $p < .01$.

These data suggest that the practice intervention was not a particularly strong one. It should be noted, though, that on some tests subjects' performance actually deteriorated from Time 1 to Time 2. The average gain score for the two groups across the five dependent measures was only .09 standard deviation. This suggests either that the tasks used in these tests are resistant to practice effects, or that performance on these tasks reaches a maximum level of proficiency after only a few trials. Also, recall that analyses of the Pilot Trial Battery cognitive paper-and-pencil tests (see Table II.35) showed gain scores that were as high or higher than those found here. Perhaps gain in scores through retesting or practice is of even less concern for computerized tests than for paper-and-pencil tests.

In summary, data from the practice experiment indicate that scores from computerized psychomotor tests appear to be quite stable over a 2-week period. Practice does have some effect on test scores, but it appears to be relatively small. Certainly it does not seem strong enough to warrant serious concern about the usefulness of the tests.

Intercorrelations of Cognitive Paper-and-Pencil Tests, Computer-Administered Tests, and ASVAB Subtests

Table II.38 contains the intercorrelations for the ASVAB subtests, paper-and-pencil cognitive measure, and computer-administered tests, which include both cognitive/perceptual and psychomotor measures. Note that we have also included scores on the AFQT. These correlations are based on the Fort Knox field test sample, but include only those subjects with test scores available on all variables (N = 168).

In examining these relationships, we first looked at the correlations between tests within the same battery. For example, correlations between ASVAB subtest scores range from .02 to .74 (absolute values). The range of intercorrelations is a bit more restricted when examining the relationships between the cognitive paper-and-pencil test scores (.27 to .67). This range of values reflects the fact that the PTB measures were designed to tap fairly similar cognitive constructs. Intercorrelations for the cognitive/perceptual computer test scores range from .00 to .83 in absolute terms. Note that the highest values appear for correlations between scores computed from the same test. Intercorrelations between psychomotor variables range from .15 to .81 in absolute terms.

Perhaps the most important question to consider is the overlap between the different groups of measures. Do the paper-and-pencil measures and computer tests correlate highly with the ASVAB or are they measuring unique or different abilities? Note that across all PTB paper-and-pencil tests, ASVAB Mechanical Comprehension appears to correlate the highest with the new tests. Across all ASVAB subtests, Orientation Test 3 yields the highest correlations.

Table II.39 summarizes the correlational data in Table II.38. The two tables lead to the conclusion that the various types of measures do not overlap excessively, and therefore do appear to each make separate contributions to ability measurement.

Factor Analysis Results

In addition to examining the intercorrelations among all the cognitive/perceptual measures and psychomotor measures, we also examined results from a factor analysis. Two variables, PS&A reaction time and Short-Term Memory reaction time, were omitted because the reaction time scores from these measures correlated very highly with their corresponding slope or intercept variables. To avoid obtaining communalities greater than one, they were omitted. Results from the seven-factor solution of a principal components factor analysis with varimax rotation are displayed in Table II.40. All loadings of .30 or greater are shown.

Factor 1 includes eight of the ASVAB subtests, six of the paper-and-pencil measures, and two cognitive/perceptual computer variables. Because this factor contains measures of verbal, numerical, and reasoning ability, we have termed this "g."

Table II.38

Intercorrelations Among the ASVAB Subtests and the Pilot Trial Battery Cognitive Paper-and-Pencil and Perceptual/Psychomotor Computer-Administered Tests: Fort Knox Sample (N = 168)

Test[a]	AFQT Score	Gen Science	Arith Reas	Word Knowledge	Parg Comp	Numb Ops	Code Spd	Auto/Shop	Math Knowledge	Mech Comp	Elec Info	Asmbl Obj	Obj Rotat	Shapes	Maze	Path	Reas 1	Reas 2	Orient 1	Orient 2	Orient 3	SRT-RT	CRT-RT	PS&A-PC	PS&A-RT	PS&A-Slp	PS&A-Int	Targ ID-PC	Targ ID-RT	STM-PC	STM-RT	STM-Slp	STM-Int
ASVAB																																
AFQT Score																																
Gen Science	61																															
Arith Reas	87	54																														
Word Know	81	66	61																													
Parg Comp	69	43	53	58																												
Numb Ops	44	02	21	06	14																											
Code Spd	30	-02	14	10	14	55																										
Auto/Shop	45	54	50	45	29	-08	-07																									
Math Know	76	60	74	62	54	19	15	43																								
Mech Comp	55	-50	62	54	36	10	-06	64	57																							
Elec Info	56	66	56	55	39	-03	01	71	59	63																						
PTB-A																																
Asmbl Obj	44	38	48	40	29	-01	19	39	48	57	38																					
Obj Rotat	30	20	33	18	18	18	13	27	29	36	32	47																				
Shapes	35	29	33	28	20	14	23	17	36	35	26	51	50																			
Maze	32	23	38	15	12	19	19	34	36	44	35	60	58	47																		
Path	35	20	37	21	17	24	31	33	35	42	32	54	59	50	63																	
Reas 1	46	33	48	41	33	04	15	29	47	51	32	66	39	41	51	52																
Reas 2	41	32	46	34	34	02	12	18	49	38	30	52	27	44	33	37	50															
Orient 1	45	41	45	35	29	14	20	40	47	50	39	52	55	54	53	58	49	41														
Orient 2	50	29	52	38	33	18	13	36	48	49	35	51	38	42	49	42	48	47	55													
Orient 3	59	51	62	48	43	05	15	55	60	63	54	61	42	42	51	46	49	53	67	56												
PTB-B																																
SRT-RT	-23	11	-20	-15	-15	-25	-07	-09	-14	-11	-08	-11	-21	-09	-22	-23	-29	-14	-17	-31	-18											
CRT-RT	-16	-05	-21	-02	-08	-20	-14	-18	-10	-14	-16	-16	-17	-19	-37	-33	-21	-12	-27	-37	-22	53										
PS&A-PC	25	-01	22	18	23	14	30	04	12	11	02	18	14	10	02	19	25	14	13	10	14	17	17									
PS&A-RT	-06	-04	-01	-02	00	-14	-15	-06	-06	-12	-18	-15	-17	-24	-26	-18	-03	01	-13	-16	-13	19	31	50								
PS&A-Slp	11	-00	17	07	09	-02	-01	-05	06	01	-14	04	-07	-12	-15	-01	12	14	00	01	-01	-03	09	52	82							
PS&A-Int	-23	-03	-30	-10	-12	-13	-15	-04	-14	-17	-01	-31	-21	-16	-28	-28	-30	-25	-28	-26	-15	32	31	-27	-06	-61						
Targ ID-PC	12	07	07	21	09	-08	10	07	06	12	01	12	05	09	-05	13	12	16	05	03	11	11	07	40	33	32	-13					
Targ ID-RT	-23	-31	-29	-19	-06	-05	-02	-34	-25	-37	-32	-41	-46	-34	-50	-45	-29	-20	-34	-37	-35	23	42	20	47	32	13	16				
STM-PC	21	-05	16	14	12	17	25	01	05	12	05	15	10	15	02	18	14	23	10	20	20	-06	04	50	17	26	-23	25	03			
STM-RT	-02	02	-10	11	05	-09	-13	-02	-01	-09	-08	-13	-20	-13	-28	-22	-06	-05	-12	-20	-06	23	40	26	49	25	23	27	47	08		
STM-Slp	12	06	09	07	10	10	06	01	08	11	-02	09	-00	05	02	10	10	09	05	06	11	-06	03	28	29	26	-11	18	13	32	39	
STM-Int	-10	-02	-17	-08	-00	-16	-17	-03	-06	-16	-08	-19	-22	-16	-31	-29	-12	-11	-17	-25	-14	28	42	10	35	11	31	10	42	-11	83	-19
Cam Shoot	-11	-13	-14	-13	01	-04	-03	-17	-03	-15	-15	-14	-17	-07	-15	-17	-08	-03	-11	-08	-08	13	10	00	00	-01	11	-09	25	-02	25	14
Nbr Mem-PC	60	42	38	41	36	35	20	33	36	35	34	35	26	25	29	32	48	31	37	33	43	-16	09	29	02	15	-20	14	-12	23	-09	02
Nbr Mem-RT	-49	-27	-46	-22	-21	-45	-29	-23	-42	-23	-31	-17	-26	-25	-34	-26	-20	-18	-35	-29	-36	21	24	11	34	29	03	10	27	07	18	15
PTB-C																																
Trk 1-Dist	-20	-21	-31	-16	-07	09	-06	-33	-24	-44	-28	-38	-34	-23	-35	-41	-29	-26	-28	-39	-36	14	25	-12	08	00	11	-04	42	-29	25	-15
Trk 2-Dist	-19	-23	-28	-15	-03	07	-05	-33	-21	-38	-29	-37	-28	-14	-32	-35	-22	-22	-25	-36	-32	11	19	-01	11	04	09	02	39	-19	25	-01
TSHT-Time	-09	-08	-16	-03	05	-02	-00	-18	-10	-21	-17	-10	-14	-02	-23	-17	-11	-01	-09	-14	-06	08	16	16	22	09	12	12	32	09	22	15
TSHT-Dist	-09	-19	-17	-09	-02	13	00	-26	-12	-31	-23	-31	-22	-14	-30	-37	-24	-14	-23	-24	-28	08	16	-07	03	-00	09	-11	32	-12	27	-16

Note: Decimals have been omitted.

[a]PTB-A = Pilot Trial Battery paper-and-pencil tests.
PTB-B = Pilot Trial Battery computerized perceptual tests.
PTB-C = Pilot Trial Battery computerized psychomotor tests.

Table II.39

Mean Correlations, Standard Deviations, and Minimum Correlations Between Scores on ASVAB Subtests and Pilot Trial Battery Tests of Cognitive, Perceptual, and Psychomotor Abilities

Types of Scores Correlated	Number of Correlations	Mean[a] Correlation	SD[a] of Correlation	Minimum[a] Correlation
ASVAB Subtests and PTB Cognitive Paper-and-Pencil Tests	110	.33	.14	.01
ASVAB Subtests and PTB Cognitive/ Perceptual Computer-Administered Tests	165	.15	.12	.00
ASVAB Subtests and PTB Psychomotor Computer-Administered Tests	44	.17	.12	.00
PTB Cognitive Paper-and-Pencil Tests and PTB Perceptual Computer-Administered Tests	150	.19	.11	.00
PTB Cognitive Paper-and-Pencil Tests and PTB Psychomotor Computer-Administered Tests	40	.24	.11	.01
PTB Perceptual Computer-Administered Tests and PTB Psychomotor Computer-Administered Tests	60	.15	.11	.00

[a]These statistics are based on absolute correlation values.

Table II.40

Principal Components Factor Analysis of Scores of the ASVAB Subtests, Cognitive Paper-and-Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer-Administered Tests[a] (N = 168)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	h^2
ASVAB								
GS	75							59
AR	75							73
WK	77							62
PC	62							47
NO						84		77
CS						62		44
AS	62							58
MK	77							70
MC	63	38	-30					68
EI	72							65
COGNITIVE PAPER-AND-PENCIL								
Assemb Obj	35	69						66
Obj Rotation		-61						49
Shapes		66						51
Maze		70						67
Path		67	-30					65
Reason 1	37	58						54
Reason 2	37	47						44
Orient 1	37	64						58
Orient 2	40	46			-30			52
Orient 3	60	52						67
PERCEPTUAL COMPUTER								
SRT-RT					63			44
CRT-RT					61			50
PS&A-PC				67	31			70
PS&A Slope				88				81
PS&A Inter				-65	50			74
Target ID-PC				40				25
Target ID-RT		-41	37		30			57
STM-PC				30			34	41
STM-Slope							41	25
STM-Int			38		51			47
Cannon Shoot-TE			32					19
No Mem-PC	53					37		52
No Mem-RT	-37					-46		54
PSYCHOMOTOR COMPUTER								
Tracking 1			86					82
Tracking 2			77					66
Target Shoot-TF						42		23
Target Shoot-Dist			64					48
Variance Explained	5.69	4.70	2.83	2.37	1.92	1.87	1.17	

Note: Decimals have been omitted from factor loadings.

[a] Note that the following variables were not included in this factor analysis: AFQT, PS&A Reaction Time, and Short-Term Memory Reaction Time.

h^2 - communality (sum of squared factor loadings) for variables.

Factor 2 includes all of the PTB cognitive paper-and-pencil measures, Mechanical Comprehension from the ASVAB, and Target Identification reaction time from the computer tests. We called this a general spatial factor.

Factor 3 has major loadings on the three psychomotor tests, with substantially smaller loadings from three cognitive/perceptual computer test variables, the Path Test, and Mechanical Comprehension from the ASVAB. Given the high loadings of the psychomotor tests on this factor, we refer to this as the motor factor.

Factor 4 includes variables from the cognitive/perceptual computer tests. This factor appears to involve accuracy of perception across several tasks and types of stimuli.

Factor 5 is not that clear, but the highest loadings are on straightforward reaction time measures, so we interpret this as a speed of reaction factor.

Factor 6 contains four variables, two from the ASVAB and two from the cognitive/perceptual computer tests. This factor appears to represent both speed of reaction and arithmetic ability.

Factor 7 contains three variables from the computer tests. These include Short-Term Memory percent correct and slope, and Target Shoot time-to-fire. This factor is difficult to interpret, but we believe it may represent a response style factor. That is, this factor suggests that those individuals who take a longer time to fire on the Target Shoot Test also tend to have higher slopes on the Short-Term Memory (lower processing speeds with increased bits of information) but are more accurate or obtain higher percent-correct values on the Short-Term Memory test.

Note that several variables have fairly low communalities. These may be due to relatively low score variance or reliability, but it could also be due to these variables having unique variance, at least when factor analyzed with this set of tests. We think this latter explanation is highly plausible for the Cannon Shoot score.

Field Test of Non-Cognitive Measures (ABLE and AVOICE)

In this section are described the field test results for the non-cognitive predictor measures, including the descriptive scale statistics and the results of the fakability analyses. The samples were different for the different analyses, and each is described in turn.

Scale Analyses

These analyses were performed to obtain descriptive scale statistics and examine the covariation among scales. Only the Fort Knox data were used.

Sample. At Fort Knox a total of 290 soldiers completed the ABLE and 287 soldiers the AVOICE. After deletion of inventories with greater than 10% missing data for both measures, and of those ABLEs where scores on the Non-Random Response Scale were less than six, a total of 276 ABLEs and 270 AVOICEs were available for analysis. For the experiment in which portions of the Pilot Trial Battery were re-administered to soldiers 2 weeks after the first administration, the total number of "Time 2" ABLE and AVOICE inventories, after the data quality screens had been applied, was 109 and 127, respectively.

Results. Summary statistics for the ABLE and AVOICE are presented in Tables II.41, II.42 (ABLE), and Table II.43 (AVOICE). As can be seen, the median alpha coefficient (internal consistency) for the ABLE content scales is .84, and the median test-retest correlation for the ABLE content scales is .79, with a range of .68 to .83. At retest or second testing, the soldiers apparently responded in a somewhat more random way. The response validity scale, Non-Random Responses, detected it and, consequently, the correlation between first and second testing was low, .37. The median alpha coefficient (internal consistency) for the AVOICE scales is .86. The median test-retest correlation for the AVOICE scales is .76.

The ABLE content scales and the AVOICE scales were separately factor analyzed, and in both cases the two-factor solution appeared to best summarize the data. As shown in Tables II.44 (for ABLE) and II.45 (for AVOICE), the temperament factors were labeled Personal Impact and Dependability, and the interest factors were labeled Combat-Related and Combat Support.

Fakability Analyses

Recall that there were four validity scales on the ABLE: Non-Random Responses, Unlikely Virtues (Social Desirability), Poor Impression, and Self-Knowledge. To investigate intentional distortion of responses, data were gathered (a) from soldiers instructed, at different times, to distort their responses or to be honest (experimental data gathered at Fort Bragg); (b) from soldiers who were simply responding to the ABLE and AVOICE with no particular directions (data gathered at Fort Knox); and (c) from recently sworn-in Army recruits at the Minneapolis Military Entrance Processing Station (MEPS).

The purposes of the faking study were to:

- Determine the extent to which soldiers can distort their responses to temperament and interest inventories when instructed to do so. (Compare data from Fort Bragg faking conditions with Fort Bragg and Fort Knox honest conditions.)

- Determine the extent to which the ABLE response validity scales detect such intentional distortion. (Compare response validity scales in Fort Bragg honest and faking conditions.)

- Determine the extent to which ABLE validity scales can be used to correct or adjust scores for intentional distortion.

Table II.41

ABLE Scale Score Characteristics: Fort Knox Field Test (N = 276 except where otherwise noted)

Scale	No. of Items	Mean Time 1	SD	Alpha	Test-Retest[a] r	Median Item-Scale r
Content Scales						
Emotional Stability	29	64.9	8.27	.86	.68	.44
Self-Esteem	15	35.1	5.25	.83	.81	.54
Cooperativeness	24	54.1	6.09	.77	.69	.42
Conscientiousness	21	48.9	5.90	.81	.73	.43
Nondelinquency	24	55.4	7.23	.84	.81	.46
Traditional Values	16	37.2	4.60	.70	.74	.45
Work Orientation	27	61.2	7.93	.85	.80	.47
Internal Control	21	50.3	6.14	.79	.75	.43
Energy Level	25	57.1	7.11	.85	.79	.47
Dominance	16	35.5	6.13	.86	.83	.56
Physical Condition	9	31.1	7.53	.87	.81	.72
Response Validity Scales						
Unlikely Virtues	12	16.6	3.39	.68	.62	.53
Self-Knowledge	13	29.6	3.54	.62	.71	.41
Non-Random Response[b]	8	7.7	.71	.56	.37	.45
Poor Impression	24	1.5	1.86	.61	.56	.33

[a]N=109 for test-retest correlations. Test-retest interval was two weeks.

[b]N=281. Statistics reported for this scale are based on sample edited for overall missing data only. "Passing" score on non-random response scale $\leq$ 6.

Table II.42

ABLE Test-Retest Results[a]: Fort Knox Field Test

Scale	Mean Time 1 (N = 276)	Mean Time 2 (N =109)	Effect Size[b]
Content Scales			
Emotional Stability	64.9	65.1	.02
Self-Esteem	35.1	34.8	-.05
Cooperativeness	54.1	54.3	.04
Conscientiousness	48.9	48.3	-.10
Nondelinquency	55.4	55.6	.02
Traditional Values	37.2	37.9	.15
Work Orientation	61.2	60.7	-.07
Internal Control	50.3	50.2	-.01
Energy Level	57.1	57.0	-.01
Dominance	35.5	34.9	-.09
Physical Condition	31.1	30.4	-.09
Response Validity Scales			
Unlikely Virtues	16.6	17.5	.27
Self-Knowledge	29.6	29.0	-.18
Non-Random Response[c]	7.7	7.2	-.65
Poor Impression	1.	1.2	-.18

[a]Test-Retest interval was two weeks.

[b]Effect Size = (Mean Time 1 - Mean Time 2)/S_2D_2 Time 1.

[c]Based on sample edited for missing data only; N_1 = 281 and N_2 = 121.

Table II.43

AVOICE Scale Score Characteristics: Fort Knox Field Test (N = 270 except where otherwise noted)

Scale	No. of Items	Mean	SD	Alpha	Test-Retest[a] r	Median Item-Scale r
Marksman	5	15.8	4.37	.79	.77	.75
Agriculture	5	14.1	3.99	.68	.69	.70
Mathematics	5	15.1	4.37	.82	.76	.79
Aesthetics	5	14.3	4.17	.77	.72	.74
Leadership	6	20.3	4.70	.81	.56	.74
Electronic Communication	7	21.1	5.73	.92	.78	.72
Automated Data Processing	7	23.4	6.56	.88	.81	.81
Teaching/Counseling	7	22.8	5.53	.82	.73	.73
Drafting	7	21.5	6.12	.85	.74	.77
Audiographics	7	23.8	5.68	.82	.76	.70
Armor/Cannon	8	22.4	6.57	.83	.74	.69
Vehicle/Equipment Operator	10	28.1	7.79	.86	.69	.70
Outdoors	9	31.7	6.41	.79	.69	.66
Infantry	10	29.1	7.13	.81	.78	.65
Science/Chemical Operations	11	29.4	8.93	.89	.79	.71
Supply Administration	13	35.0	10.44	.92	.82	.75
Office Administration	16	45.2	13.20	.94	.86	.73
Law Enforcement	16	48.1	11.84	.88	.78	.63
Mechanics	16	50.0	14.68	.95	.80	.80
Electronics	20	60.0	17.48	.96	.74	.77
Heavy Construction/Combat	23	65.8	17.90	.94	.76	.70
Medical Services	24	68.5	18.79	.95	.84	.69
Food Service	17	48.2	11.16	.89	.71	.64

[a]N=127 for test-retest correlations.

Table II.44

ABLE Factor Analysis[a]: Fort Knox Field Test

	I Personal Impact	II Dependability	h^2
Self-Esteem	.80	.30	.73
Energy Level	.73	.46	.74
Dominance (Leadership)	.72	.13	.54
Emotional Stability	.67	.26	.52
Work Orientation	.67	.51	.71
Nondelinquency	.20	.81	.70
Traditional Values	.19	.73	.57
Conscientiousness	.39	.72	.67
Cooperativeness	.46	.60	.57
Internal Control	.44	.50	.44
			6.19

Note: h^2 = communality, the sum of squared factor loadings for a variable.

[a]Principal factor analysis, varimax rotation.

Table II.45

AVOICE Factor Analysis[a]: Fort Knox Field Test

Scale	I Combat Support[b]	II Combat-Related[c]	h^2
Office Administration	.85	-.13	.73
Supply Administration	.78	.11	.62
Teaching/Counseling	.76	.11	.59
Mathematics	.74	.09	.55
Medical Services	.73	.18	.57
Automated Data Processing	.71	.10	.51
Audiographics	.64	.17	.44
Electronic Communication	.64	.36	.54
Science/Chemical Operations	.61	.43	.55
Aesthetics	.61	.04	.37
Leadership	.58	.35	.46
Food Service	.54	.19	.33
Drafting	.54	.34	.41
Infantry	.10	.85	.74
Armor/Cannon	.13	.84	.73
Heavy Construction/Combat	.17	.84	.73
Outdoors	.02	.74	.55
Mechanics	.17	.74	.58
Marksman	.05	.73	.54
Vehicle/Equipment Operator	.17	.73	.56
Agriculture	.18	.64	.44
Law Enforcement	.27	.61	.44
Electronics	.45	.57	.52
			12.49

Note: h^2 = communality, the sum of squared factor loadings for a variable.

[a] Principal factor analysis, varimax rotation.

[b] Conventional, Social, Investigative, Enterprising, Artistic constructs.

[c] Realistic construct.

- Determine the extent to which distortion is a problem in an applicant setting. (Compare MEPS data with Fort Bragg and Fort Knox data.)

Subjects. The participants in the experimental group were 425 enlisted soldiers in the 82nd Airborne brigade at Fort Bragg. Comparison samples were MEPS candidates (N = 126) and the Fort Knox soldiers described earlier (N = 276).

Procedure and Design. Four faking and two honest conditions were created:

- Fake Good on the ABLE
- Fake Bad on the ABLE
- Fake Combat on the AVOICE
- Fake Non-combat on the AVOICE
- Honest on the ABLE
- Honest on the AVOICE

The significant parts of the instructions for the six conditions were as follows:

ABLE - Fake Good
Imagine you are at the Military Entrance Processing Station (MEPS) and you want to join the Army. Describe yourself in a way that you think will ensure that the Army selects you.

ABLE - Fake Bad
Imagine you are at the Military Entrance Processing Station (MEPS) and you do not want to join the Army. Describe yourself in a way that you think will ensure that the Army does not select you.

ABLE - Honest
You are to describe yourself as you really are.

AVOICE - Fake Combat
Imagine you are at the Military Entrance Processing Station (MEPS). Please describe yourself in a way that you think will ensure that you are placed in an occupation in which you are likely to be exposed to combat during a wartime situation.

AVOICE - Fake Non-combat

Imagine you are at the Military Entrance Processing Station (MEPS). Please describe yourself in a way you think will ensure that you are placed in an occupation in which you are unlikely to be exposed to combat during a wartime situation.

AVOICE - Honest
You are to describe yourself as you really are.

The design was a 2x2x2 with repeated measures on faking and honest conditions which were counterbalanced. Thus, approximately half the experimental group, 124 soldiers, completed the inventories honestly in the morning and faked in the afternoon, while the other half (121) completed the inventories honestly in the afternoon and faked in the morning. The first between-subjects factor consisted of these two levels: Fake Good/Want Combat and Fake Bad/Do Not Want Combat. Order was manipulated in the second between-subjects factor such that the following two levels were produced: Faked responses then honest responses, and honest responses then faked responses.

Faking Study Results - Temperament Inventory. A multivariate analysis of variance (MANOVA) on the Fort Bragg data on the temperament inventory shows that all the Fake by Set interactions are significant, indicating that soldiers can, when instructed to do so, distort their responses. The order of conditions in which the participant completed the ABLE also affected the results. Table II.46 shows the mean scores for the various experimental conditions.

Another research question was the extent to which the response validity scales detected intentional distortion. As can be seen in Table II.47, the response validity scale Social Desirability (Unlikely Virtues) detects Faking Good on the ABLE; the response validity scales Non-Random Response, Poor Impression, and Self-Knowledge detect Faking Bad. According to these data, the soldiers responded more randomly, created a poorer impression, and reported that they knew themselves less well when told to describe themselves in a way that would increase the likelihood that they would not be accepted into the Army.

We also examined the extent to which the response validity scales Social Desirability and Poor Impression could be used to adjust ABLE content scale scores for Faking Good and Faking Bad. Social Desirability was partialled from the content scales in the Fake Good condition and Poor Impression from the content scales in the Fake Bad condition.

Table II.48 shows the adjusted mean differences, which clearly show that these response validity scales can be used to adjust content scales. However, two important unknowns remain: Do the adjustment formulas developed on these data cross-validate, and do they increase criterion-related validity?

As noted earlier, in an effort to explore the extent to which intentional distortion may be a problem in an applicant setting, the ABLE and AVOICE were administered at the Minneapolis MEPS. However, the sample of 126 recruits who completed the inventories were not true "applicants," in that they had just recently been sworn into the Army. To approximate the applicant response set as closely as possible, recruits were allowed to believe that their scores on these inventories might impact on their Army careers. Recruits were then asked to complete the ABLE and AVOICE, after which they were debriefed.

To examine the extent to which recruits actually believed their ABLE and AVOICE scores would have an effect on their Army career, a single item was filled out by each recruit prior to debriefing. Of the 126 recruits in this

Table II.46

Honesty and Faking Effects, ABLE Content Scales: Fort Bragg

Scale	Honest First[a]			Fake Good First[a]			Fake Bad First[a]			Estimated Effect Size Honest vs. Good	Estimated Effect Size Honest vs. Bad
	N	M	SD	N	M	SD	N	M	SD		
Emotional Stability	120	66.1	7.8	54	70.3	10.2	54	50.1	10.8	-.49	1.81
Self-Esteem	115	34.2	4.7	54	38.2	5.4	54	22.2	5.8	-.69	2.48
Cooperativeness	121	53.2	6.3	54	55.5	8.8	54	36.7	10.4	-.32	2.12
Conscientiousness	116	46.3	5.8	54	49.6	8.4	54	31.7	8.7	-.49	2.13
Nondelinquency	115	53.1	6.2	54	54.8	10.2	54	36.8	9.6	-.22	2.19
Traditional Values	116	36.7	4.6	54	38.7	6.5	54	23.6	6.1	-.38	2.56
Work Orientation	120	59.3	7.6	54	64.7	10.3	54	40.8	11.7	-.63	2.04
Internal Control	115	49.5	6.3	54	50.9	8.2	54	35.6	8.9	-.20	1.92
Energy Level	116	57.5	6.9	54	61.4	9.1	54	37.9	9.9	-.51	2.46
Dominance (Leadership)	116	35.6	5.6	54	40.3	5.6	54	24.5	6.6	-.84	1.87
Physical Condition	116	33.0	7.4	54	35.4	7.7	54	18.3	8.6	-.32	1.88

[a] Mean scores are based on persons who responded to this condition first.

Table II.47

Honesty and Faking Effects, ABLE Response Validity Scales: Fort Bragg

ABLE Response Validity Scale	Honest First[a]			Fake Good First[a]			Fake Bad First[a]			Effect Size Honest vs. Fake Good	Effect Size Honest vs. Fake Bad
	N	M	SD	N	M	SD	N	M	SD		
Unlikely Virtues (Social Desirability)	109	15.8	3.1	57	20.1	5.8	56	17.8	4.8	-1.02	-.53
Self-Knowledge	109	29.6	3.6	57	29.7	4.1	56	21.8	5.2	.03	1.85
Non-Random Response	109	7.6	1.0	57	7.0	1.8	56	2.8	2.2	.45	3.16
Poor Impression	109	1.5	2.1	57	1.7	2.2	56	14.6	7.9	-.09	-2.67

[a] Values are based on the sample that completed the questionnaires under the condition of interest first.

Table II.48

Effects of Regressing Out Response Validity Scales (Social Desirability and Poor Impression) on Faking Condition ABLE Content Scale Scores: Fort Bragg

Content Scales	Fake Good		Fake Bad	
	Adjusted Standardized Mean Difference[a]	Correlation With Social Desirability[b]	Adjusted Standardized Mean Difference [a]	Correlation with Poor Impression[b]
Emotional Stability	-.14	.14	-.14	-.41
Self-Esteem	-.64	.19	.77	-.40
Cooperativeness	.06	.30	.38	-.47
Conscientiousness	-.17	.31	.31	-.38
Nondelinquency	.13	.31	.63	-.42
Traditional Values	-.24	.25	1.00	-.40
Work Orientation	-.33	.30	.32	-.38
Internal Control	.03	.15	.22	-.44
Energy Level	-.12	.24	.45	-.41
Dominance (Leadership)	-.63	.25	.32	-.38
Physical Condition	-.07	.20	.35	-.39

[a]Standard mean differences are [Mean (Honest) minus Mean (Fake)]/SD (Honest).

[b]Correlations are average of correlations for first administration under Honest and relevant Fake condition.

sample, 57 responded "yes" to the question of whether they believed scores would have an impact, 61 said "no," and 8 wrote in that they didn't know. Thus, while the MEPS sample was not a true "applicant" sample, its make-up was reasonably close to it (recently sworn-in recruits, close to half of whom believed their ABLE and AVOICE scores would affect their Army career).

Table II.49 shows mean scores for MEPS recruits and the two "honest" conditions of this study, Fort Bragg and Fort Knox. In total, these results suggest that intentional distortion may not be a significant problem in an applicant setting. (Faking or distortion in a draft situation cannot even be estimated in the present U.S. situation.)

Overall, the ABLE data show that:

- Soldiers can distort their responses when instructed to do so.
- The response validity scales detect intentional faking.
- An individual's Social Desirability scale score can be used to adjust his or her content scale scores to reduce variance associated with faking.
- Faking or distortion may not be a significant problem in an applicant setting.

Faking Study Results - Interest Inventory. We divided the Interest scales into the two groups, Combat-Related and Combat Support, that emerged from the factor analyses and multivariate analysis of variance (MANOVA) on the Fort Bragg data. Nine of the 11 Combat-Related AVOICE scales are sensitive to intentional distortion, and 9 of the 12 Combat Support scales are sensitive to intentional distortion. The interactions of Fake by Set by Order are significant at $p = .05$, indicating that order of conditions in which the participants completed the AVOICE also affected the result. Tables II.50 and II.51 show mean scores for the various conditions when the particular condition was the first administration.

When told to distort their responses so that they are not likely to be placed in combat-related occupational specialties (MOS)--that is, instructed to Fake Non-combat--soldiers tended to decrease their scores on all scales. Scores on 19 of 24 interest scales were lower in Fake Non-combat as compared to the honest condition. In the Fake Combat condition, soldiers in general increased their Combat-Related scale scores and decreased their Combat Support scale scores.

The ABLE response validity scales were then used to adjust AVOICE scale scores for Faking Combat and Faking Non-combat. Comparing these differences to the unadjusted differences revealed that these adjustments have little effect, perhaps because the response validity scales consisted of items from the ABLE and the faking instructions for the ABLE and AVOICE were different (i.e., Fake Good and Fake Bad vs. Fake Combat and Fake Non-combat).

Again the question was explored of whether applicants might tend to distort their responses to the AVOICE. The mean scores for the MEPS recruits and the two Honest conditions, Fort Bragg and Fort Knox, showed no particular pattern to the mean score differences.

Table II.49

Comparison of Fakability Results from Fort Bragg (Honest), Fort Knox, and MEPS (Recruits) ABLE Scales

ABLE Scale	Fort Bragg Honest[a] N	Fort Bragg Honest[a] Mean	MEPS (Recruits) N	MEPS (Recruits) Mean	Fort Knox N	Fort Knox Mean	Total SD	Degrees of Freedom	F	P
Response Validity Scales										
Social Desirability (Unlikey Virtues)	116	15.91	121	16.63	276	16.60	3.21	2,510	2.15	.12
Self-Knowledge	116	29.54	121	28.03	276	29.64	3.63	2,510	9.10	.00
Non-Random Response	116	7.58	121	7.79	276	7.75	.64	2,510	3.73	.02
Poor Impression	116	1.50	121	1.05	276	1.54	1.84	2,510	3.15	.04
Content Scales										
Emotional Stability	112	66.22	118	66.03	272	65.05	7.86	2,499	1.18	.31
Self-Esteem	112	34.77	118	34.04	272	35.12	5.00	2,499	1.93	.15
Cooperativeness	112	53.33	118	54.60	272	54.19	6.05	2,499	1.34	.26
Conscientiousness	112	46.37	118	46.49	272	48.97	5.86	2,499	12.24	.00
Nondelinquency	112	53.24	118	54.36	272	55.49	6.91	2,499	4.48	.01
Traditional Values	112	36.67	118	36.97	272	37.28	4.50	2,499	.77	.46
Work Orientation	112	59.71	118	58.37	272	61.40	7.73	2,499	6.90	.00
Internal Control	112	49.48	118	51.90	272	50.37	6.13	2,499	4.75	.01
Energy Level	112	57.56	118	56.67	272	57.19	6.95	2,499	.48	.62
Dominance (Leadership)	112	35.54	118	32.84	272	35.41	6.05	2,499	8.69	.00
Physical Condition	112	32.96	118	28.27	272	31.08	7.49	2,499	12.10	.00

[a] Scales are based on persons who responded to the Honest Condition first.

Table II.50

Effects of Faking, AVOICE Combat Scales: Fort Bragg

AVOICE Combat Scales	Honest[a]			Fake Combat[a]			Fake Noncombat[a]			Effect Size Honest vs Combat	Effect Size Honest vs Combat
	N	M	SD	N	M	SD	N	M	SD		
Marksman	122	18.1	4.5	59	20.2	3.9	60	12.8	5.9	-.49	1.06
Agriculture	124	15.0	3.8	59	12.9	3.6	60	15.1	4.0	.56	-.03
Armor/Cannon	124	24.2	5.8	59	28.9	7.6	60	15.1	6.3	-.73	1.53
Vehicle/Equipment	124	28.7	6.4	59	26.6	7.9	60	23.5	8.0	.30	.75
Outdoors	123	36.0	6.1	59	38.3	6.0	60	25.7	10.2	-.38	1.34
Infantry	123	33.5	6.8	59	37.8	8.2	59	20.5	8.4	-.59	1.77
Law Enforcement	124	53.3	10.8	59	54.5	12.1	60	42.3	12.5	-.11	.97
Heavy Construction	124	70.5	16.3	59	68.9	15.0	59	58.7	16.4	.10	.72
Mechanics	124	50.7	12.7	59	44.6	15.2	60	47.3	13.6	.45	.26
Electronics	124	58.1	18.3	59	50.3	17.3	60	56.8	18.0	.43	.07
Adventure	108	37.5	4.3	56	38.1	3.7	54	26.8	6.6	-.15	2.06

[a] Values are based on the sample that completed the questionnaire under the condition of interest first.

[b] Effect Size = (Mean Honest minus Mean Combat, or Noncombat)/SD Total.

Table II.51

Effects of Faking, AVOICE Combat Support Scales: Fort Bragg

AVOICE Combat Support Scales	Honest[a]			Fake Combat[a]			Fake Noncombat[a]			Effect Size Honest vs Combat	Effect Size Honest vs Combat
	N	M	SD	N	M	SD	N	M	SD		
Mathematics	120	14.2	4.7	56	11.8	4.7	59	15.6	5.0	.51	-.29
Aesthetics	120	14.6	4.1	57	12.1	4.6	59	17.1	5.5	.59	-.54
Leadership	124	22.3	4.2	59	21.5	4.1	59	17.3	5.8	.19	1.05
Electronic Communications	123	21.1	6.1	59	21.8	7.0	60	14.2	5.6	-.11	1.16
Automated Data Processing	122	20.4	6.7	58	15.5	7.2	59	23.8	7.4	.71	-.49
Teaching/Counseling	124	23.8	5.7	59	20.7	5.6	60	21.0	5.6	.55	.49
Drafting	124	22.3	6.1	59	18.4	6.2	60	21.5	5.5	.64	.14
Audiographics	124	23.5	5.6	59	18.7	6.2	60	20.7	5.6	.83	.50
Science/Chemical Operations	123	28.0	8.4	59	28.0	9.2	60	25.8	9.6	0	.25
Supply Administration	124	30.5	9.8	59	26.4	9.6	60	35.3	11.9	.42	-.46
Office Administration	123	38.5	13.3	59	31.2	12.3	59	49.5	17.2	.56	-.75
Medical Services	124	67.8	18.3	59	60.4	17.5	60	61.0	17.8	.41	.37
Food Service	122	38.0	10.2	59	31.0	10.6	59	45.8	16.3	.68	.62

[a]Values are based on the sample that completed the questionnaire under the condition of interest first.

[b]Effect Size = (Mean Honest minus Mean Combat, or Noncombat)/SD Total.

Overall, the AVOICE data show that:

- Soldiers can distort their responses when instructed to do so.
- The ABLE Social Desirability and Poor Impression scales are not as effective for adjusting AVOICE scale scores as they are for adjusting ABLE content scale scores.
- Faking or distortion may not be a significant problem in an applicant setting.

Summary of Field Test Results

The data on field tests presented in this section were crucial for the final revisions of the Pilot Trial Battery (PTB) before it was used in the Concurrent Validation. The revisions based on the field tests are described in Section 7.

Section 7

TRANSFORMING THE PILOT TRIAL BATTERY INTO THE TRIAL BATTERY

The entire Pilot Trial Battery, as administered at the field tests, required approximately 6.5 hours of administration time. However, the Trial Battery, which was the label reserved for the predictor battery to be used in the full-scale Concurrent Validation, had to fit in a 4-hour time slot.

Three general principles, consonant with the theoretical and practical orientation that had been used since the inception of the project, guided the revision and reduction decisions:

1. Maximize the heterogeneity of the battery by retaining measures of as many different constructs as possible.

2. Maximize the chances of incremental validity and classification efficiency.

3. Retain measures with adequate reliability.

Using all accumulated information, the final decisions were made in a series of meetings attended by the project staff and by the Scientific Advisory Group. Considerable discussion was generated at these meetings, but the group was able to reach a consensus on the reductions and revisions to be made to the Pilot Trial Battery.

The recommendations for revisions and the reasons for their adoption are described in the following pages.

Changes to Cognitive Paper-and-Pencil Tests

Changes to the cognitive paper-and-pencil tests are summarized in Table II.52.

The Spatial Visualization construct in the PTB was measured by three tests: Assembling Objects, Object Rotation, and Shapes. The Shapes Test was dropped because the evidence of validity for job performance for tests of this type was judged to be less impressive than for the other two tests. The Object Rotation Test was not changed. Eight items were dropped from the Assembling Objects Test by eliminating those items that were very difficult or very easy, or that had low item-total correlations. The time limit for Assembling Objects was not changed; the effect was to make Assembling Objects more a power test than it was prior to the changes.

The Spatial Scanning construct was measured by two tests, Mazes and Path. The Path Test was dropped and the Mazes Test was retained with no changes. Mazes showed higher test-retest reliabilities than Path (.71 vs. .64), and gain scores were lower (.24 SD units for Mazes vs. .62 SD units for Path), which was desirable. Also, Mazes was a shorter test than Path (5.5 minutes versus 8 minutes).

Table II.52

Summary of Changes to Paper-and-Pencil Cognitive Measures in the Pilot Trial Battery

Test Name	Changes
Assembling Objects	Decrease from 40 to 32 items.
Object Rotation	Retain as is with 90 items.
Shapes	Drop Test.
Mazes	Retain as is with 24 items.
Path	Drop Test.
Reasoning 1	Retain as is with 30 items. New name REASONING TEST.
Reasoning 2	Drop Test.
Orientation 1	Drop Test.
Orientation 2	Retain as is with 24 items. New name ORIENTATION TEST.
Orientation 3	Retain as is with 20 items. New name MAP TEST.

The Figural Reasoning construct was measured by Reasoning Test 1 and Reasoning Test 2. Reasoning 1 was evaluated as the better of the two tests because it had higher reliabilities for both internal consistency (alpha = .83 vs. .65 and separately timed, split-half coefficients = .78 vs. .63) and test-retest (.64 vs. .57), as well as a higher uniqueness estimate (.49 vs. .37). Reasoning 1 was retained with no item or time limit changes and Reasoning 2 was dropped. Reasoning Test 1 was renamed Reasoning Test.

Three tests in the PTB measured the Spatial Orientation construct. Orientation Test 1 was dropped because it showed lower test-retest reliabilities (.67 vs. .80 and .84) and higher gain scores (.63 SD units vs. .11 and .08 SD units). In addition, we modified the instructions for Orientation Test 2 because field test experience had indicated that the PTB instructions were not as clear as they should be. Orientation Test 2 was renamed Orientation Test. Orientation Test 3 was retained with no changes and renamed the Map Test.

Changes to Computer-Administered Tests

Besides the changes made to specific tests, several improvements were made to the computer battery as a whole.

General Improvements

The more general changes were as follows.

1. Virtually all test instructions were modified:

   - Most instructions were shortened considerably.
   - Names of buttons, slides, and switches on the response pedestals were written in capital letters whenever they appeared.
   - Where possible, the following standard outline was followed in preparing the instructions:

     - Test name
     - One-sentence description of the purpose of the test
     - Step-by-step test instructions
     - One practice item
     - Brief restatement of test instructions
     - Two or three additional practice items
     - Instructions to call test administrator if there are questions about the test.

2. Whenever the practice items had a correct response, the subject was given feedback.

3. Rest periods were eliminated from the battery. This was possible because virtually every test was shortened.

4. The computer programs controlling test administration were merged into one super-program, eliminating the time required to load the programs between tests.

5. The format and parameters used in the software containing test times were reworded, so that the software was more "self-documented."

6. The total time allowed for subjects to respond to a test item (in other words, response time limit) was set at 9.0 seconds for all reaction time tests (Simple and Choice Reaction Time, Short-Term Memory, Perceptual Speed and Accuracy, and Target Identification). In the PTB version, the response time limit had varied from test to test, for no particular reason.

7. Also, with regard to the reaction time tests, the software was changed so that the stimulus for an item disappears when the subject lifts his or her hand from the home button (to make a response). Subjects had been instructed not to lift their hands from the home button until they had determined the correct response, so that separate measures of decision and movement time could be obtained.

However, more than a few of the field test subjects continued to study the item stimulus after leaving the home buttons. By causing the item to disappear, we hope to eliminate that problem.

Changes to Specific Tests

Changes to the individual computer-administered tests are summarized in Table II.53.

No changes were recommended for Simple Reaction Time. However, the order of the pretrial intervals (the interval between the time the subject depresses the home buttons and the trial stimulus appears) was randomized.

Fifteen items were added to Choice Reaction Time in an attempt to increase the test-retest reliability for mean reaction time on this test.

Twelve items were eliminated from the Perceptual Speed and Accuracy Test (reduced from 48 to 36 items), primarily to save time. Internal consistency estimates were high for scores on this test (.83, .96, .88, and .74 for percent correct, mean reaction time, slope, and intercept, respectively), and reduction in the number of items did not seem to be cause for concern.

Several changes were made to the Target Identification Test. First, one of the two item types, the "moving" items, was eliminated. Field test data showed that scores of the "moving" and stationary items correlated .78, and the moving items had lower test-retest reliabilities than stationary items (.54 vs. .74). All target objects were made the same size (50% of the objects depicted as possible answers) since field test analyses indicated size had no appreciable effect on reaction time. A third level of angular rotation was added so that the target objects were rotated either 0°, 45°, or 75°. Theoretically, and as found in past research, reaction time is expected to increase with greater angular rotation. Two of the item parameters were not changed (position of correct response object and direction of target object). Finally, the number of items was reduced from 48 to 36 to save time. Internal consistency and test-retest estimates indicated that the level of risk attached to this reduction was acceptable. (For mean reaction time, the internal consistency estimate was .96 and the test-retest estimate was .67.) The test, as modified, then had two items in each of 18 cells determined by crossing angular rotation (0°, 45°, 75°), position of correct response object (left, center, or middle of screen), and direction of target object (left-facing or right-facing).

One item parameter (probe delay period) was eliminated from the Short-Term Memory Test, while two others were retained [item type (symbolic vs. letter) and item length (one, two, or five objects)]. Analyses of field test data showed that probe delay period did not significantly affect mean reaction time scores. To save time, 12 items were eliminated. Two of the three most important scores for this test appeared to have high enough reliabilities to withstand such a reduction.

The number of items on the Cannon Shoot Test was reduced from 48 to 36, again to save time. Internal consistency and test-retest reliabilities for the time error score were high enough (.88 and .66, respectively) to warrant

Table II.53

Summary of Changes to Computer-Administered Measures in the Pilot Trial Battery

Test Name	Changes
COGNITIVE/PERCEPTUAL TESTS	
Demographics	Eliminate race, age, and typing experience items. Retain SSN and video experience items.
Simple Reaction Time	No changes.
Choice Reaction Time	Increase number of items from 15 to 30.
Perceptual Speed & Accuracy	Reduce items from 48 to 36. Eliminate word items.
Target Identification	Reduce items from 48 to 36. Eliminate moving items. Allow stimuli to appear at more angles of rotation.
Short-Term Memory	Reduce items from 48 to 36. Establish a single item presentation and probe delay period.
Cannon Shoot	Reduce items from 48 to 36.
Number Memory	Reduce items from 27 to 18. Shorten item strings. Eliminate item part delay periods.
PSYCHOMOTOR TESTS	
Target Tracking 1	Reduce items from 27 to 18. Increase item difficulty.
Target Tracking 2	Reduce items from 27 to 18. Increase item difficulty.
Target Shoot	Reduce items from 40 to 30 by eliminating the extremely easy and extremely difficult items.

such reduction without the expectation of a significant impact on reliability. Also, the items were modified so that all targets are visible on the screen at the beginning of the trial and so that the subject is given at least 2 seconds to view the speed and direction of the target before the target reaches the optimal fire point.

Two modifications were made to Number Memory to reduce test administration time. The item part delay period was made a constant (1 second) rather than treated as a parameter with two levels (0.5 and 2.5 seconds), and the item string length (number of parts in an item) was changed from four, six, or eight parts to two, three, or four parts. These changes drastically reduced the time required to complete the test. As a result, the reduction in the number of items that had been recommended was not necessary. The Trial Battery version of this test had 28 items, constructed so that there were 13 replications of the four arithmetic operations (add, subtract, multiply, and divide).

Similar kinds of changes were made to Target Tracking Test 1 and Target Tracking Test 2. Since internal consistency and test-retest reliability estimates were relatively high, the number of items was reduced from 27 to 18. The difficulty of the test items was increased by increasing the speed of the crosshairs and the target; this was done because field test data indicated that the mean distance score was positively skewed. Also, the ratio of target to crosshairs speed, rather than target speed, was used as a test parameter. It seemed that, given a particular crosshairs speed, the ratio would be a better indicator of item difficulty than the actual target speed.

Several changes were made to the Target Shoot Test. First, all test items were classified according to three parameters: crosshairs speed, ratio of target to crosshairs speed, and item complexity (i.e., number of turns/mean segment length). Then, items were revised to achieve a balanced number of items in each cell when the levels of these parameters were crossed. This had the result of "un-confounding" these parameters so that analyses could be made to see which parameters contributed to item difficulty. Second, extremely difficult items were eliminated and item presentation times (the time the target was visible on the screen) were increased to a minimum of 6 seconds (and a maximum of 10 seconds). This was done to eliminate a severe missing data problem for such items (as much as 40%) discovered during field tests. Missing data occurred when subjects failed to "fire" at a target. These "no-fires" were found to occur where the target moved very rapidly or made many sudden changes in direction and speed, or the item lasted only a few seconds. The number of items was reduced from 40 to 30 to save testing time, primarily by eliminating the extremely easy items.

Changes to Non-Cognitive Measures (ABLE and AVOICE)

Changes to the non-cognitive measures (ABLE and AVOICE) are summarized in Table II.54.

Table II.54

Summary of Changes to Pilot Trial Battery Versions of Assessment of Background and Life Experiences (ABLE) and Army Vocational Interest Career Examination (AVOICE)[a]

Inventory/Scale Name	Changes
ABLE-Total	Decrease from 270 to approximately 209 items.
AVOICE-Total	Decrease from 309 to approximately 214 items.
AVOICE Expressed Interest Scales	Drop.
AVOICE Single Item Holland Scales	Drop.
AVOICE Agriculture Scale	Drop.
Work Environment Preference Scales	Move to criterion measure booklet (delete from AVOICE booklet).

[a] In addition to the changes outlined in this table by inventory scale, it was recommended that all ABLE item response options be standardized as three-option responses and all AVOICE item response options be standardized as five-option responses.

Time constraints required a 26% reduction in the total number of ABLE and AVOICE items. The goal was to decrease items on a scale-by-scale basis, while preserving the basic content of each scale. The strategy adopted to accomplish this for each scale was to:

1. Sort items into content categories.
2. Rank order items within category, based on item-scale correlations.
3. Drop items in each category with the lowest item-scale correlations until desired number of items for that scale had been deleted.

The total number of AVOICE items was decreased from 309 to 214. Thirty-eight of these 214 are items on the Work Environment Preference scales. It was decided to take this whole section out of the AVOICE booklet and include it in one of the criterion measure booklets, where a bit more administration time was available.

A decision was also made to delete the Agriculture scale, the six single-item Holland scales, and the eight Expressed Interest items. Reductions made on the remaining AVOICE scales were accomplished using the same strategy as that for the ABLE, decreasing scale lengths while preserving heterogeneity.

Description of the Trial Battery and Summary Comments

Table II.55 shows the final array of tests for the Trial Battery. These are the measures that were the product of the revisions just described.

The Trial Battery was designed to be administered in a period of 4 hours and was used during the Concurrent Validation phase of Project A, in which data collection began in FY85. Data collected in that phase will allow the first look at the validity of Trial Battery measures against job performance criteria. It will also allow replication, on a much larger sample, of many of the analyses described in the development and testing sections presented in this report.

Table II.55

Description of Measures in the Trial Battery

COGNITIVE PAPER-AND-PENCIL TESTS	Number of Items	Time Limit (minutes)
Reasoning Test	30	12
Object Rotation Test	90	7.5
Orientation Test	24	10
Maze Test	24	5.5
Map Test	20	12
Assembling Objects Test	32	16

COMPUTER-ADMINISTERED TESTS	Number of Items	Approximate Time
Demographics	2	4
Reaction Time 1	15	2
Reaction Time 2	30	3
Memory Test	36	7
Target Tracking Test 1	18	8
Perceptual Speed and Accuracy Test	36	6
Target Tracking Test 2	18	7
Number Memory Test	28	10
Cannon Shoot Test	36	7
Target Identification Test	36	4
Target Shoot Test	30	5

NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES	Number of Items	Approximate Time
Assessment of Background and Life Experiences (ABLE)	209	35
Army Vocational Interest Career Examination (AVOICE)	176	20

Part III

CRITERION DEVELOPMENT

The sections included in Part III describe the development work for each major criterion measure, the revisions made on the basis of pilot data, the procedure used for the major criterion field tests, the results of the field tests, and the final revisions made on the basis of those results for use in Concurrent Validation.

Section 1

INTRODUCTION TO CRITERION DEVELOPMENT

The overall goals of training and job performance (i.e., criterion) measurement in Project A are to define the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor. The specific measures are to be used as criteria against which to validate selection and classification tests and are not intended to serve as operational performance appraisal methods. That is, the research participants will be informed that these performance measures are not part of their formal performance appraisal and will not be entered into their personnel file. However, this does not mean that the Project A measures cannot be modified to serve as useful operational performance appraisals in future contexts; we certainly hope that they can be.

The general procedure for criterion development in Project A followed a basic cycle of a comprehensive literature review, conceptual development, scale construction, pilot testing, scale revision, field testing, and proponent (management) review. The specific measurement goals are to:

- Make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency.
- Compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multitrait, multimethod approach).
- Develop rating scale measures of performance factors that are common to all first-tour enlisted MOS (Army-wide measures).
- Develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance.

Given these intentions, the criterion development effort focused on three major methods: hands-on job sample tests, multiple-choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in developing the rating methods.

Modeling Performance

The development efforts to be described were guided by a particular "theory" of performance. The intent was to proceed through an almost continual process of data collection, expert review, and model/theory revision. Two iterations of this procedure are described in this report. The basic outline of the initial model is described in the following paragraphs.

Multidimensionality

A first basic point that should generate no argument is that job performance is multidimensional. There is not one attribute, one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is perhaps a bit more arguable to go on from there and assert that job performance is a construct (which implies a "theory" of performance), and is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization. For example, a manager could make contributions to organizational goals by working out congruent short-term goals for his subordinates, and thereby guiding them in the right direction, or by praising them for a job well done, and thereby increasing subsequent effort levels. Each of these activities probably requires different knowledges and skills, which are in turn most likely a function of different abilities.

Consequently, for any particular job, one fundamental task of performance measurement is to describe the basic factors that comprise performance. That is, how many such factors are there and what is their basic nature?

Two General Factors

For the population of entry-level enlisted positions in the Army, we postulated that there are two major types of job performance factors. The first is composed of performance components that are specific to a particular job. That is, measures of such components would reflect specific technical competence or specific job behaviors that are not required for other jobs. For example, typing correspondence would be a performance component for an administrative clerk (MOS 71L) but not for a tank crewman (MOS 19E). We have called such components "MOS-specific" criterion factors.

The second type of performance includes components that are defined and measured in the same way for every job. These have been referred to as "Army-wide" criterion factors. Examples might be performance on the common tasks for which every soldier is responsible or proficiency in peer leadership.

For the MOS-specific components we anticipated that there would be a relatively small number of distinguishable subfactors (or constructs) of technical performance that would be a function of different abilities or skills and that would be reflected by different task content. The criterion construction procedures were designed to identify technical performance factors that reflected different task content.

The Army-wide concept incorporates the basic notion that total performance is much more than task or technical proficiency. It might include such things as contribution to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity. A much more detailed description of the initial working model for the Army-wide segment of performance can be found in Borman, Motowidlo, Rose, and Hanser (1987a).

In sum, the working model of total performance with which the project began viewed performance as multidimensional within the two broad categories of factors or constructs. The job analysis and criterion construction methods were designed to "discover" the content of these factors via an exhaustive description of the total performance domain, several iterations of data collections, and the use of multiple methods for identifying basic performance factors.

Factors Versus a Composite

Saying that performance is multidimensional does not preclude using just one index of an individual's contributions to make a specific personnel decision (e.g., select/not select, promote/not promote). As argued by Schmidt and Kaplan (1971) some years ago, it seems quite reasonable for the organization to scale the importance of each major performance factor relative to a particular personnel decision that must be made, and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision. That is, the way in which performance information is weighted is a value judgment on the organization's part. The determination of the specific combinational rules (e.g., simple sum, weighted sum, non-linear combination) that best reflect what the organization is trying to accomplish is in large measure a research question.

In sum, it makes sense to assert that performance in a particular job is made up of several relatively independent components and then ask how each component relates to some continuum of overall utility. It is quite possible for people with different strengths and weaknesses on the performance factors to have very similar overall utility for the organization.

A Structural Model

If performance is characterized in the above manner, then a more formal way to model performance is to think in terms of its latent structure. The usual common factor model of the latent structure is open to criticism because all of the criterion (i.e., performance) measures may not be at the same level of explanation or they may be so qualitatively different that putting them into the same correlation matrix does not seem appropriate. For example, combining the dichotomous variable stay vs. leave (voluntarily) with a hands-on job performance test score seems like a strange thing to do. Also, two criteria may not be functionally independent. One might be a cause of another; for example, individual differences in training performance may be a cause of individual differences in job performance.

Considerations such as the above have led some people to propose structural equation modeling as a way to portray the criterion space and the associated predictor space more meaningfully (e.g., Bentler, 1980; James, Muliak, & Brett, 1982).

From this perspective, the aims of criterion analysis are to use all available evidence, theory, and professional judgment to (a) identify the variables that are necessary and sufficient to explain the phenomena of

interest, and (b) specify the nature of the relationships between pairs of variables in terms of whether they 1) are correlated because one is a cause of another, 2) are correlated because both are manifestations of the same latent property, or 3) are independent. The more explicitly the causal directions and the predicted magnitude of the associations can be specified, the greater the potential power of the model. That is, it more clearly outlines the kinds of data to be collected and the kinds of analyses to be done, and it provides a much more explicit framework within which to interpret empirical results.

Within the structural equation framework there are two general kinds of models, one dealing with manifest variables (operational measures) and one with latent variables (constructs). The most thorough portrayal of a domain presumably involves both; certainly in Project A we have assumed that it does. The proposal and research plans have dealt explicitly with criterion constructs and criterion measures. What we really want to model, in terms of identifying the necessary and sufficient variables and their causal interrelationships, are the more "fundamental" underlying constructs. What we in fact will have are operational measures that represent the constructs.

A few points--some general, some specific--should be made about this view. First, it is true that we simply know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter. For personnel psychologists it is almost second nature to talk about predictors in terms of constructs. However, investigation of job performance constructs seems to have been limited to those few studies dealing with synthetic validity and those using the critical incidents format to develop performance factors. Relative to the latter, the occupations receiving the most attention have been managers, nurses, firefighters, police officers, and perhaps college professors (cf. Landy & Farr, 1983). Relatively little attention has been given to conceptualizing performance in clerical, technical, or skilled jobs.

Second, the usual textbook illustration of a latent structural equation model (e.g., James et al., 1982) shows each latent variable being represented by one or more manifest operational measures. However, in our situation, just as it is easy to think of examples where a predictor test score could be a function of more than one latent variable (e.g., the score on computerized two-hand tracking apparatus could be a function of several latent psychomotor "factors"), the same will be true of criterion measures. Most of them will not be unidimensional.

Third, we would be hard-pressed to defend placing the criterion variables on some continuum from immediate to intermediate to ultimate as a means for portraying their relative importance or functional interrelationships. For example, although there are good reasons for developing hands-on performance measures, we would not be willing to defend hands-on performance scores as the "most ultimate" measure.

Unit vs. Individual Performance

Finally, people do not usually work alone. Individuals are members of work groups or units and it is the unit's performance that frequently is the most central concern of the organization. However, determining the individual's contribution to the unit's score is not a simple problem. Further, variation in unit performance is most likely a function of a number of factors besides the "true" level of performance of each individual. The quality of leadership, weather conditions, or the availability of spare parts are examples of such additional sources of variation in unit performance. In addition, there probably are, somewhere, interactions between the characteristics of individuals and the characteristics of units or situations.

For two major reasons, Project A has not incorporated unit effectiveness in its model of performance. First, the project is focused on the development of a new selection/classification system for entry-level personnel and is concerned with improving personnel decisions about individuals, and not units. The task is to maximize the average payoff per individual selected. The Army cannot make differential assignments based on differences in weather, leadership climate, and so on. Future conditions cannot be predicted with any certainty and during a tour of duty an individual will serve in several different units. Consequently, personnel assignments must be optimal when averaged across all such conditions. By design, they should not take situational interactions into account. Operationally, these sources of variation must be dealt with by other means (e.g., leadership training). However, in a research context, Project A is attempting to investigate these additional sources of variation via a systematic description of the work environment. The Army Work Environment Questionnaire (Olsen, Borman, Robertson, & Rose, 1984) asks job incumbents to describe 14 dimensions of their job situation using a 44-item questionnaire.

The second major reason for not using unit performance as a criterion is the prohibitive cost. It simply was not possible to develop reliable and valid field exercises for assessing unit performance in a representative sample of MOS within a reasonable time frame. In isolated instances it might be possible to take advantage of regularly scheduled exercises or use existing performance records that a particular unit (e.g., maintenance depot) might keep. However, it proved not possible to obtain such data in any systematic way. Even if it could be done, it would not be easy to establish the correspondence between individual performance and unit effectiveness.

What we have chosen to do is to try to identify the factors, or means, by which individuals contribute to unit performance and to assess individual performance on those factors via rating methods. At the same time we are attempting to determine how much of the variance in individual performance is accounted for by the situational characteristics assessed by the Army Work Environment Questionnaire.

Plan for Part III

With the above discussion as background, we now turn to describing the development steps for each of the major performance measures. Once the initial array of criteria has been described, the procedures and results of

the full-scale criterion field tests will be summarized. Finally, the revisions of the performance measures that were made on the basis of the field test results and the reviews by the Army proponents will be outlined.

At that point the final array of performance measures to be used in the Concurrent Validation will have been established. Again, it is the intent of the project that, within the limits of its money and time, this array will describe the entire performance space and utilize all feasible means of assessing performance.

Section 2

DEVELOPMENT OF MEASURES OF TRAINING SUCCESS[1]

The purpose of this section is to describe the development of the achievement tests used to measure training success for the 19 MOS in the Project A sample. The tests were developed in three batches (A, B, and Z). Batch A and Batch B are defined as described in Part I and Batch Z is composed of the remaining 10 MOS. The complete lineup is shown in Table III.1. The methods used to develop materials for the three batches were essentially the same except for modification, as described below, to take advantage of experience with prior batches. All measures were pilot tested with a sample of 50 soldiers at the conclusion of Advanced Individual Training (AIT). Batches A and B were also field tested with job incumbents; the tests in Batch Z were not.

The Measurement Model

The Construct of Training Success

The Project A Research Plan defines training success in terms of the individual trainee's achievement. The original Statement of Work used the term in a similar way, that is, to refer to specific training measures taken on soldiers in the course of training, such as those included in the TRADOC Educational Data System (TREDS) and the Automated Instruction Management System (AIMS). Many of these measures include both hands-on and cognitive instruments.

Training success encompasses both the outcomes of formal training and organizational socialization. Organizational socialization is defined as the way in which soldiers accommodate to their role as soldiers and "learn the ropes," such as the attitudes, standards, and patterns of behavior expected of soldiers in general and of soldiers in an assigned MOS. Organizational socialization is achieved through formal training, of course, but it also is developed through a variety of tactics outside of the regular classroom, including role modeling, drill, stressful experiences, NCO leadership, and other practices designed to produce appropriate military attitudes.

A wide variety of potentially useful measures either are available or could be created to assess three major elements of training success: (a) the knowledge component, (b) the hands-on, or performance, component, and (c) the organizational socialization component. The achievement tests, described

---

[1]This section is based primarily on an ARI Technical Report 757, Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS, by Robert H. Davis, Gregory A. Davis, John N. Joyner, and Maria Veronica de Vera, and a supplementary ARI Research Note, in preparation, which contains the report appendixes.

below, are designed to measure the knowledge component of formal training experience, specifically AIT. This component of training success includes two types of knowledge: (a) knowledge about the job, as taught in AIT, and (b) knowledge about a wide range of "common skills" that cut across all MOS and that all soldiers are expected to know.

Table III.1.

Military Occupational Specialties (MOS) Included in Batches A, B, and Z

Batch A		Batch B	
13B	Cannon Crewman	11B	Infantryman
64C	Motor Transport Operator	19E	Armor Crewman
71L	Administrative Specialist	31C	Radio Teletype Operator
95B	Military Police	63B	Light Wheel Vehicle Mechanic
		91A	Medical Specialist

Batch Z[a]	
12B	Combat Engineer
16S	MANPADS Crewman
27E	TOW/Dragon Repairer
51B	Carpentry/Masonry Specialist
54E	NBC Specialist
55B	Ammunition Specialist
67N	Utility Helicopter Repairer
76W	Petroleum Supply Specialist
76Y	Unit Supply Specialist
94B	Food Service Specialist
19K	M1 Abrams Armor Crewman[b]

[a]Not field tested with job incumbents.

[b]Developed for longitudinal validation; not included in the Concurrent Validation.

Relationship Between Training Content and Job Content

Within the military, there is a very close relationship between training content and tasks performed on the job. Skill Level 1 soldiers within any given MOS may work at quite different jobs--that is, jobs that emphasize different skills--but it is almost always the case that the knowledges and skills necessary for the performance of a job at Skill Level 1 are taught in AIT. As a matter of doctrine training must be job-related, and in developing

training objectives and materials Army personnel make every effort to ensure that they are job-related. As a result, if a content-valid test is created based on curricular materials alone, one can assume that most of the items will be job-related. While school curricula do sometimes include topics or tasks that are unrelated to the job, this is the exception rather than the rule.

Classes of Items

It seems clear that some trainees learn important job skills that are not taught in the schools. As a result of extracurricular activities, outside study, generalization, or all three, a trainee may develop some job skills in the school setting that are not taught as part of the curriculum. From the perspective of criterion development, one might hypothesize that the exceptional--that is, most successful--trainee is one who goes beyond the formal curriculum and learns such knowledge and skills. In education, the jargon term for this phenomenon is incidental learning.

Similarly, military training performance is predictive of later military job performance because (a) training performance reflects general learning ability (and hence identifies who will acquire knowledge on the job), (b) the information acquired in training is itself a significant factor in job performance, or, more likely, (c) both. Accordingly, we constructed two subsets of test items--one reflecting training content and the other job content. Where a sufficient number of test items could be developed for both classes, scores on the two types of items may shed light on the relationships between success in training and success on the job. That is, is the correlation between training performance and job performance a function of achievement during training, incidental learning during training, or individual differences in basic abilities that are present before training starts?

The Meaning of Content Validity

Although definitions of content validity differ, the literature stresses three critical components: clarity of the content domain, representativeness of content, and relevance of content.

By domain clarity, we mean that the content domain should be defined unambiguously. Essentially, this means that the boundaries of the domain from which test content is drawn should be clearly defined and understood. Experts may differ as to the appropriate boundaries, and the differences may become matters of disagreement in the course of test construction. But once the boundaries are defined, experts should be able to agree as to whether or not items fall inside or outside of those boundaries. At the outset, we operationally defined the content domain in the following way. For training content, the domain was described by Programs of Instruction (POIs) lesson plans, technical publications, Soldier's Manuals, and the Common Task Manual. For the job, content was specified by Army Occupational Surveys (AOSPs), technical publications, Soldier's Manuals, and the Common Task Manual.

The issue of content representativeness refers to the question of whether the domain has been adequately sampled. Operationally, establishing content representativeness involves a strategy for arriving at item budgets, that is, budgets for the number of items on a test to cover different parts of the domain. When people disagree about such matters, the question is normally resolved in terms of the level of expertise of those making the decision. In the case of the Project A achievement tests, the procedure for developing item budgets as representative samples of training and job content was determined by test construction experts (i.e., project staff) but the content of the budget was evaluated by subject matter experts (job incumbents and trainers--SME).

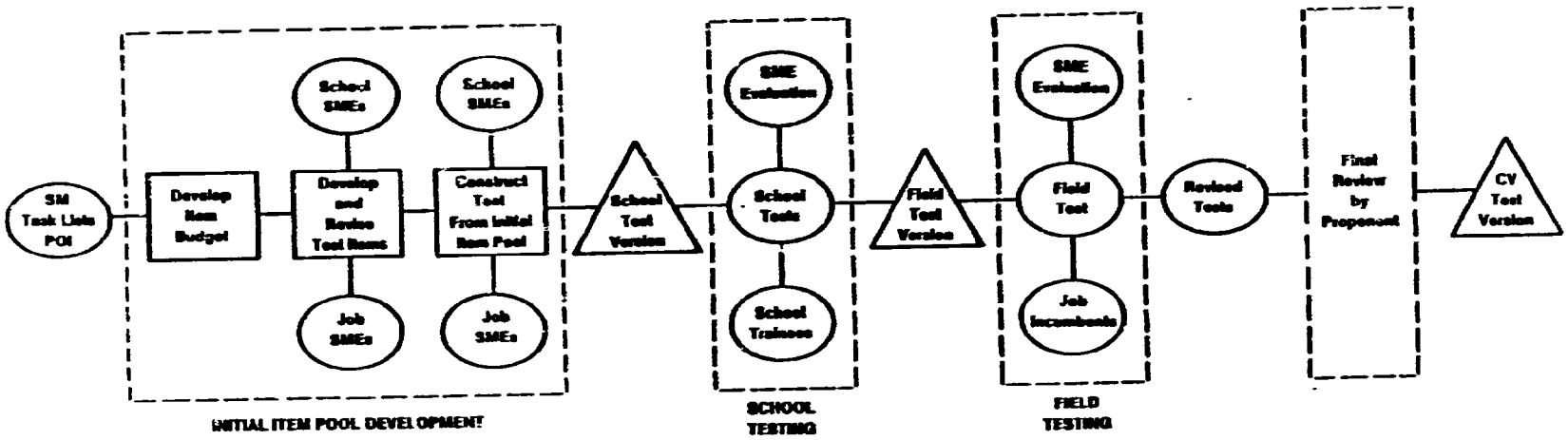
It is the SME evaluations that provide the judgments of content relevance. In the narrowest sense, we may simply ask whether or not specific items are relevant to the two facets of the content domain already identified, that is, training achievement and job performance. The question may also be extended to explore the relevance of items under different circumstances or scenarios (e.g., peacetime, readiness, and combat). How this was accomplished is described later in this section.

Test Development Procedure

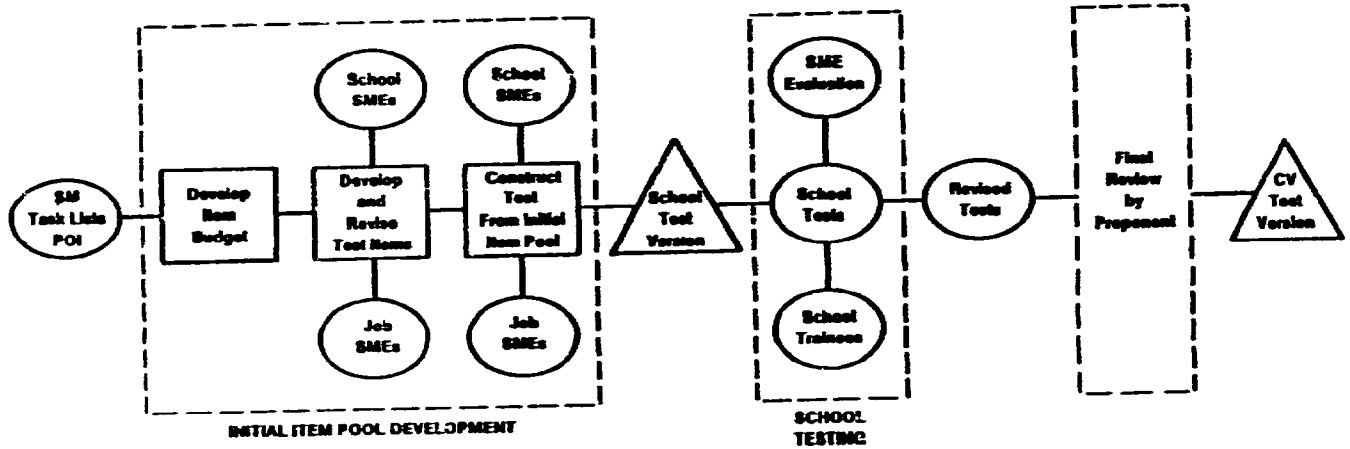
The principal steps in the construction of the training achievement tests were as follows:

1. Development of the initial item pool
2. Review by job incumbents
3. Review by school trainers
4. Administration to trainees
5. Preparation of the item pools for administration to job incumbents
6. Administration to job incumbents (Field Tests, Batches A and B only)
7. Review by TRADOC proponent agencies
8. Preparation of the item pools for administration to job incumbents in the Concurrent Validation.

Each of these steps is described briefly below. Although each test went through many revisions during this process, it is perhaps easiest to think in terms of the three versions shown in Figure III.1: (a) the initial item pool, (b) the version administered to incumbents in the field test, and (c) the version administered to incumbents in the Concurrent Validation. Figure III.1 also summarizes the differences in developmental procedures between Batches A/B and Batch Z.



BATCHES A & B DEVELOPMENT PROCESS



BATCH Z DEVELOPMENT PROCESS

Figure III.1. Development process for tests of achievement in training.

Procedures were modified somewhat on the basis of experience as the tests were developed. For example, all item pools were reviewed by groups of SMEs. However, after the first few group reviews, it was apparent that a preliminary review by one SME for accuracy, correct use of technical language, currency, and appropriateness could greatly facilitate the group review. Accordingly, this step was introduced in the process. Similarly, as reviews progressed, a concern for racial and gender balance within SME groups led to the development and implementation of guidelines for taking racial and gender aspects into account in assigning SMEs to review groups. A second informal review was scheduled for all items that had been reviewed before the implementation of the guidelines.

Development of the Initial Item Pool

Development of the item pool for each MOS involved three elements: refinement of the AOSP task list, calculation of a test item budget, and item drafting itself.

Refinement of the AOSP Task List. The Army Occupational Survey Program uses a questionnaire checklist of several hundred items to survey job incumbents about specific job tasks that they do or do not perform. Related tasks are combined into duty areas on the basis of expert judgment by the job proponent. The number of duty areas in each of the 19 MOS included in the present study ranged from 15 to 23. One of the key statistics reported as part of the AOSP is the percentage of soldiers at different skill levels who are performing the task activity. As described below, this statistic was used to prepare a test item budget prior to drafting items.

Before the AOSP reports were used, however, several actions were taken to refine the item information. For Batches A and B, the AOSP task lists were edited as follows:

- Ninety-nine percent confidence intervals were computed for the mean percentage performing all tasks. Tasks with a very low percent performing (equal to or less than the lower boundary of the confidence interval) were deleted from consideration.

- The remaining task statements were then reviewed by four to six SMEs (experienced NCOs in that MOS) to:

  -- Delete AOSP statements for any of three reasons: They were no longer part of the job due to changes in doctrine or equipment; they were not really tasks, and should not have been included in the AOSP listing (e.g., administrative labels that were misconstrued as tasks); or they were sets of tasks (i.e., they contained only individual tasks that were already in the domain).

  -- Confirm the project staff's grouping of AOSP task statements into the task specified in the Soldier's Manual.

Calculation of Item Budgets. To ensure that the content of item pools were representative of tasks performed and that it covered the entire MOS rather than aspects easiest to write items about, an item budget was drafted

based on the duty areas into which the AOSP survey is divided. As noted above, the number of duty areas in the 19 AOSP surveys analyzed ranges from 15 to 23. It was expected that during tryout, revision, and field testing, items would be eliminated from the pool because of faulty construction or lack of discriminatory or predictive power. To allow for item attrition, the initial target was 225 draft items for each MOS, even though the final version of the test was expected to be closer to 150 items. Survey data on percentage performing were used in building the budget as described below.

Step 1--Determine the match between AOSP duty areas and training objectives. A matrix was prepared to display the duty areas of the AOSP versus the subdivisions of the POI, each of which covers a number of training objectives. Three outcomes were possible: (a) some duty areas matched Army training lessons completely; (b) some duty areas did not match any training lesson; (c) some training lessons did not match any duty area. The majority of the item budget, 200 items, was allocated to the first two categories.

Step 2--Distribute the first 200 items. To determine a target number of items for each duty area, the 200 items budgeted to the job performance domain were distributed across the duty areas in proportion to the mean percentage of incumbents reported by the AOSP as performing the tasks that composed the duty area.

Within each of the AOSP duty areas, items were budgeted in proportion to how much they were emphasized in training: The greater the overlap between the AOSP tasks (within a duty area) and the training objectives (within the POI), the more items were written to represent job/training content.

The remaining items (out of the original 200) were assigned to job-only content. For example, if 20 items were assigned to a duty area and the duty area had a total of eight tasks, six of which matched objectives on the POI, then 15 (6/8 x 20) training/job items and 5 job-only items (15 + 5 = 20) would be written for the task.

Step 3--Distribute the remaining items (25 or fewer). The remainder of the item budget for a given MOS was reserved for items not related to any area of the AOSP task list, but covering training content as defined by the POI. Within the portion of the training performance domain that did not match any portion of the job performance domain, the allocation of test items was based on the amount of training time devoted to particular content.

Drafting of Items. After item budgets were established, written materials dealing with job training activities were examined for information that could be transformed into multiple-choice test items. Four sources were used: the AOSP task lists, training materials (POIs, lesson plans, lesson guides, etc.), technical publications (Army regulations, Technical Manuals, Field Manuals, etc.), and the Soldier's Manual for each MOS. The Soldier's Manual is a description of the tasks that each MOS holder is to have mastered to be considered qualified at a given skill level.

Using these various documents and the item budgets, multiple-choice items were drafted for all MOS. The item-writing group included the research staff, a retired Army lieutenant colonel, and other contract item-writers.

Review by Job Incumbents

To prepare the item pool for review by a panel of job incumbents, the pool was first reviewed by one subject matter expert, usually a senior officer, who purged the item pool of its more glaring faults. The items were then reviewed by job incumbents, which required a series of site visits. On each visit, incumbents reviewed the items for technical accuracy and appropriate vocabulary, and rated item content for importance and relevance to Skill Level 1 soldiers.

The entire line-up of SMEs for the various review stages is shown in Table III.2. Analysis indicated that minority groups were adequately represented in the SME samples. For example, Table III.3 shows the expected and observed frequencies of SMEs by race, compared to the percentage of active duty soldiers in the Army in each racial category.

Table III.2

Number of Subject Matter Experts Participating in Training Achievement Test Reviews, and Locations of Reviews

	Refinement of Task List		Job Incumbent Review		School Trainer Review	
MOS	No. of SME	Location	No. of SME	Location	No. of SME	Location
Batch A						
13B	5	Fort Ord	7	Fort Ord	7	Fort Sill
64C	4	Fort Ord	4	Fort Ord	6	Fort Dix
71L	4	Fort Ord	6	Fort Ord	6	Fort Jackson
95B	5	Fort Ord	8	Fort Sill/Dix	10	Fort McClellan
Batch B						
11B	5	Fort Ord	5	Fort Ord	6	Fort Benning
19E	5	Fort H. Liggett	5	Fort H. Liggett	6	Fort Knox
31C	5	Fort Ord	5	Fort Ord	6	Fort Gordon
63B	5	Fort Ord	5	Fort Ord	6	Fort Dix
91A	5	Fort Ord	5	Fort Ord	6	Fort Sam Houston
Batch Z						
12B	5	Fort Ord	6	Fort Lewis	6	Fort L. Wood
16S	5	Fort Ord	5	Fort Lewis	6	Fort Bliss
27E	4	Fort Ord	6	Fort Lewis	6	Redstone Arsenal
51B	4	Fort Ord	4	Fort Lewis	4	Fort L. Wood
54E	5	Fort Ord	5	Fort Lewis	5	Fort McClellan
55B	5	Fort Ord	6	Fort Lewis	5	Redstone Arsenal
67N	5	Fort Ord	6	Fort Lewis	6	Fort Rucker
76W	5	Fort Ord	6	Fort Ord	6	Fort Lee
76Y	5	Fort Ord	6	Fort Ord	6	Fort Lee
94B	5	Fort Ord	8	Fort Sill/Dix	10	Fort McClellan

19K			7	Fort Knox	10	Fort Knox

Table III.3

Distribution of Soldiers in Four Race Categories, Army-Wide and Among Subject Matter Expert Reviewers for Training Achievement Tests

Race	Army-Wide Percent Active Duty[a]	Expected Frequency in SME Sample	Observed Frequency in SME Sample
Caucasian	61.8	142.8	121
Black	30.5	70.4	74
Hispanic	4.0	9.2	33
Other	3.7	8.6	3
			231

[a]Source: Dr. Mark J. Eitelberg, personal communication.

Item Quality. To establish the technical accuracy and appropriateness of the draft items, job incumbents were asked:

- Would the item be clear to someone taking the test?
- Is the option indicated really the correct answer?
- Is there more than one correct option?
- Are the distractors realistic and believable?
- Is each technical term commonly used and easily understood?
- Are there other commonly used terms that should be included to make the questions clearer?

Items were then revised in accordance with the responses from incumbents.

Importance Ratings. Incumbents were next asked to rate the importance of each item in three different contexts: combat (Scenario 1), combat readiness (Scenario 2), and garrison duty (Scenario 3). The scenarios used to describe these three contexts are shown in Figure III.2. A 5-point scale ranging from "Very Important" (5) to "Of Little Importance" (1) was used to collect importance ratings.

Table III.4 shows the mean item importance under each of the three scenarios. Also shown are the interrater reliabilities for the pooled ratings. The "item pool" is defined as those items that were taken to incumbents for the importance ratings--that is, the first version of the test.

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 increasing international tension. Hostilities have not broken out. The Corps maneuver terrain is rugged, hilly, and wooded, and weather is expected to be wet and cold. 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 stationed on a post in the Continental United States. The unit has personnel and equipment sufficient to make it mission capable for training and evaluation and installation support missions. 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 III.2. Alternative scenarios used for judging importance of tasks and items for training achievement tests.

Table III.4

Mean Item Importance Ratings by Job Incumbents for Three Scenarios (Initial Item Pool for Training Achievement Tests)

MOS	Number		Mean Item Importance Rating[a]			Interrater Reliability		
	Raters	Items	Scenario 1	Scenario 2	Scenario 3	Scenario 1	Scenario 2	Scenario 3
Batch A								
13B	7	259	3.9	3.8	3.8	.33	.34	.17
64C	4	216	4.3	4.8	4.9	.74	.75	.67
71L	4	119	2.5	2.9	3.2	.74	.74	.48
95B	5	221	2.9	3.1	2.9	.88	.74	.85
Batch B								
11B	5	200	3.9	3.8	3.7	.73	.69	.70
19E	5	219	3.5	3.5	3.7	.66	.66	.70
31C	5	200	4.0	3.9	3.8	.53	.47	.45
63B	5	282	2.1	2.6	3.1	.69	.67	.53
91A/B	5	306	2.1	2.4	2.9	.74	.75	.71
Batch Z								
12B	6	229	2.6	3.0	3.9	.90	.89	.81
16S	5	208	4.1	4.1	4.8	.88	.87	.75
27E	5	233	3.4	3.7	3.8	.90	.90	.89
51B	4	228	3.5	3.6	4.6	.81	.52	.00
54E	7	162	3.8	3.8	3.3	.73	.70	.72
55B	5	236	3.3	3.5	3.6	.64	.73	.70
67N	6	221	3.2	3.7	3.9	.88	.92	.91
76W	5	214	3.9	4.4	4.8	.63	.34	.26
76Y	5	198	2.1	2.3	4.4	.90	.88	.43
94B	5	213	3.2	3.2	3.3	.64	.63	.34

19K	6	250	3.3	3.3	3.1	.92	.92	.90

[a] Rated on a 5-point scale from "Of Little Importance" (1) to "Very Important" (5).

Two points about the ratings data are worth highlighting. First, a relatively large percentage of the items were rated as very important. Second, when importance ratings under the three scenarios are compared, a lower percentage of items were rated as very important when using the combat scenario than when using the garrison scenario (33.1 vs. 43.1%) and a higher percentage of items were considered to be of little importance (22.8 vs. 11.2%). A 2 x 2 contingency table comparing item frequencies (Garrison & Combat vs. Rating 1 & 5) yields a chi-square of 224.09, $p = .004$.

Mean interrater reliabilities were reasonably high for the combat and combat readiness scenarios, .74 and .71 respectively, but somewhat lower for the garrison scenario, .60.

Relevance Ratings. To establish the relevance of the draft test items, incumbents were asked, "Do Skill Level 1 personnel in this MOS need to use this knowledge on the job?" It was recognized that an MOS comprises many jobs, or duty positions, and that incumbents in different duty positions might disagree about item relevance because they defined the job differently. The procedure followed was to favor inclusion. If any one respondent in the group asserted that the knowledge was required for job performance, then the item was flagged as job-relevant.

A by-product of the total review was the identification of tasks or duty areas that were not included in the AOSP data but were part of incumbents' responsibilities or that were included in the AOSP report but were no longer part of the MOS. Some items were drafted for the former category after the site visit. To maintain the relative distribution of items across duty areas, additional items were also drafted to replace discarded items.

Review by School Trainers

The item pool was also reviewed by trainers at one of the training sites for the MOS. As with the review by job incumbents, the trainers reviewed items for technical accuracy and appropriate vocabulary, and rated item content for importance and relevance. It was during such site visits that pilot tests were conducted with trainees, as described in the next subsection.

To obtain a measure of item importance from the trainers' point of view, SMEs were given the following instructions:

> Look at each of the test questions and ask yourself how important it is that a trainee in the course learn the knowledge represented by this question.

Trainers used the same scale as incumbents to rate item importance. Table III.5 shows the mean rating for items in the item pool. The table also contains interrater reliabilities for all MOS.

Overall, trainers tended to rate items significantly higher than did incumbents. Mean importance rating by trainers for the pool was 4.18 (median = 4.03) while the mean of the means across scenarios for incumbents on the initial item pool was 3.52 (median = 3.58) (Wilcoxon $Z = 3.38$, $p = .001$).

This same trend appears in the proportions of items rated very important and of little importance. Trainers rated a mean of 54.4% of the items in the item pool as very important, compared with incumbents who gave a rating of very important to 33.1% of the items on the combat scenario and 43.1% of the items on the garrison scenario. Incumbents rated 22.8% of the items as being of little importance on the combat scenario and 11.2% on the garrison scenario; trainers, however, rated only 4.1% of the items as of little importance.

Table III.5

Mean Item Importance of Ratings by Trainers (Initial Item Pool for Training Achievement Tests)

MOS	Number Raters	Number Items	Mean Rating[a]	Interrater Reliability
Batch A				
13B	6	297	4.4	.72
64C	7	215	4.2	.78
71L	5	122	3.8	.95
95B	5	122	4.7	.50
Batch B				
11B	7	200	3.8	.52
19E	6	214	4.0	.64
31C	6	192	3.7	.69
63B	6	238	3.3	.61
91A/B	6	299	3.9	.81
Batch Z				
12B	6	221	4.0	.87
16S	5	208	4.1	.61
27E	6	219	4.0	.73
51B	4	218	4.8	.57
54E	6	220	3.9	.66
55B	6	227	4.7	.32
67N	4	215	4.5	.18
76W	3	214	4.7	.00
76Y	6	132	5.0	.32
94B	6	200	3.8	.68

19K	6	202	4.1	.75

[a]Rated on a five-point scale from "Of Little Importance" (1) to "Very Important" (5).

Mean trainer interrater reliability across MOS was .58 (median = .62). This compares with a mean of .67 for incumbents (median = .70) across all three scenarios.

To establish the relevance of the draft test items to training, trainers were asked the following:

> Can trainees be expected to have the knowledge represented in the items as a result of training?

As with job relevance, the procedure favored inclusion. If any one of the trainers responded affirmatively, then the item was flagged as training-relevant. At this point, relevance data were available for all items with respect to the job alone (from SME incumbents), training alone (from SME trainers), or both. Where the two judgments overlapped, items were considered relevant to both job and training. Items added in subsequent revisions after these judgments were made were not rated for relevance.

Table III.6 is based on relevance data obtained from job incumbents and from trainers and shows the distribution of the various classes of items for each MOS on the pilot test administered to trainees in the schools. The Not Rated category consists of items added to the pool after relevance ratings had been collected. Percentages have been computed for the Job-Only, Training-Only, and Job-and-Training categories, using the total of these three as the divisor.

As would be expected, many more items were rated as Job-and-Training (2,843 or 75.5%) than as either Job Only (676 or 17.9%) or Training Only (249 or 6.6%). Also, there are substantial differences in the range of items in these three categories. Of particular interest is the comparison between Job Only (range = 0-78) and Training Only (range = 0-140). The large range for Training Only is accounted for solely by MOS 91A; without this one MOS the range would be 0-91. MOS 91A is the designation for Medical Specialists, and incumbents appear to believe that many items which trainers consider relevant are not relevant to the job.

Given the doctrinal emphasis on relating training to the job, it is not surprising that (with the exception of MOS 91A) few items were rated as Training Only, despite the effort on the part of the item writers to create such items.

Administration to Trainees

After review by job incumbents and trainers, test items were administered to groups of trainees in their last week of training. A sample of trainees was also interviewed after the test to obtain information about the clarity and comprehensibility of the items. Specific questions included the following:

- Did you have any difficulty understanding the question? Were there any words or phrases which were difficult to understand?

Table III.6

Number and Percentage of Items Rated Relevant to Job and Training (Initial Item Pool for Training Achievement Tests)

MOS	Job Only N	Job Only %	Training Only N	Training Only %	Job and Training N	Job and Training %	Not Relevant N	Not Rated[a] N
Batch A								
13B	70	41.4	5	3.0	94	55.6	6	62
64C	78	36.8	0	0.0	134	63.2	0	16
71L	42	34.4	4	3.3	76	62.3	0	0
95B	64	31.5	8	3.9	131	64.5	11	20
Batch B								
11B	68	39.5	14	8.1	90	52.3	21	25
19E	32	16.2	9	45.7	156	79.2	2	5
31C	47	26.3	15	8.4	117	65.4	5	8
63B	48	23.0	8	3.8	153	73.2	2	4
91A	0	0.0	140	54.9	115	45.1	5	5
Batch Z								
12B	7	3.4	0	0.0	197	96.6	0	23
16S	11	5.4	0	0.0	191	94.6	0	6
27E	1	0.5	19	9.3	185	90.2	0	15
51B	0	0.0	0	0.0	202	100.0	0	16
54E	0	0.0	1	0.5	207	99.5	0	15
55B	0	0.0	5	2.4	206	97.6	0	16
67N	1	0.5	0	0.0	208	99.5	0	8
76W	68	31.8	12	5.6	134	62.6	0	0
76Y	78	39.2	0	0.0	121	60.8	0	1
94B	61	31.1	9	4.6	126	64.3	8	2
Total	676	17.9	249	6.6	2,843	75.5	60	247

[a]Items added to the pool after relevance ratings had been collected.

- Do you agree with the correct answer? Is there a better way to state the answer?
- (For items derived from tasks performed in training) Is it necessary to know the answer to this question to perform the task in training?
- (For items derived from tasks performed in training) Is the item a fair measure of a soldier's ability to perform the task?

The results of this test administration to trainees are shown in Table III.7. All results shown are based on items relevant to training, that is, Job-and-Training and Training-Only items. Items relevant only to the job are not included.

Table III.7

Results From Training Achievement Tests Administered to Trainees

MOS	Number of Subjects	Number of Items	Mean Number Correct	SD	Range	Alpha	Mean Percent Correct
Batch A							
13B	50	104	54.4	10.2	44	.81	52.3
64C	50	130	69.0	13.7	60	.87	53.1
71L	70	71	39.3	7.4	31	.79	55.3
95B	50	105	69.6	10.6	46	.85	66.2
Batch B							
11B	51	111	53.4	13.7	74	.91	48.1
19E	50	169	102.0	17.4	86	.92	60.4
31C	49	135	78.3	14.6	71	.90	58.0
63B	60	162	67.1	19.8	78	.92	41.4
91A	49	255	128.1	40.4	201	.97	50.2
Batch Z							
12B	50	214	118.1	16.6	78	.88	55.4
16S	71	197	120.0	19.0	112	.91	60.9
27E	43	219	131.3	21.5	102	.92	59.9
51B	50	218	120.5	22.0	107	.93	55.2
54E	46	220	131.6	19.8	75	.91	59.6
55B	48	227	153.6	21.6	101	.92	67.7
67N	47	214	122.5	19.9	108	.91	57.3
76W	32	146	67.1	15.1	58	.89	46.0
76Y	50	122	68.8	19.0	84	.94	56.1
94B	45	168	76.7	18.2	74	.90	45.6

When tests were administered in the schools, the targeted number of subjects was 50 at each school. The actual number to whom the tests were administered ranged from 32 for MOS 76W to 71 for MOS 16S; the mean was 50.1. The mean for coefficient alpha was .90.

An index of difficulty was computed by dividing the mean number of items correct by the number of items, that is, the percentage of items on a test that were correct on average. This percentage ranged from 41.4 for MOS 63B to 67.7 for MOS 55B. The mean percentage correct was 54.5.

Preparation of Batch A and Batch B Training Achievement Tests for Field Tests With Job Incumbents

After all the SME judgments were made and trainee tryouts completed, the items were revised in accordance with the SME and trainee comments. For the Batch A and Batch B MOS, the item pools were prepared for administration to job incumbents in the criterion field tests. Data from the field test administration were later used (along with data from the administration of the items to trainees, relevance data, and importance data) to convert the pools of draft items into the standardized training knowledge tests.

As the item pools were cut and items added or changed in these early test construction steps, the descriptive characteristics of the overall pool--that is, importance and relevance--inevitably changed as well. Items were dropped if they were judged to be of little importance or no relevance. The nature of the item budget was preserved by adding new items if necessary. The characteristics of the field test versions in terms of importance and relevance are reported in the following subsection. These data parallel those reported for the initial item pools.

SME Importance and Relevance Ratings: Field Test Version

Table III.8 shows the number of items, mean item importance rating for the three scenarios by job incumbents, and incumbent interrater reliability for the field test versions of the tests. Since the field tests included only Batches A and B, the data reported are for 9 of the 19 MOS. Most of these tests were culled of items prior to the field test and are consequently shorter than tests in the item pool. The basis for the culling has already been described. In addition, prior to the field test some items were added on which importance data had not been collected and for which no importance ratings were available.

As would be expected, the pattern of importance ratings across scenarios by job incumbents was little affected by the culling procedure. There was also little difference in the mean across all scenarios between the item pool (3.40) and field test versions (3.43).

Table III.9 shows the ratings by trainers on the average importance of items retained for the field test. The table also contains trainer interrater reliabilities for Batches A and B.

As expected for the culled tests, mean importance ratings were very slightly higher for field test versions than for the item pools for both incumbents (3.43 vs. 3.40) and trainers (4.02 vs. 3.97). As already discussed in connection with the item pool, trainers rated item importance higher overall than did incumbents.

Table III.8

Mean Item Importance Ratings by Job Incumbents for Three Scenarios
(Field Test Version of Training Achievement Tests)

MOS	Number		Mean Item Importance Rating[a]			Interrater Reliability		
	Raters	Items	Scenario 1	Scenario 2	Scenario 3	Scenario 1	Scenario 2	Scenario 3
Batch A								
13B	7	180	3.9	3.9	3.8	.57	.58	.36
64C	4	210	4.3	4.8	4.9	.74	.75	.67
71L	4	119	2.6	2.9	3.3	.74	.75	.48
95B	5	221	3.0	3.1	2.9	.88	.74	.85
Batch B								
11B	5	143	3.9	3.8	3.7	.73	.71	.76
19E	5	202	3.6	3.6	3.8	.64	.63	.66
31C	5	187	4.0	3.9	3.8	.50	.45	.41
63B	5	216	2.2	2.6	3.2	.72	.64	.49
91A/B	5	260	2.0	2.4	2.9	.75	.74	.70

[a]Rated on a 5-point scale from "Of Little Importance" (1) to "Very Important" (5).

Table III.9

Mean Item Importance Ratings by School Trainers (Field Test Version of Training Achievement Tests)

MOS	Number Raters	Number Items	Mean Item Importance Rating[a]	Interrater Reliability
Batch A				
13B	6	152	4.40	.73
64C	7	145	4.24	.78
71L	5	122	3.78	.95
95B	5	90	4.72	.50
Batch B				
11B	7	144	3.93	.41
19E	6	202	4.06	.62
31C	6	187	3.70	.62
63B	6	216	3.44	.48
91A	6	260	3.95	.78

[a]Rated on a five-point scale from "Of Little Importance" (1) to "Very Important" (5).

Table III.10 contains the relevance data for the version of the test administered to incumbents in the field tests. The distribution across relevance categories is similar to that noted in connection with the pilot test version in Table III.6.

Field Test Instruments

At this stage the nine training achievement tests for the MOS in Batch A and Batch B were deemed ready for field testing with job incumbents. The field test procedure is described in Section 8, and the field test results and the subsequent modification of the tests are described in Section 9.

Up to this point the 10 tests for the 10 MOS in Batch Z followed the same developmental steps as for the tests in Batches A and B. However, as noted previously, the Batch Z instruments were not field tested with job incumbents. Consequently, the Concurrent Validation versions of these 10 tests retain more items than do the 9 A/B tests. Additional item analyses will be carried out for Batch Z on the basis of the data from the Concurrent Validation sample.

Copies of the 19 MOS tests as used in Concurrent Validation are contained in the ARI Research Note under preparation.

Table III.10

Number and Percentage of Items Rated Relevant to Job and Training (Field Test Version of Training Achievement Tests)

MOS	Job Only N	Job Only %	Training Only N	Training Only %	Job and Training N	Job and Training %	Not Relevant N	Not Rated[a] N
Batch A								
13B	70	41.2	5	2.9	95	55.9	6	59
64C	80	37.2	0	0.0	135	62.8	0	13
71L	42	34.4	4	3.3	76	62.3	0	8
95B	64	31.5	8	3.9	131	64.5	11	20
Batch B								
11B	68	39.5	14	8.1	90	52.3	21	26
19E	32	16.2	9	4.6	156	79.2	2	5
31C	47	26.3	15	8.4	117	65.4	5	20
63B	48	23.0	8	3.8	153	73.2	2	8
91A	0	0.0	140	54.9	115	45.1	5	5
Total	451	26.2	203	11.8	1068	62.0	52	164

[a]Items added to the tests after relevance ratings had been collected.

Section 3

DEVELOPMENT OF TASK-BASED MOS-SPECIFIC CRITERION MEASURES[1]

The MOS-specific criterion measures described in this section concern the assessment of performance on a sample of job tasks that were identified as representative of all job tasks in the MOS. The general procedure was to develop a careful description of all the major tasks that comprise the job, draw a sample of these tasks, and develop multiple measures of performance on each task.

Objectives

The major objective is to develop reliable and valid task-based measures of first-tour performance in the nine Batch A and Batch B MOS. Such measures are intended to reflect that part of the performance domain having to do with job-specific technical competence. Hands-on (job sample) tests, paper-and-pencil knowledge tests, and ratings measures were developed for each job task.

While no one measure can be assumed in advance to be a better estimate of the job incumbent's "true" performance, intercorrelations among the measures are of interest for what they tell us about the common and unique variance. Consequently, another objective is to compare alternative measures in terms of their construct validity for assessing task proficiency.

As noted in Part I, nine MOS were selected for study (see Table III.11). These nine MOS were chosen to provide maximum coverage of the total array of knowledge, ability, and skill requirements of Army jobs.

Development Procedure

The design strategy for the MOS-specific measures involved, for each MOS, selection of approximately 30 tasks that accurately sampled the job domain. The total number of tasks was dictated primarily by time constraints. While the time required to assess performance on individual tasks would differ with the nature of the task, a total of 30 tasks for each MOS seemed reasonable as a planning figure.

For each MOS, all 30 tasks would be assessed with written knowledge tests. Fifteen of the 30 tests would also be assessed with hands-on tests. Finally, task performance ratings would be obtained for the 15 tasks measured with the hands-on job sample tests, and job history data covering recency and frequency of performance would be researched for all 30 tasks.

---

[1] This section is based primarily on ARI Technical Report 717, Development and Field Test of Task-Based MOS-Specific Criterion Measures, by Charlotte H. Campbell, Roy C. Campbell, Michael G. Rumsey, and Dorothy C. Edwards, and the supplementary ARI Research Note in preparation, which contains the report appendixes.

Table III.11

Military Occupational Specialties (MOS)
Selected for Criterion Test Development

Batch A	
13B	Cannon Crewman
64C	Motor Transport Operator
71L	Administrative Specialist
95B	Military Police

Batch B	
11B	Infantryman
19E	Armor Crewman
31C	Single Channel Radio Operator
63B	Light Wheel Vehicle Mechanic
91A	Medical Specialist

The MOS-specific task tests and the auxiliary instruments were developed and field tested for the Batch A MOS before we began developing the tests in the remaining five MOS (Batch B). While the procedures were the same for Batch A and Batch B, some lessons learned from Batch A development were applied to Batch B. Across all nine MOS some individual variation was necessary because of particular circumstances in an MOS; however, variations were slight. The general procedure was composed of eight major activities:

- Define task domain.
- Collect SME judgments.
- Analyze SME judgments.
- Select tasks to be tested.
- Assign tasks to test mode.
- Construct hands-on and knowledge tests.
- Conduct pilot tests and make revisions.
- Construct auxiliary instruments.

These eight major activities are discussed in the following subsections.

Definition of Task Domain

Defining task domain involved dealing with the entire population of tasks for an MOS. The job task descriptions of first-tour (Skill Level 1) soldiers were derived from three primary sources:

1. MOS-Specific Soldier's Manuals (SM). Each MOS Proponent, the agency responsible for prescribing MOS policy and doctrine, prepares and

publishes a Soldier's Manual that lists and describes tasks, by Skill Level, that soldiers in the MOS are doctrinally responsible for knowing and performing. The number of tasks varies widely among the nine MOS, from a low of 17 Skill Level 1 (SL1) tasks to more than 130 SL1 tasks.

2. Soldier's Manual of Common Tasks (SMCT) (FM 21-2, 3 October 1983).[2] The SMCT describes tasks that each soldier in the Army, regardless of his or her MOS, must be able to perform. The 1983 version contains 78 SL1 tasks and "supersedes any common tasks appearing in MOS-specific Soldier's Manuals" (p. vii).

3. Army Occupational Survey Program (AOSP). The AOSP obtains task descriptions by surveying job incumbents with a questionnaire checklist that includes several hundred items. The items are obtained from a variety of sources (e.g., the Proponent school), and include and expand the doctrinal tasks from the preceding two sources. The AOSP is administered periodically to soldiers in all skill levels of each MOS by the U.S. Army Soldier Support Center. The analysis of responses by means of the Comprehensive Occupational Data Analysis Program (CODAP) provides the number and percentage of soldiers at each skill level who report that they perform each task. The number of tasks or activities in the surveys for the nine MOS of interest ranged from 487 to well over 800.

While the above sources provided the main input to the MOS job descriptions, Proponent agencies were also contacted directly to determine whether other relevant tasks existed. The number of additional tasks thus generated was not large, but the added tasks were sometimes significant. For example, the pending introduction of new equipment added tasks that had not yet appeared in the written documentation.

Completion of the above process resulted in a not very orderly accumulation of tasks, part tasks, steps, and activities. To bring some order to this collection, a six-step refinement process was conducted for each MOS.

1. Identify AOSP activities performed at SL1. The assumption for this step was that every activity included in an AOSP survey that had a non-zero response frequency among SL1 soldiers, after allowing for error in the survey, was performed at SL1. The procedure for estimating the error was to compute the boundaries of a confidence interval about zero. Tasks or activities with frequencies above the confidence interval boundary were considered to have non-zero frequencies and were retained. The percentage of tasks/activities dropped from each AOSP by this application was about 25% for each MOS.

---

[2]For Batch A MOS, the version of Field Manual 21-2 in effect during task selection was the 1 December 1982 edition, containing 71 tasks.

2. Group AOSP statements under SM tasks. An AOSP questionnaire item was referenced to an SM task if the item duplicated the SM task or was subsumed under the SM task as a step or variation in conditions, even if doctrine did not specifically identify the activity as an SLI responsibility.

3. Group AOSP-only statements into tasks. Since some AOSP questionnaire items could not be matched with any SM task, the next step was to edit such AOSP items so that they were similar in format to the SM task statements but were still a clear portrayal of additional task content not contained in the SM.

4. Consolidate domain (Proponent review). The first three steps resulted in a fairly orderly array of tasks for each MOS. With this task list it was feasible to go to the Proponent agency for each MOS for verification. At each Proponent a minimum of three senior NCOs or officers reviewed the list and eliminated tasks that had been erroneously included in the domain. While specific reasons for dropping tasks varied with each MOS, the general categories were:

   - Tasks specific to equipment that was being changed.
   - Tasks eliminated by current doctrine not yet reflected in available publications.
   - Collective tasks actually performed by crews/squads, platoons, or even companies/batteries.
   - Tasks specific to equipment variations that should be combined.
   - Tasks specific to the mission of the Reserve Component. While for most units there are no discriminable task differences between active duty and Reserve Component organizations, this is not true for all MOS.

   The full consolidated domain list of tasks, with supporting AOSP statements for each MOS, is contained in Appendix B, ARI Research Note in preparation.

5. Delete tasks that pertain only to restricted duty positions. The SM for most MOS contains tasks for individual duty positions within the MOS. For example, the 64C Motor Transport Operator can be a Dispatcher; the 95B Military Policeman can be a Security Guard. For most duty positions, incumbents move freely in and out of the position, the performance of the duty position tasks being dependent on whether they are assigned the position or not. Other positions are more permanent. Restricted Duty Positions were operationally defined as those for which the award of an Additional Skill Identifier (ASI) or Special Skill Identifier (SSI) and at least 1 week of specialized training were required. Five duty positions in two MOS were affected.

6. Delete Higher Skill Level (HSL) and AOSP-only tasks with atypically low frequencies. The first step in this process was to translate AOSP frequencies into task frequencies. Generally, when AOSP and task statements matched, the AOSP frequency for the matching statement was applied to the task. If there was no match, the most frequent step or condition was the basis for the task frequency. However, in some cases, frequencies were aggregated to account for equipment differences.

   The general approach for identifying low-frequency tasks was to compare frequency distributions of the SL1 tasks with the HSL and AOSP-only tasks. A four-step procedure identified the atypically infrequent tasks to be eliminated:

   - List the response frequencies of SL1 tasks from the AOSP/CODAP.
   - List the response frequencies of HSL/AOSP-only tasks.
   - Test groups (lists) for difference, using Mann-Whitney U test.
   - If the groups were different, and the HSL/AOSP-only group had tasks with lower response frequencies (which they did in all cases), eliminate those low-frequency tasks until group differences were not significant at .01 level.

The result of this process was a final task list for each MOS. It included all SL1 MOS and common tasks with non-zero frequencies (or no AOSP/CODAP frequency) and HSL/AOSP-only tasks performed by SL1 soldiers. Table III.12 shows the reduction of the task list during each phase and the reasons for the reduction by MOS. The nine final task lists are contained (with data from the SME judgments, detailed below) in Appendix C, ARI Research Note in preparation.

Collection of SME Judgments

After the MOS domains were refined, every domain comprised more than 100 tasks. To select 30 representative tasks for each MOS, more information was needed. MOS Proponent agencies were asked to provide subject matter experts (SMEs) regarding the tasks on the task list. Requirements were that they be in the grade E-6 or above (i.e., second or third tour) or officers in the grade O-3 (captain) or above. Recent field experience supervising SL1 personnel was an additional requirement. For Batch A MOS, 15 SMEs in each MOS were requested. For Batch B, some modifications were made in the review process (detailed below) and 30 SMEs in each of these MOS were requested. Collection of SME data required approximately 1 day for each MOS. The number of SMEs obtained for each MOS and samples of all instructions provided to SMEs are contained in Appendix D, ARI Research Note in preparation.

Three types of judgments were obtained from the SMEs:

Task Clustering. Each task was listed on a 3" x 5" card along with a brief description. SMEs were told to sort the tasks into groups so that all the tasks in each group were alike, and each group was different from the

Table III.12

Effects of Domain Definition on MOS Task Lists

AOSP Review	13B	64C	71L	95B	11B	19E	31C	63B	91A
AOSP Statements	669	677	822	546	822	609	656	633	685
Deleted - Zero Frequency	67	169	329	197	188	103	134	84	267
Deleted by SME	--	--	58	--	--	--	--	195	61
AOSP Statements Used	602	508	435	369	634	506	522	354	357

Domain Consolidation	13B	64C	71L	95B	11B	19E	31C	63B	91A
Tasks in MOS	378	166	203	304	357	338	267	188	251
Nonapplicable Systems	-	-	-	-	-	50	-	-	-
Eliminated By Doctrine	23	-	-	-	16	14	97	10	12
Collective Tasks	25	-	-	-	5	-	-	-	-
Combined Systems	57	-	-	-	-	-	-	-	-
Reserve Component Tasks	-	-	-	-	15	-	-	-	-
Tasks in Domain	273	166	203	304	321	274	170	178	239

Domain Reduction	13B	64C	71L	95B	11B	19E	31C	63B	91A
Tasks in Domain	273	166	203	304	321	274	170	178	239
Restricted Duty Position	44	-	42	-	-	-	-	-	-
Preliminary Sort	-	-	-	176	-	-	-	-	-
Low Frequency-HSL/AOSP Only	53	47	-	-	90	39	-	-	-
Domain Tasks For SME Judgments	177	119	161	128	231	235	170	178	239

other groups. For the Batch B MOS, common tasks were grouped for the SMEs, based on the clustering derived from the Batch A data. SMEs were permitted to add to or break up the groups as they saw fit.

Task Importance. To set the context for the Batch A MOS task importance judgments, all SMEs were given a European scenario that specified a high state of training and strategic readiness but was short of involving actual conflict. After collection of Batch A data, concern was expressed as to the scenario effect on SME judgments. As a result, for Batch B MOS three scenarios were used. An "Increasing Tension" scenario identical to that used in Batch A was retained, and a "Training" scenario specifying a stateside environment and a "Combat" scenario (European non-nuclear) were developed. Sample MOS definitions for the three scenarios are given in Figure III.3. These scenarios are similar to those used in the training achievement test development (see Section 2).

The SMEs for each Batch B MOS were randomly divided into three groups and each group was given a different scenario as a basis for judgments. However, for MOS 63B (Light Wheel Vehicle Mechanic) only 11 SMEs were available and a repeated measures procedure was used; that is, each 63B SME rated task importance three times, using each of the three scenarios in counterbalanced order. To make their judgments, SMEs were asked to rate the importance of the task in performing the MOS job in support of the unit mission under the appropriate scenario.

Slightly different procedures were used in Batch A and Batch B. For Batch A MOS, the judges were given the tasks on individual cards, identical to those used in task clustering, and told to rank the tasks from Most Important to Least Important. For Batch B MOS, judges were provided a list of the tasks, with descriptions, and asked to rate them on a 7-point scale from "1 = Not at all important for unit success" to "7 = Absolutely essential for unit success."

Task Performance Difficulty. To arrive at an indication of expected task difficulty, SMEs were asked to sort a "typical" group of 10 soldiers across five performance levels based on how they would expect a typical group of SLI soldiers to perform on each task.

Analysis of SME Judgments. The judgment data were analyzed and the following products were obtained:

> Cluster Membership. Task clusters were identified by means of a factor analysis of a cross-product matrix derived from the SMEs' task similarity clusterings.
>
> Importance. The importance rank of each task, averaged across SMEs, was analyzed by computer. For Batch A MOS, a single importance score was obtained. For Batch B MOS, a rank ordering of average importance ratings was generated under each of the three scenarios.

Your personnel unit is deployed to Europe as part of a U.S. Corps during a deteriorating political and military situation. The Corps mission is to defend and maintain the host country's border in the event that hostilities escalate. The enemy approximates a combined arms army and has nuclear and chemical capability. Air parity does exist. The Corps has drawn all equipment and is fully operational. In support of the Corps Personnel Operations Center, your unit is responsible for supporting the functions of strength accounting, replacement operations, casualty reporting, personnel management, personnel actions, and personnel records.

--Neutral or "Increasing Tension" Scenario for MOS 71L

Your soldiers are assigned to support the activities of the installation Provost Marshal on a large Army post in the midwestern United States. Post activities include a basic training center, an officer/enlisted training school, and maneuver units under a separate brigade organization. There is a permanent on-post dependent population of approximately 4,000. Provost activities supported include physical security and crime prevention, investigations, traffic and game warden operations, K-9 section, AWOL apprehension/civil liaison, vehicle and weapons registration, and operation of the installation detention facility. 95B personnel also must complete individual soldier training, SQT testing, command and maintenance inspections, and team/unit training.

--Training (CONUS) Scenario for MOS 95B

Your tank battalion 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.

--Combat Scenario for MOS 19E

Figure III.3. Scenarios used in SME ratings of task importance for task-based MOS-specific tests.

Difficulty. To calculate judged task difficulty, the mean of the distribution of 10 hypothetical soldiers across the five performance levels of each task, averaged across SMEs, was computed.

Performance Variability. The standard deviation of the distribution of 10 hypothetical soldiers across the five performance levels on each task was averaged across SMEs. This statistic is intended to be an indicator of the variability in performance that would be expected of a task.

Selection of Tasks To Be Tested

While the methods used for selecting tasks were similar for Batch A and Batch B MOS, there were enough differences to warrant their being outlined separately.

Batch A Test Task Selection. From five to nine project staff, including the individual who had prime responsibility for that particular MOS, participated in the selection process for each MOS. The task selection panel was provided the data summaries of the SME judgments and asked to make an initial selection of 35 tasks to represent each MOS. No strict rules were imposed on the analysts in making their selections, although they were told that high importance, high performance variability, a range of difficulty, and frequently performed tasks were desirable, and that each cluster should be sampled.

To capture the policy that each staff person used, task selections were first regressed on the task characteristics data to identify individual selection policies. The equations were then applied to the task characteristics data to provide a prediction of the task selections each individual would have made if his or her selections were completely consistent with a linear model.

In the second phase of selection, analysts were given their original task selections and the selections predicted by their regression-captured policies. They were directed to review and justify discrepancies between their observed and predicted selections. Analysts independently either modified their selections or justified their original selections. The rationale for intentional discrepancies was identified and the regression equations adjusted.

The next phase was a Delphi-type negotiation among analysts to converge their respective choices into a list of 35 tasks for each MOS. Information on the choices and rationale provided by each analyst in the preceding phase was distributed to all analysts, and each made a decision to retain or adjust his or her decisions, taking into account opinions others had expressed. Decisions and revisions were collected, collated, and redistributed as needed until near consensus was reached. For all MOS, three iterations were necessary.

The resulting task selection lists were mailed to each Proponent; a briefing by Project A staff was provided if requested. A Proponent representative then coordinated a review of the list by Proponent personnel designated as having the appropriate qualifications. After some minor Proponent-recommended adjustments, the final list of 30 tasks was selected. These are listed in Appendix F, ARI Research Note in preparation.

Batch B Test Task Selection. Based on experiences with Batch A selection, some modifications were introduced in the selection process for Batch B. One primary concern was to involve Proponent representatives more actively in the selection process. Also, the Batch A experience showed analysts' selections to be non-linear. Analysts qualified their selections on the basis of knowledge of the MOS or the tasks, information not directly represented in the data; they used non-linear combinations that often differed within each cluster. Therefore, the decision was made to drop the regression analysis for Batch B selection.

The panel for Batch B selection consisted of five to nine members of the project staff, as in Batch A, combined with six military personnel (NCO and officers) from each MOS. These six were in the grade of E-6 or higher with recent field experience, and were selected to provide minority and gender (for applicable MOS) representation to the task selection process.

The materials provided the selection panel were the same variables generated by the SME judgments. Again, no strict rules were imposed. However, panel members were provided a target number of tasks to be selected from each cluster (calculated in proportion to the total number of tasks in each cluster). A second adjustment prescribed a minimum of two tasks per cluster to permit estimation of the correlation among tasks in the cluster. Within these constraints the Delphi procedure was again used to reach consensus. The tasks selected for Batch B MOS are listed in Appendix F, ARI Research Note in preparation.

Assignment of Tasks to Test Mode

The initial development plan required that, for each MOS, knowledge tests be developed for all 30 tasks, and hands-on tests for 15 of these tasks (since such testing for the all 30 tasks would exceed the hands-on resources). The considerations that constrained selection for hands-on testing were:

- Fifteen soldiers must complete all 15 hands-on tests in 4 hours. No single test is to take more than 20 minutes.
- Scorer support would be limited to eight NCO scorers.
- The hands-on test site must be within walking distance of the other test activities.
- Equipment requirements must be kept within reason if units are to support the requirements.

- The test must be administrable in a number of installations. Tasks must not be affected by local operating procedures.

On the basis of these constraints, in each MOS a project staff member prepared an anticipated hands-on test approach for each of the 30 tasks in the MOS. Working independently, each of the five project analysts first reviewed the suggested test approach, and modified it as he/she deemed necessary. The analyst then assigned points to each task to indicate hands-on test suitability, using the following three areas of consideration:

- Skill Requirements - Analysts determined a numerical value for skill requirements based on the number of steps requiring physical strength, control, or coordination. Each skill step was counted as one point.

- Omission - This rating considered the likelihood that a soldier would omit a required step. For a step to have "omission value":

  - A soldier must be able to complete the procedure (albeit incorrectly) without performing the step.

  - Nothing in the test situation must cue the soldier to do the step.

  Each "omission step" received a numerical rating of one.

- Time Value - When "doctrine" (usually the SM) specified a time limit for task performance, the task was awarded a numerical value of two. Where no doctrinal time limit has been established but where time would be a reliable indication of task proficiency, the task was awarded a numerical value of one.

Following the individual ratings, analysts met in group discussions and proceeded task by task to resolve differences until a consensus was reached and a single numerical score was assigned to each task. The tasks were then rank ordered, and a final feasibility check was conducted to ensure that the top 15 rated tasks fell within the 4-hour time limit. As an example, the tasks selected and assigned for MOS 11B are shown in Figure III.4.

Construction of Hands-On and Knowledge Tests

For both hands-on and knowledge tests, the primary source of information was task analysis data. Task analyses were derived from the Soldier's Manuals, technical manuals, and other supporting Army publications, as well as SME input and direct task observation where necessary. Much of the development effort involved having specific staff members work on both types of tests for the same tasks.

HANDS-ON AND KNOWLEDGE TESTS	KNOWLEDGE TESTS ONLY
1. Put on Field or Pressure Dressing	1. Perform CPR
2. Perform Operator Maintenance on M16A1	2. Administer Nerve Agent Antidote
3. Load, Reduce Stoppage, Clear M60 Machine Gun	3. Call for/Adjust Indirect Fire
4. Set Headspace/Timing on .50 Cal Machine Gun	4. Navigate on Ground
5. Engage Targets With Hand Grenades	5. Put on Protective Clothing
6. Prepare Dragon for Firing	6. Collect/Report Information — SALUTE
7. Prepare Range Card for M60 Machine Gun	7. Camouflage Self and Equipment
8. Engage Targets With LAW	8. Recognize Armored Vehicles
9. Put on/Wear M17 Gas Mask	9. Move Under Direct Fire
10. Operate Radio AN/PRC-77	10. Estimate Range
11. Operate as Station in Radio Net	11. Perform PMCS-M113 or 1/4 Ton
12. Install/Fire/Recover Claymore	12. Drive Wheeled or Track Vehicle
13. Techniques of MOUT	13. Hasty Firing Position, Urban Terrain
14. Zero AN/PVS-4 on M16A1	14. Establish Observation Post
15. Conduct Surveillance w/o Electronic Devices	15. Select Overwatch Position
	16. Place AN/PVS-5 into Operation

Figure III.4. Infantryman (MOS 11B) tasks selected for Hands-On/Knowledge Testing.

Hands-On Test Development. The model for hands-on test development emphasized four activities:

- Determine test conditions. Test conditions are designed to maximize the standardization of the test between test sites and among soldiers at the same test site. Test conditions are determined for the test environment, equipment, location, and task limits.

- List performance measures. The performance measures are the substantial elements of the task to be tested and the behaviors to be rated GO/NO-GO by the scorer. Performance measures are defined as either product or process depending on what the scorer is directed to observe to score the behavior. Performance measures must adhere to the following principles:

  - Describe observable behavior only.

  - State a single pass/fail behavior.

  - Contain only necessary actions.

- Contain a standard (how much or how well).

- State an error tolerance limit if variation in behavior is permissible.

- Include a scored time limit if, and only if, the task or step is doctrinally time-constrained; that is, the Soldier's Manual specifies a time limit for performing the task.

- Include a sequence requirement if, and only if, sequence is doctrinally required.

- State examinee instructions. The instructions must be kept very short and very simple; any information not absolutely essential to performance must be excluded. Examinee instructions are read verbatim to the soldier by the scorer and may be repeated at any time. These written instructions are the only verbal communications the scorer is allowed to have with the soldier during the test.

- Develop scorer instructions. These instructions tell the scorer how to set up, administer, and score the test. They cover both usual and unusual situations, and ensure standardized administration and scoring.

Examples of one hands-on task from the MOS 71L and MOS 11B protocols are shown in Figures III.5 and III.6.

Knowledge Test Development. The format of the knowledge tests was dictated by their proposed use. For example, free-response formats demand more of the soldier's literacy skills and are more difficult to score reliably than are multiple-choice formats, which are easier to score and are familiar to most soldiers. However, multiple-choice formats are difficult to develop because of inherent cueing, particularly between items, and the need to develop alternatives that are likely and plausible but clearly wrong. Because of the large quantity of data to be gathered in the project, machine scoring is essential. Therefore, a multiple-choice, single-correct-response format was selected.

Test administration constraints dictate the number of tasks to be tested and the time available for testing. For Project A, all tasks selected (approximately 30 per MOS) would be tested in the knowledge mode. Four hours were allocated to the knowledge testing block for the field trials, to be reduced to 2 hours for Concurrent Validation testing. Allowing an average of slightly less than 1 minute to read and answer one item dictated an average of about nine items per task.

Knowledge test development was based on the same information that was available for hands-on development. However, three distinct characteristics of multiple-choice performance knowledge test items are that they:

- Are performance-based. Most tasks, of course, cannot elicit full job-like behavior in the knowledge mode and therefore must be tested using performance-based items. These items require the examinee to select an answer describing how something should be done.

SCORESHEET

TYPE A MEMORANDUM

Time

Start: _____

Finish: _____

SCORESHEET

Scorer: _____ Soldier: _____

Date: _____ SSN: _____

Soldier ID: ___ _____

NOTE TO SCORER: Tell the soldier: "ASSUME YOU HAVE JUST RECEIVED THIS DRAFT TO BE TYPED AS A MEMORANDUM IN FINAL FORM. YOU MAY REFER TO THE SUPPLEMENT BOOK AND THE DICTIONARY IF YOU WANT. YOU CAN MAKE CORRECTIONS. WORK AS FAST AND AS ACCURATELY AS YOU CAN. HOW MANY COPIES OF THE ORIGINAL WOULD YOU MAKE?" NUMBER OF COPIES: _____

PERFORMANCE MEASURES	GO	NO-GO
1. Correct number of copies (2) • 1 white • 1 yellow manifold	____	____
2. One inch left and right margins • ± 1 space	____	____
3. Correct letterhead • 5th line below top • Centered • DA all caps, other initial caps	____	____
4. Correct reference (office) symbol-AZAK-YD • Left margin • 4th line below last letterhead line	____	____
5. Correct date - 10 October 1984 or 10 OCT 84 • End right margin • Same line as reference	____	____

Figure III.5. Administrative Specialist (MOS 71L) Hands-On Performance Test sample (Page 1 of 2).

PERFORMANCE MEASURES	GO	NO-GO
6. Correct memo addressee • 4th line below reference • Left margin • All capitals • 2nd addressee below 1st	____	____
7. Correct SUBJECT line • Left margin • 2nd line below last addressee • Colon after SUBJECT, 2 spaces	____	____
8. Correct body • 5th line below subject, left margin • Paragraphs numbered • Numbers and all lines in left margin • Single space • Double space between paragraphs	____	____
9. Correct Authority line • 2nd line below last body line • Left margin • All caps • Colon after FOR THE COMMANDER	____	____
10. Correct signature block • 5th line below Authority line • Begins at center • Name (all caps), rank (initial or all caps), branch (all caps) • Title initial caps	____	____
11. Corrections neat and clean Number of typographical errors: _____ • Sample typos: strikeovers, misspelling, incorrect punctuation, incorrect spacing.	____	____

Figure III.5. Administrative Specialist (MOS 71L) Hands-On Performance Test sample (Page 2 of 2).

Scorer: _____ Soldier: _____

Date: _____ ID: _____

Check:	Yes	NO
Know Soldier?	___	___
Soldier in CO.?	___	___
Supervise Soldier?	___	___

SCORESHEET

LOAD, REDUCE A STOPPAGE AND CLEAR AN M60 MACHINEGUN

INSTRUCTIONS TO SOLDIER: For this test you must load, fire, apply immediate action and clear the M60 machinegun. Do not go on to the next procedure until I tell you to. First, you must load and fire the machinegun. Begin.

PERFORMANCE MEASURES:	GO	NO-GO	COMMENTS
Load			
1. Placed safety in FIRE.	___	___	_____
2. Pulled cocking handle to the rear, locking bolt to the rear.	___	___	_____
3. Returned cocking handle forward.	___	___	_____
4. Placed safety in SAFE.	___	___	_____
5. Raised cover.	___	___	_____
6. Lifted rear of gun slightly to observe into chamber.	___	___	_____
7. Positioned belt with double loop toward gun and split link down.	___	___	_____
8. Placed round in feed tray groove.	___	___	_____
9. Closed cover.	___	___	_____
10. Lifted up on cover to insure cover locked.	___	___	_____
11. Placed safety in FIRE.	___	___	_____
12. Pulled trigger.	___	___	_____
13. Performed steps in sequence.	___	___	_____
Seconds to load machinegun:	_____		

Figure III.6. Infantryman (MOS 11B) Hands-On Performance Test sample (Page 1 of 2).

PERFORMANCE MEASURES:	GO	NO-GO	COMMENTS
Immediate Action			
INSTRUCTIONS TO SOLDIER: You have been firing the machinegun and the weapon suddenly stops firing. Apply immediate action. Begin.			
14. Pulled cocking handle to the rear, locking bolt to rear. (Round ejects)	___	___	_____
15. Returned cocking handle forward.	___	___	_____
16. Pulled the trigger.	___	___	_____
17. Performed steps in sequence.	___	___	_____
Seconds to perform immediate action:	_____		
INSTRUCTIONS TO THE SOLDIER: You must now unload and clear the machinegun. Begin.			
18. Pulled cocking handle to the rear, locking bolt to the rear.	___	___	_____
19. Returned cocking lever forward.	___	___	_____
20. Placed safety in SAFE.	___	___	_____
21. Opened cover.	___	___	_____
22. Removed ammunition belt.	___	___	_____
23. Lifted rear of gun slightly to observe into chamber.	___	___	_____
24. Closed cover.	___	___	_____
25. Placed safety in FIRE.	___	___	_____
26. Pulled cocking handle to rear.	___	___	_____
27. Pulled trigger and eased cocking handle forward. (Must hold onto handle; bolt must not slam forward.)	___	___	_____
28. Placed safety in SAFE.	___	___	_____
29. Performed steps in sequence.	___	___	_____
Seconds to unload and clear machinegun:	_____		

Figure III.6. Infantryman (MOS 11B) Hands-On Performance Test sample (Page 2 of 2).

A prevalent pitfall in performance knowledge test development is a tendency to cover information about why a step or action is done or rely on technical questions about the task or equipment. Just as in the hands-on tests, the objective of the knowledge test is to measure the soldier's ability to perform a task. The knowledge or recall required by the test item must not exceed what is required of a soldier when he or she is actually performing the task. Because of this performance requirement, knowledge tests must present job-relevant stimuli as much as possible, and the liberal use of quality illustrations is essential.

- Identify performance errors. Performance-based knowledge tests must focus on what soldiers do when they fail to perform the task or steps in the task correctly.

- Present likely alternatives. The easiest answer to write is the correct alternative. The approach here focuses on identifying what it is soldiers do wrong when they perform a step; that is, if they do not perform the step correctly, what is it that they do perform? This information becomes the basis for the other alternatives.

Knowledge tests were constructed by project personnel with experience in test item construction and expertise in the MOS/task being tested. Test items were reviewed internally by a panel of test experts to ensure consistency between individual developers. The following general guidelines were used in construction:

- Stem length for items was usually restricted to two lines. Where needed, a "Situation" was separately described if it could be applied to two or more items.

- Item stems were designed so that the item could be answered based on the stem alone, that is, without reference to the alternatives.

- Illustrations were used where they could duplicate job cues. Where necessary, illustrations were also used as alternatives or to provide a job-related reference. All illustrations were drawings.

- Each task tested was a separate entity, clearly identified by task title and distinct from other tested tasks.

- For items that allowed or required use of publications on the job, abstracts were prepared. If the publication was lengthy (e.g., tables of vehicle maintenance checks), the abstract was provided as a separate handout. Brief abstracts, of one page or less, were appended to the test. Materials needed in performance knowledge tests, such as maps, protractors, and scratch paper, were also provided.

- Test items within a task test were arranged in the sequence in which they would normally occur when the soldier performed the task.

- Completed tests were checked for inter-item cueing.
- All correct alternatives were authenticated as correct by a citable reference.

In four of the nine MOS, some of the tasks that incumbents perform are affected by the type of equipment to which they are assigned. For these MOS it was necessary to develop separate tracked versions of tests covering the specific items of equipment involved.

Pilot Tests and Revisions

Following construction of the tests, arrangements were made through the Proponent for troop support for a pilot testing of the hands-on and knowledge tests. This procedure was conducted by the test developer and involved the support of four NCO scorers/SMEs, five MOS incumbents in SL1, and the equipment dictated by the hands-on test.

Pilot of Hands-On Tests. The following activities were performed:

1. Test Review - The four NCO scorers independently reviewed the instructions to scorer, and the scoresheets. The developer noted comments or questions that could be clarified by changes or additions to the materials.

2. Test Set-Up - One of the scorers set up the test as directed in the prepared instructions. The developer noted deficiencies or necessary changes in the instructions.

3. Scoring - One of the incumbents performed the test while the four scorers scored the test independently. After the test, all four scoresheets were compared. Discrepancies in scoring were discussed and the reasons ascertained. Some scorer discrepancies were the result of a scorer's physical position relative to the incumbent, but many required changing a performance measure or the instructions to scorers, or even changing the test or performance procedure itself. If possible, these changes were made before the next incumbent was tested. Normally, variations in incumbent performances occur naturally, but to ensure variation the developer could cue incorrect performance without the scorers' knowledge. Testing continued with the other incumbents, followed by scoresheet review and revision. The incumbents were included in the review process to assist in determining how they actually performed.

4. Examinee Debriefings - Incumbents were interviewed to determine whether the instructions provided them adequate guidance on what they were expected to do.

5. Time Data - Performance times were kept on all incumbents, as were station and test set-up times.

Based on the pilot test information, a final version of each hands-on test was prepared. These tests are contained in Appendix G, ARI Research Note in preparation.

Pilot of Knowledge Tests. The knowledge tests were pilot tested at the same time as the hands-on measures. The same four NCO hands-on scorers and five MOS incumbents were utilized but the procedure was different for the two groups.

1. NCO SME - The test developer went through each test, item by item, with all four NCOs simultaneously. The specific questions addressed were:

   - Would the SL1 soldier be expected to perform this step, make this decision, or possess this knowledge in the performance of this task on the job?
   - Is the keyed alternative correct?
   - Are the "incorrect" alternatives actually incorrect?
   - Is there anything in local or unit SOP that would affect the way this task item is performed?
   - Are the illustrations clear, necessary, and sufficient?
   - Is there any aspect of this task that is not covered in the test but should be covered?

2. Incumbents - The five incumbents took the test as actual examinees. They were briefed as to the purpose of the pilot test and told to attempt to answer all items. The tests were administered by task and the time to complete each task test was recorded individually. After each task test the incumbents were debriefed. The following questions were addressed:

   - Were there any items where you did not understand what you were supposed to do or answer?
   - Were there any illustrations that you did not understand?
   - (Item by item) This is what is supposed to be the correct answer for Item____. Regardless of how you answered it, do you agree or disagree that this choice should be correct?

Revisions based on SME and incumbent inputs were made to the tests. On the basis of the times obtained for the incumbents, the tests were divided into four sections or booklets of approximately equal lengths for Batch A MOS; for Batch B, tests were divided into two booklets. The purpose of dividing tests into several booklets and varying the order of administration among groups of soldiers was to distribute fatigue effects. These revised versions of the tests are contained in Appendix H, in ARI Research Note in preparation.

Construction of Auxiliary Instruments

Task-Specific Performance Rating Scales. Development of hands-on and knowledge tests provided two methods of measuring the sample of 15 tasks. As a third method, the soldier's peers and supervisors were asked to rate the soldier's performance on those same 15 tasks by means of a 7-point numerical rating scale. The intent was to assess performance on the same set of 15 tasks with three different methods. The rating scales were developed for administration during the field tests.

Job History Questionnaire. Although soldiers in a given MOS share a common pool of potential tasks, their actual task experience may vary substantially. The most widespread reason for this difference is assignment options in the MOS. The options may be formal, such as when an SL1 Armor Crewman may be a driver or a loader on the tank, or they may be informal specializations, such as when one Administrative Specialist types orders while another types correspondence. A more extreme reason for task difference in an MOS occurs when soldiers are assigned to duties not typically associated with their MOS. For example, an Armor Crewman may be assigned to drive a 1/4-ton truck, or a Medical Specialist may perform clerical tasks. Such soldiers are not given Special or Additional Skill Identifiers, nor are they considered to be working in a secondary MOS: They are simply tankers who drive trucks or medics who type and file. The likelihood of differences in task experience is further increased by differences in unit training emphasis where training schedules at battalion, company, and platoon level emphasize different tasks. As a result of these circumstances, soldiers' experiences vary, even within an MOS and location.

Given that the central thrust of Project A is the validation of selection and classification measures, any differential task experience that affects performance is a contaminating variable. That is, if the differences in task experiences of sampled soldiers are wide enough to have an impact on task performance, experience effects may also be strong enough to mask predictor relationships with performance. In this case, measures of experience would need to be incorporated into validation analyses so that predictor-criterion relationships could be assessed independent of experience.

To assess the likely impact of experience effects on task performance, and consequently on the Concurrent Validation strategies, a Job History Questionnaire was developed to be administered to each soldier. Specifically, soldiers were asked to indicate how recently and how frequently (in the preceding 6 months) they had performed each of the 30 tasks selected as performance criteria. A copy of the questionnaire for the MOS 13B Cannon Crewman is included as Appendix J, ARI Research Note in preparation.

Field Test Instruments

At this point the initial versions of the hands-on job sample tests and the multiple-choice knowledge tests had been developed, pilot tested, and revised. The 7-point task performance rating scales and the Job History

Questionnaire had been constructed. These instruments were included in the complete criterion array and field tested on samples of approximately 150 job incumbents from each of the Batch A and Batch B MOS.

The field test procedures for the MOS-specific task performance measures are described in Section 8, and the field test results and subsequent modifications of the various measures are described in Section 10.

Section 4

DEVELOPMENT OF MOS-SPECIFIC BEHAVIORALLY ANCHORED RATING SCALES (BARS)[1]

A major component of Project A criterion development is devoted to using the critical incident method to identify the basic set of performance factors that describe total job performance. Total performance has been conceptualized as composed of two types of factors, those that have the same meaning and interpretation across jobs and those that are specific to a particular job--that is, they are specific to the job's content. This chapter deals with the job-specific factors and the rating scales developed to measure them.

The procedure used to identify MOS-specific job factors was derived in large part from procedures outlined by Smith and Kendall (1963) and by Campbell, Dunnette, Arvey, and Hellervik (1973). Smith and Kendall recommended conducting critical incident workshops that involve, as a first step, naming and defining the major components of performance for the job in question. Workshop participants are then asked to write samples of effective and ineffective performance for each of the major components they have identified.

Campbell et al. suggested a slight modification to the Smith and Kendall procedure, recommending that performance categories be generated after participants have had an opportunity to write several incidents. In this way, participants are not constrained by a priori performance categories and are more likely to write performance examples that represent all job requirements.

In this procedure the next step involves editing the written critical performance incidents. These edited incidents are then used to identify the major dimensions of the job by asking supervisors and incumbents to read the performance incidents and make two ratings for each. Raters assigned each incident to a performance dimension and then indicated the level of performance on the dimension represented by that incident. The final product is a set of behaviorally defined and anchored performance dimensions that focus on the duties and standards of a specific job or MOS.

The purpose of this part of Project A is to develop behaviorally anchored performance rating scales (BARS) that assess job-specific performance factors for the nine MOS in Batch A and Batch B.

---

[1]This section is based primarily on an ARI Technical Report in preparation, Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS, by Jody L. Toquam, Jeffrey J. McHenry, VyVy A. Corpe, Sharon R. Rose, Steven E. Lammlein, Edward Kemery, Walter C. Borman, Raymond Mendel, and Michael J. Bosshardt, and a supplementary ARI Research Note, also in preparation, which contains the report appendixes.

Development Procedure

Each of the nine MOS was assigned to a specific member of the research staff. This individual assumed responsibility for (a) conducting workshops to collect performance incidents for the assigned MOS, (b) editing incidents, (c) preparing retranslation exercises, (d) developing performance rating scales, and (e) revising the scales for use in the Concurrent Validation efforts. Thus, a single researcher became an "expert" concerning the job duties and requirements involved in the assigned MOS.

Workshop Participants

Incumbents, or first-term enlistees, from target MOS were not, as a rule, included in the workshops, because their experience with the job was relatively limited. Almost all participants were noncommissioned officers (NCOs) who were directly responsible for supervising first-term enlistees and who had spent 2 to 4 years as first-termers in these MOS themselves. Consequently, most workshop participants were familiar with the job requirements from both an incumbent and a supervisor perspective.

To ensure thorough coverage and representation of the critical behaviors comprising each MOS, workshops for each MOS were conducted at six CONUS (Continental United States) Army posts. Each post was asked to designate from 10 to 16 NCOs for each target MOS. Thus, the goal was to obtain input from about 60 to 96 supervisors for each MOS. The total number of NCOs participating in the performance incident workshops by MOS is shown in Table III.13. The total array of posts at which workshops were held is shown in Table III.14.

Table III.13

Participants in MOS-Specific BARS Workshops

MOS		Number of Participants
Batch A		
13B	Cannon Crewman	88
64C	Motor Transport Operator	81
71L	Administrative Specialist	63
95B	Military Police	86
Batch B		
11B	Infantryman	83
19E	Armor Crewman	65
31C	Radio Teletype Operator	60
63B	Light Wheel Vehicle Mechanic	75
91A	Medical Specialist	71

Table III.14

Locations and Dates of MOS-Specific BARS Workshops

Location	Dates
Batch A	
Fort Ord	25-26 August 1983
Fort Polk	29-30 August 1983
Fort Bragg	12-13 September 1983
Fort Campbell	15-16 September 1983
Fort Hood	13-14 October 1983
Fort Carson	31 October - 1 November 1983
Batch B	
Fort Lewis	9-11 January 1984
Fort Stewart	11-13 January 1984
Fort Riley	16-18 January 1984
Fort Bragg	27-29 February 1984
Fort Bliss	12-14 March 1984
Fort Sill	14-16 March 1984

Collection of Data on Performance Incidents

After a workshop group was convened, research staff members serving as workshop leaders described Project A and briefed participants on the purpose of the workshop. This led to a discussion of the different types of performance rating scales available and to a discussion of the advantages of using behaviorally anchored rating scales to assess job performance. Leaders then described how the results from the day's activities would be used to develop this type of rating scale for that particular MOS.

Workshop leaders then provided instruction for writing performance incidents and distributed performance incident forms. Participants were asked to generate accounts of performance incidents, using examples provided as guides. Participants were asked to avoid writing about activities or behaviors that reflect general soldier effectiveness (e.g., following rules and regulations, military appearance), as these requirements have been identified and described in another part of the project.

After about 4-5 hours, performance incident writing was halted and workshop leaders began generating discussion about the major components or activities comprising the job. During this discussion, participants were asked to identify the major job performance categories, which workshop leaders recorded on a blackboard or flipchart. When participants indicated that all possible performance categories had been identified, the workshop leader asked them to review the list and consider whether or not all job

duties were indeed represented. The leader also asked participants to consider whether each category represented first-term enlistee job requirements or requirements of more experienced soldiers.

Following this discussion, participants were asked to review the performance incidents they had written and to assign them to one of the job categories or dimensions that appeared on the blackboard or flipchart. The workshop leader tallied the total number of incidents in each category. Those categories with very few incidents were the focus of the remainder of the workshop; participants were asked to spend the remaining time generating performance incidents for those categories.

Results from the performance incident workshops are reported in Table III.15 for Batch A MOS and in Table III.16 for Batch B MOS. The number of participants and the number of performance incidents generated are reported by MOS and by post. The mean numbers of incidents generated, the total number of participants, and total number of incidents are also reported by MOS and by post.

The schedule permitted research staff members time to edit and review performance incidents between data collection activities. For example, following the data collection activities at Fort Bragg and Fort Campbell, performance incidents were edited, content analyzed, and sorted into categories. These categories were then integrated with those generated during the earlier workshops and discussed with participants in subsequent workshops held at Fort Hood and Fort Carson.

A similar iterative procedure was used to generate performance dimensions for the Batch B MOS.

Retranslation Activities

A primary objective of the retranslation exercise is to verify that the performance dimension system provides a thorough and comprehensive coverage of the critical job requirements. The evidence for this verification is high agreement among judges that specific incidents represent particular components (factors) of performance, that all hypothesized factors can be represented by incidents, and that all incidents in the sample can be assigned to a factor (if they cannot, factors may be missing).

A second objective involves constructing the performance anchors for each dimension. Participants are asked to rate the level of performance described in the incident. These ratings are then used to help construct behavioral anchors that describe typical performance at different effectiveness levels within a single performance dimension.

Retranslation procedures employed for Batch A MOS differed from those for Batch B MOS, as noted in the following description.

Table III.15

BARS Performance Incident Workshops: Number of Participants and Incidents Generated by MOS and by Location - Batch A

Location	MOS 13B	64C	71L	95B	Total By Location
Fort Ord					
N - Participants	14	10	5	14	43
N - Incidents	195	80	59	213	547
Mean Per Participant	13.9	8.0	11.8	15.2	12.7
Fort Polk					
N - Participants	12	15	15	15	57
N - Incidents	150	240	210	235	835
Mean Per Participant	12.5	16.0	14.0	15.7	14.7
Fort Bragg					
N - Participants	13	14	11	17	55
N - Incidents	235	221	218	225	899
Mean Per Participant	18.1	15.8	19.8	13.2	16.4
Fort Campbell					
N - Participants	13	13	10	11	47
N - Incidents	195	191	154	238	778
Mean Per Participant	11.5	13.6	17.1	15.9	14.2
Fort Hood					
N - Participants	13	13	10	11	47
N - Incidents	180	183	133	92	588
Mean Per Participant	13.9	14.1	13.3	8.4	10.7
Fort Carson					
N - Participants	19	15	13	14	61
N - Incidents	204	232	215	180	831
Mean Per Participant	10.7	15.5	16.5	12.9	13.6
Total By MOS					
N - Participants	88	81	63	86	318
N - Incidents	1159	1147	989	1183	4478
Mean Per Participant	13.2	14.2	15.7	13.8	14.1

Table III.16

BARS Performance Incident Workshops: Number of Participants and Incidents Generated by MOS and by Location - Batch B

Location	MOS 11B	19E	31C	63B	91A	Total By Location
Fort Lewis						
N - Participants	16	11	8	10	11	56
N - Incidents	211	180	124	172	130	817
Mean Per Participant	18.3	16.4	15.5	17.2	11.8	14.6
Fort Stewart						
N - Participants	14	15	15	16	16	76
N - Incidents	216	275	256	208	249	1204
Mean Per Participant	15.4	18.3	17.1	13.0	15.6	15.8
Fort Riley						
N - Participants	18	7	10	11	8	54
N - Incidents	216	123	127	133	90	689
Mean Per Participant	12.0	17.6	12.7	12.1	11.3	13.8
Fort Bragg						
N - Participants	13	14	16	15	13	71
N - Incidents	231	190	220	250	217	1,108
Mean Per Participant	17.8	13.6	13.8	16.7	16.7	15.6
Fort Sill[a]						
N - Participants	8	4	3	9	10	34
N - Incidents	26	0	13	32	20	91
Mean Per Participant	3.3		4.3	3.6	2.0	2.7
Fort Bliss[a]						
N - Participants	14	14	8	14	13	63
N - Incidents	93	70	39	71	55	328
Mean Per Participant	6.6	5.0	4.9	5.1	4.2	5.2
Total By MOS						
N - Participants	83	65	60	75	71	354
N - Incidents	993	838	779	866	761	4,237
Mean Per Participant	12.0	12.0	13.0	11.6	10.7	12.0

[a] Participants at these posts spent most of the time completing retranslation booklets rather than generating critical incidents.

Retranslation Material and Procedure for Batch A. The Smith and Kendall (1963) procedure calls for including individuals familiar with the target job as participants in the retranslation process. For the Batch A MOS, we planned to include the participants from the earlier workshops in the retranslation phase; most of these persons were supervisors of the target incumbents, rather than incumbents. Participants were informed during the workshops that we would contact them via mail to complete another phase of the project.

After taking a count of the incidents, we decided that it was impractical to ask participants to rate all performance incidents generated for their MOS; the number of incidents per MOS ranged from 761 to 1,183. Instead, we asked participants to retranslate only a subset of the total incidents.

Return rates across all Batch A MOS were such that, on the average, only about 20% of the participants completed the retranslation task. This number of ratings proved insufficient for analyses. To increase the number of retranslation ratings, we conducted retranslation workshops at Fort Meade, Maryland, utilizing NCOs from the four MOS who were familiar with first-term enlistee job requirements. Further, project staff members from HumRRO who were familiar with the job requirements of one or more MOS also completed retranslation booklets.

Procedures for Batch B. Because of the low return rate for Batch A MOS, the procedures were modified for Batch B. Activities scheduled for the final two workshops, conducted at Fort Sill and Fort Bliss, varied from those described previously. At these workshops, participants spent the first 2 hours generating performance incidents describing MOS-specific job behaviors, then spent the remainder of their day completing retranslation booklets.

Participants were asked to complete as many retranslation booklets as possible. In general, each individual completed about one-and-a-half to two booklets. Also during this session, participants were asked to retranslate the performance incidents generated earlier during that session. Hence, we obtained retranslation ratings for all performance incidents generated at the first four workshops as well as retranslations for the new incidents generated at that particular workshop.

Construction of Initial Rating Scales

Table III.17 summarizes the number of ratings obtained from the retranslation exercise for Batch A and Batch B. The retranslation data were analyzed separately for each MOS. The process included computing for each incident (a) the number of raters, (b) percent agreement among raters in assigning incidents to performance dimensions, (c) mean effectiveness rating, and (d) standard deviation of the effectiveness ratings. Percent agreement values, mean effectiveness ratings, and standard deviations are provided for all performance incidents in Section 3 of the MOS appendixes in the ARI Research Note in preparation.

Table III.17

BARS Retranslation Exercise: Number of Forms Developed for Each MOS and Average Number of Raters Completing Each Form

MOS	Number of Forms	Number of Incidents/Form		Average Number of Raters/Form
		Average	Total	
Batch A				
13B	4	171	684	17.0
64C	5	191	955	12.6
71L	4	190	760	14.0
95B	5	229	1145	7.6
Batch B				
11B	2	274	548	19.0
19E	3	201	603	9.7
31C	3	235	705	9.0
63B	3	230	690	16.0
91A	3	210	630	17.7

The next step in the process involved identifying those performance incidents in which raters agreed reasonably well on performance dimension assignment and effectiveness level. For each MOS, we identified performance incidents that met the following criteria: (a) at least 50% of the raters agreed that the incident depicted performance in a single performance dimension, and (b) the standard deviation of the mean effectiveness rating did not exceed 2.0. These incidents were then sorted into their assigned performance dimensions. Results from this sorting are presented for each MOS in Table III.18.

After the incidents had been sorted into performance dimensions, the percentage agreement values were examined to identify dimensions that raters found confusing or difficult to distinguish from one another. On the basis of these data, some dimensions were dropped and some were collapsed.

After modifying the dimension system using results from the retranslation exercise, we developed behavioral anchors for each dimension. This involved sorting effective performance incidents with mean values of 6.5 or higher, average performance with mean values of 3.5 to 6.4, and ineffective performance with mean values from 1.0 to 3.4. We reviewed the content of the incidents in each of these three areas and then summarized the information in each to form three behavioral anchors depicting effective, average, and ineffective performance (see example in Figure III.7).

Table III.18

Behavioral Examples Reliably Retranslated Into Each Dimension on the BARS Measures

Dimension	Number of Examples
Cannon Crewman (13B)	
A. Loading out equipment	49
B. Driving and maintaining vehicles, howitzers, and equipment	195
C. Transporting/sorting/storing and preparing ammunition for fire	108
D. Preparing for occupation and emplacing howitzer	44
E. Setting up communications	24
F. Gunnery	99
G. Loading/unloading howitzer	32
H. Receiving and relaying communications	19
I. Recording/record keeping	29
J. Position improvement	14
	613
Motor Transport Operator (64C)	
A. Driving vehicles	158
B. Vehicle coupling	46
C. Checking and maintaining vehicles	181
D. Using maps/following paper routes	27
E. Loading cargo and transporting personnel	75
F. Parking and securing vehicles	32
G. Performing administrative duties	42
H. Self-recovering vehicles	20
I. Safety-mindedness	80
J. Performing dispatcher duties	15
	676
Administrative Specialist (71L)	
A. Preparing, typing, and proofreading documents	183
B. Distributing and dispatching incoming/outgoing documents	63
C. Maintaining office resources	73
D. Posting regulations	44
E. Establishing and/or maintaining files IAW TAFFS	50
F. Keeping records	94
G. Safeguarding and monitoring security of classified materials	43
H. Providing customer service	30
I. Preparing special reports, documents, drafts, and other materials	19
J. Sorting, routing and distributing incoming/outgoing mail	28
K. Maintaining Army Post Office equipment	2
L. Keeping Post Office records	20
M. Maintaining security of mail	9
	658
Military Police (95B)	
A. Traffic control and enforcement on post and in the field	63
B. Providing escort security and physical security	128
C. Making arrests, gathering information on criminal activity, and reporting on crimes	173
D. Patrolling and crime/accident prevention activities	236
E. Promoting confidence in the military police by maintaining personal and legal standards and through community service work	118
F. Using interpersonal communication (IPC) skills	87
G. Responding to medical emergencies and other emergencies of a non-criminal nature	50
	855
Infantryman (11B)	
A. Ensuring that all supplies and equipment are field-ready and available and well-maintained in the field	73
B. Providing leadership and/or taking charge in combat situations	33
C. Navigating and surviving in the field	53
D. Using weapons safely	38
E. Demonstrating proficiency in the use of all weapons, armaments, equipment and supplies	91
F. Maintaining sanitary conditions, personal hygiene, and personal safety in the field	24
G. Preparing a fighting position	29
H. Avoiding enemy detection during movement and in established defensive positions	22
I. Operating a radio	27
J. Performing reconnaissance and patrol activities	37
K. Performing guard and security duties	75
L. Demonstrating courage and proviciency in engaging the enemy	5
M. Guarding the processing POWs and enemy casualties	15
	522

(Continued)

Table III.18 (Continued)

Behavioral Examples Reliably Retranslated Into Each Dimension on the BARS Measures

Dimension	Number of Examples
Armor Crewman (19E)	
A. Maintaining tank hull/suspension system and associated equipment	123
B. Maintaining tank turret system/ fire control system	37
C. Driving/recovering tanks	80
D. Stowing and handling ammunition	39
E. Loading/unloading guns	30
F. Maintaining guns	43
G. Engaging targets with tank guns	45
H. Operating and maintaining communication equipment	36
I. Establishing security in the field	33
J. Navigating	11
K. Preparing/securing tank	27
	504
Radio Teletype Operator (31C)	
A. Inspecting equipment and trouble-shooting problems	50
B. Pulling preventative maintenance and servicing equipment	79
C. Installing and preparing equipment for operation	162
D. Operating communications devices and providing for an accurate and timely flow of information	142
E. Preparing reports	33
F. Maintaining security of equipment and information	57
G. Locating and providing safe transport of equipment to sites	50
	575
Light-Wheel Vehicle Mechanic (63B)	
A. Inspecting, testing, and detecting problems with equipment	47
B. Troubleshooting	63
C. Performing routine maintenance	23
D. Repair	101
E. Using tools and test equipment	68
F. Using technical documentation	56
G. Vehicle and equipment operation	18
H. Recovery	36
I. Planning/organizing jobs	15
J. Administrative duties	41
K. Safety mindedness	89
	557
Medical Specialist (91A)	
A. Maintaining and operating Army vehicles	51
B. Maintaining accountability of medical supplies and equipment	28
C. Keeping medical records	31
D. Attending to patients' concerns	15
E. Providing accurate diagnoses in a clinic, hospital, or field setting	11
F. Arranging for transportation and/or transporting injured personnel	44
G. Dispensing medications	42
H. Preparing and inspecting field site or clinic facilities in the field	34
I. Providing routine and ongoing patient care	95
J. Responding to emergency situations	142
K. Providing instruction to Army personnel	18
	511

A. TRAFFIC CONTROL AND ENFORCEMENT

Controlling traffic and enforcing traffic laws and parking rules.

1 2	3 4 5	6 7
• Often uses hand/arm signals that are difficult to understand, at times resulting in unnecessary accidents; often fails to wear reflectorized gear; overlooks hazardous traffic conditions; sleeps on duty; pays excessive attention to things unrelated to the job.	• Usually does a reasonable job when directing traffic by using adequate hand/arm signals and/or wearing reflectorized gear.	• Consistently uses appropriate hand/arm signals; always wears reflectorized gear; generally monitors traffic from plain-view vantage points; consistently refrains from behaviors such as reading and prolonged conversation on non-job related topics.
• May display excess leniency or harshness when citing offenders, allowing their military rank, race, and/or sex to influence his/her actions; makes many errors when filling out citations.	• Makes few errors when filling out citations; usually does not allow an offender's race, sex, and/or military rank to interfere with good judgment.	• Always uses emergency equipment (e.g., flares, barricades) to highlight unsafe conditions and ensures that hazards are removed or otherwise taken care of.

Figure III.7. Sample Behavioral Summary Rating Scale for Military Police (95B).

It is important to note that for each MOS we developed Behavioral Summary Scales. Traditional behaviorally anchored rating scales contain specific examples of job behaviors for each effectiveness level in a performance dimension. Behavioral Summary Scales, on the other hand, contain anchors that represent the behavioral content of ALL performance incidents reliably retranslated for that particular level of effectiveness. This makes it more likely that a rater using the scales will be able to match observed performance with performance on the rating scale (Borman, 1979).

After developing the performance rating scales for each MOS, we submitted the scales to review by a project research staff member familiar with the development process. Results from this review were used to clarify performance definitions and behavioral anchors. The final set of performance rating scales administered in field test sessions are included in Section 4 of the MOS appendixes in the ARI Research Note in preparation.

Revisions After Retranslation

The categorization of the original critical incident pool produced a total of 93 initial performance dimensions with a range of 7-13 dimensions per MOS. Based on the retranslation results, a number of the original performance dimensions were redefined, omitted, or combined. From the original set, six were omitted and four were lost through combination. One of the omissions was due to the fact that too few critical incidents were retranslated into it by the judges. The other five were omitted because the factor represented tasks that were well beyond Skill Level 1 or were from a very specialized low-density "track" within the MOS (e.g., MOS 71L F5-Postal Clerk).

Field Test Versions of MOS-Specific BARS

In sum, the results from the retranslation exercises were used to evaluate and modify the performance dimension system that had been developed for each MOS. The final set of behaviorally anchored rating scales for the nine MOS for use in the field test contained from 6 to 12 performance dimensions. Each of the performance dimensions includes behavioral anchors describing ineffective, average, and effective performance. Raters were asked to use these anchors to evaluate ratees on a scale ranging from 1 (ineffective performance) to 7 (effective performance).

Before the rating scales were tried out in the field, one additional scale was constructed for each MOS rating booklet. On this scale raters are asked to evaluate an incumbent's overall performance across all MOS-specific performance dimensions. This final rating scale is virtually the same for all MOS; it includes three anchors depicting ineffective, average, and effective performance.

Rating scale booklets that provided raters with performance dimension titles, definitions, and behavioral anchors were assembled for each MOS. The rating booklets were designed so that raters could evaluate up to five ratees in each. The booklets do not include instructions for using the scales to make performance ratings; instead, oral instructions were given during the field test rating sessions.

The field test samples and procedures are described in Section 8. Field test results and subsequent modifications of the BARS are described in Section 11.

Section 5

DEVELOPMENT OF ARMY-WIDE RATING SCALES[1]

The principal objective for this part of Project A's criterion development work is to construct a set of critical incident-based rating scales that will assess the major performance factors in the Army-wide, or non-job-specific, portion of the total performance space. Another objective is to develop rating scales that focus on specific common tasks that all first-term soldiers are required to perform. The procedures for developing each of these two kinds of rating scales will be described in turn.

Development of Army-Wide Behavior Rating Scales

The development of the Army-wide behavior rating scales followed the same general procedure as for the MOS-specific BARS (described in Section 4) and those details will not be repeated here. What is presented below are the procedures and findings that are specific to the Army-wide scales.

Behavior Analysis Workshops and Procedures

Seventy-seven officers and NCOs participated in six 1-day workshops intended primarily to elicit behavioral examples of soldier effectiveness that were not MOS-specific. Table III.19 describes the workshop participant groups. A total of 1,315 behavioral examples were generated in the six workshops. Details relevant to this data collection appear in Table III.20.

Duplicate examples and examples that did not meet the criteria specified (e.g., the incident described the behavior of an NCO rather than a first-term soldier) were dropped from further consideration. The remaining 1,111 examples were edited to a common format and content analyzed by project staff to form preliminary dimensions of soldier effectiveness. Specifically, three researchers independently read each example and grouped together those examples that described similar behaviors. The sorted examples were then reviewed and the groupings were revised until each author arrived at a set of dimensions that were homogeneous with respect to their content.

After discussion among project staff and with a small group of officers and NCOs at Fort Benning, a consensus was reached on a set of 13 dimensions. These were then submitted to retranslation.

Retranslation of the Behavioral Examples

The retranslation task was divided into five parts, with each part requiring a judge to evaluate 216-225 behavioral examples. Judges were

---

[1]This section is based primarily on ARI Technical Report 716, Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program, Elaine D. Pulakos and Walter C. Borman (Eds.), and the supplementary ARI Research Note 87-22, which contains the report appendixes.

Table III.19

Participants in Behavioral Analysis Workshops for Army-Wide Rating Scales

Rank	n	Gender	n
NCO (N = 30)		NCO	
SP4	1	Male	28
E-5	5	Female	2
E-6	13		
E-7	11		
Officer (N = 47)		Officer	
First Lt.	3	Male	44
Captain	29	Female	3
Major	15		

Table III.20

Soldier Effectiveness Examples Generated for Army-Wide Behavior Rating Scales

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

provided with definitions of each of 13 dimensions to aid in the sorting, and with a 1-9 effectiveness scale (1 = extremely ineffective; 5 = adequate/average; 9 = extremely effective) to guide the effectiveness ratings. The retranslation materials, including all 1,111 edited behavioral examples, appear in Appendix B, ARI Research Note in preparation. Sixty-one officer and NCO judges completed retranslation ratings.

Retranslation Results

Table III.21 shows the number of behavioral examples reliably retranslated for each of the 13 dimensions. The criteria established for acceptance--greater than 50% agreement for the sorting of an incident into a single dimension, and a standard deviation of less than 2.0 for the distribution of judges' effectiveness ratings for one incident--left 870 of the 1,111 examples (78%) included for subsequent scale development work.

The results in Table III.21 were seen as satisfactory, in that sufficient numbers of reliably retranslated examples were available to develop behavioral definitions of each dimension. Two pairs of dimensions were combined, resulting in a total of 11 Army-wide dimensions. Leading Other Soldiers and Supporting Other Unit Members were combined to form Leading/Supporting; Attending to Detail and Maintaining Own Equipment were collapsed to form Maintaining Assigned Equipment. The two combinations seemed appropriate because of the conceptual similarity of each of the dimension pairs.

For each of the 11 dimensions, the reliably retranslated behavioral examples were then divided into three categories of effectiveness levels, and behavioral summary statements were written to capture the content of the specific examples at low (1-3.49), average (3.5-6.49), and high (6.5-9) performance levels. Development of the behavioral summary statements is the critical step in forming Behavior Summary Scales (Borman, 1979).

Additional Army-Wide Scales

In addition to the 11 Army-wide BARS, two summary rating scales were prepared. First, an overall effectiveness scale was developed to obtain overall judgments of a soldier's effectiveness based on all of the behavioral dimension ratings. Second, an NCO potential scale was developed to assess each soldier's likelihood of being an effective supervisor as an NCO.

Final List of Army-Wide Behavioral Rating Scales

The 11 Army-wide BARS that were retained plus the overall performance and NCO potential scales provided the following behavioral rating scales for the field test:

A. Technical Knowledge/Skill
B. Effort
C. Following Regulations and Orders
D. Integrity
E. Leadership

Table III.21

Behavioral Examples Reliably Retranslated[a] Into Each Dimension for Army-Wide Behavior Rating Scales

Dimensions	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[b]	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[b]	59
K. Developing own job and soldiering skills	40
L. Effectively leading and providing motivation to other soldiers[c]	71
M. Supporting other unit members[c]	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.

[b] These two dimensions were subsequently combined to form a Maintaining Assigned Equipment dimension.

[c] These two dimensions were subsequently combined to form a Leadership dimension.

F. Maintaining Assigned Equipment
G. Maintaining Living/Work Areas
H. Military Appearance
I. Physical Fitness
J. Self-Development
K. Self-Control
Overall Effectiveness
NCO Potential

Development of Army-Wide Common Task Dimensions

Rating scales covering the common task domain were developed from tasks appearing in the Skill Level 1 Common Task Soldier's Manual. Because this manual specifies tasks that all first-term soldiers are expected to be able to perform, it seemed an appropriate source of Army-wide common task dimensions.

To develop these dimensions, a senior staff member content analyzed the specific tasks contained in the manual (e.g., Read and Report Total Radiation Dose; Repair Field Wire) and identified 13 common task areas that appeared to reflect in summary form all of the specific tasks. Examples of common task areas are See: Estimating Range and Combat Techniques: Moving Under Direct Fire.

Ratings consisted of evaluating how well each ratee typically performed each task on a 7-point scale, from 1 = "Poor: does not meet standards and expectations for adequate performance in this task area" to 7 = "Excellent: exceeds standards and expectations for performance in this task area." In addition, raters were given the option of choosing a "0," indicating that they had not observed a soldier performing in the task area. The 13 common task dimensions are:

A. See: Identifying Threat (armored vehicles, aircraft)
B. See: Estimating Range
C. Communicate: Send a Radio Message
D. Navigate: Using a Map
E. Navigate: Navigating in the Field
F. Shoot: Performing Operator Maintenance Weapon (e.g., M16 rifle)
G. Shoot: Engaging Target With Weapon (e.g., M16)
H. Combat Techniques: Moving Under Direct Fire
I. Combat Techniques: Clearing Fields of Fire
J. Combat Techniques: Camouflaging Self and Equipment
K. Survive: Protecting Against NBC Attack
L. Survive: Performing First Aid on Self and Other Casualties
M. Survive: Knowing and Applying the Customs and Laws of War

Field Test Instruments

On the basis of the above development steps, the Army-wide BARS scales and the Common Task Rating Scales were deemed ready for field testing in the Batch A and Batch B MOS. Field test procedures are described in Section 8, and field test results and subsequent modifications to the instruments are described in Section 12.

Section 6

DEVELOPMENT OF THE COMBAT PERFORMANCE PREDICTION RATING SCALE[1]

This section describes the development of a combat performance prediction scale, designed to evaluate performance under degraded conditions and the increased confusion, workload, and uncertainty of a combat environment. Such conditions would be expected for many soldiers near a battle area, even though it is likely that only a small percentage of the total Army force will directly participate in combat. Clearly, a soldier's judged effectiveness in a combat environment represents a potentially important indicator of overall effectiveness (Sadacca & Campbell, 1985).

This scale, like the Army-wide rating scales, was intended to be appropriate for any MOS. It is the only criterion that specifically addresses combat performance for all Project A MOS. It is also the only instrument expressly designed to measure performance under adverse conditions.

In developing this rating scale, we recognized that this rating task may pose some unusual difficulties for raters. First, although raters may often observe soldiers in garrison/field performance, opportunities to observe performance under adverse conditions may have been limited. Second, the majority of peer and supervisor raters have never experienced combat, so they are being asked to predict how soldiers would perform in a situation that the raters themselves may not know first-hand.

Unlike the Army-wide rating scales, which are behavioral summary scales, the Combat Performance Prediction Scale takes the form of a summated scale. This type of instrument is a series of scaled items (critical incidents), each followed by a response format. The items represent the positive and the negative aspects of each behavioral dimension. Items are presented in random order (across dimensions) on the rating form to preclude a response-set bias (for either the dimension or the direction of the item).

A major consideration in selecting the summated format for the prediction scale was the expected high correlations, attributable primarily to method variance, between this scale and the Army-wide scales if similar formats had been used for the two types. This was of particular concern given the subjective nature of the judgments that raters would be asked to make on the prediction scale. Another consideration was that we felt it was more reasonable to ask raters how likely it was that the soldiers they were rating would perform a given act, than to ask them to predict whether or not these soldiers would actually perform the act at a particular performance level, under combat conditions. Summing across rating items (acts) yields a score that measures the rater's assessment of the probability of how the ratees would act under combat-like conditions.

---

[1]This section is based primarily on an unpublished manuscript, "Development of Combat Performance Prediction Scales," by Barry Riegelhaupt and Robert Sadacca.

Scale Development

Development of a Conceptual Framework

The starting point for our combat scale development work was to build a conceptual model of combat effectiveness. We began with a set of behaviors that were not directly related to task performance, but were related to the broader concept of individual effectiveness in combat. In particular, elements that would be potentially important contributors to organizational effectiveness in Army combat units were considered. From the Army's perspective, being a good combat soldier means performing tasks in a technically proficient manner and displaying such characteristics as motivation, personal discipline, and physical fitness that are valued Army-wide. Within this framework, there may be additional elements that contribute to a soldier's combat effectiveness in the unit. The initial step of developing a conceptual framework was seen as useful for guiding thinking during subsequent empirical work to identify and define all elements of the combat effectiveness domain.

The preliminary set of combat performance dimensions is shown in Figure III.8. They are the result of preliminary hypotheses about behaviors that might be important to combat effectiveness. They were developed from a review of relevant literature (Anderson, 1984, a,b,c; Brown & Jacobs, 1970; Fiedler & Anderson, 1983; Frost, Fiedler, & Anderson, 1983; Henriksen et al., 1980; Hollander, 1954, 1965; Kern, 1966; Sterling, 1984) and insights provided by combat veterans on the Project A staff.

While the conceptual framework was considered important to subsequent development, we also believed strongly that an empirical strategy should be used to examine the combat effectiveness domain. Accordingly, a variant of the critical incidents or behavioral analysis (Smith & Kendall, 1963) approach was employed to identify dimensions of combat effectiveness. The many behavioral examples emerging from this step were content analyzed, and then submitted to a retranslation and scaling procedure. Following field testing, the best items were selected and the scale to be used in Concurrent Validation was developed.

Critical Incident Workshops

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 combat effectiveness. Officer and NCO participants in critical incident workshops were asked to provide behavioral examples (positive and negative) relevant to first-term combat effectiveness; examples were to be appropriate for and applicable to any MOS.

Forty-six officers and NCOs participated in one of the four 1-day critical incident workshops. All participants were combat veterans, the large majority with experience in Vietnam. In each workshop, the leader, a member of the Project A research staff, first described Project A and

A. Esprit de corps

Ability and desire to foster a common spirit of devotion and enthusiasm among members of a group/unit; identification with group/unit goals; commitment to maintaining and enhancing the reputation of the unit.

B. Initiative/Flexibility

Ability and willingness to identify and seize the opportunity to create novel solutions to combat problems; reasoned acceptance of risk.

C. Intelligence/Common Sense

Ability to size up a situation accurately by using all available information; willingness to evaluate the opinions of experienced personnel before making decisions.

D. Commitment/Devotion/Responsibility

Willingness to sacrifice personal gain for the good of the unit and its members; devotion to accomplish one's duty; willingness to take responsibility for the safety of self and others, for the maintenance of weapons and equipment, etc.

E. Physical and Moral Courage

Ability to face danger with confidence and emotional stability.

F. Obedience/Allegiance to Superiors

Ability and willingness to obey orders, for example, to advance on enemy positions, to dig in, etc.

G. Tactical/Technical Knowledge

Ability to follow standard operating procedures; knowledge of and ability to coordinate weapons, ammunition, equipment, and personnel.

H. Psychological/Physical Effects of Combat

Reaction to stress associated with shooting and killing enemy soldiers, losing a team/unit leader, seeing others wounded or killed, waiting for orders between battles, uncertainty of the situation, etc.

I. Interpersonal Communications

Ability to interact with others on a one-to-one or group level.

J. Decisiveness

Ability to make decisions based often on limited, incomplete, and unreliable information.

K. Personal Example

Ability to set a good personal example for others.

Figure III.8. Preliminary set of combat performance dimensions.

explained how the prediction of combat performance was an integral part of the project. Participants were then led into a discussion of how combat effectiveness could be defined in terms of more specific dimensions--that is, what categories can be used to define what is meant by combat effectiveness?

The workshop leader next presented the preliminary set of dimensions (see Figure III.8) and discussed overlap, semantic differences, and possible additions. This approach permitted workshop participants to think about combat effectiveness from their own perspective, and then compare that with our notions. Perhaps the most important function served was to establish a context for the behavioral examples the participants would be writing.

The workshop leader then distributed the instructions on how to write behavioral examples. These materials had a modeling orientation showing participants improperly written examples and then these examples corrected to the proper form. After review of these materials, participants were asked to write a behavioral example, which was reviewed and corrected as needed by the workshop leaders. 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.

A total of 361 behavioral examples was generated in the four workshops (Table III.22). After duplicative examples and those that were specific to officers, MOS, or equipment were eliminated, 158 usable examples remained. Since some of these examples might be eliminated during subsequent scale development work, it was desirable to have a larger set of items available. A review of a set of examples that had been used in the Army-wide rating scale retranslation workshops revealed 73 that described behavior in a combat-type situation; most of them described effective or ineffective behavior under adverse conditions during training and field exercises. These examples were added to the 158 usable examples from the combat workshops. The distribution is shown in Table III.23.

The examples were edited to a common format and used to revise the preliminary list of dimensions of combat effectiveness. Three researchers independently read each example and grouped those that described similar behaviors. The examples were then reviewed and the groupings revised until the researchers arrived at a set of homogeneous behavior categories. The content analysis of the incidents resulted in a reduction of the number of dimensions from 11 to 8. The revised dimensions are shown in Figure III.9. Employing the eight dimensions and 231 behavioral examples, materials were developed for retranslation and scaling workshops.

Retranslation and Scaling Workshops

Retranslation provides a way of checking on the clarity of individual behavioral examples and of the dimension system. In retranslation, persons familiar with the target domain make two judgments about each example: (a) the dimension or category it belongs to based on its content, and (b) the level of effectiveness or ineffectiveness it reflects. Examples for which there is disagreement either on category membership or on effectiveness level may not be stated clearly, and may need to be revised or eliminated from

further consideration. Also, confusion between two or more content categories in the sorting of several examples may reflect poorly formed and/or defined aspects of the dimension system.

The retranslation task was performed by 16 officer and NCO judges, all of whom were combat veterans. Judges were provided with definitions of each dimension to aid in sorting behaviors, and a 1-9 effectiveness scale (1 = extremely ineffective; 5 = average effectiveness; and 9 = extremely effective) for use in rating the level of positive or negative performance.

Table III.22

Combat Performance Workshop Participants and Examples Generated

Workshop	Participants	Number of Examples Generated
1	11 Field Grade Officers	80
2	10 NCOs	32
3	15 Field Grade Officers	166
4	10 NCOs	83
Total	46	361

Table III.23

Number of Edited Examples of Combat Behavior

	Combat Workshops	Army-Wide Workshops	Total
Positive	96	42	138
Negative	62	31	93
Total	158	73	231

A. Cohesion/Commitment to Others
- Ability and desire to foster a common spirit of devotion and enthusiasm among members of a group
- Concern for the physical/emotional welfare of the individual members of the group
- Commitment to maintaining/enhancing the effectiveness of the group

B. Intelligence/Common Sense
- Ability to learn quickly and apply the newly acquired knowledge/skill in a novel situation
- Ability to size up a situation and use available resources to make a decision
- The exercise of appropriate judgment

C. Self-Discipline/Responsibility
- Willingness to accept responsibility for the accomplishment of the task at hand
- Concern for conditions that jeopardize the safety of self and others
- Concern for the maintenance of weapons and equipment, etc.

D. Physical/Medical Condition
- Ability and willingness to maintain both physical and medical fitness
- Physical endurance as demonstrated by little or no reduction in performance even after or during prolonged or strenuous activities
- Concern for proper health care/hygiene to avoid sickness and disease

E. Mission Orientation
- Willingness to make sacrifices and endure hardships to accomplish mission
- Commitment and dedication to accomplishing one's assigned duties/responsibilities
- Willingness to accept a reasonable amount of risk in the pursuit of mission accomplishment

F. Technical/Tactical Knowledge
- Ability to follow SOP
- Knowledge of and ability to coordinate weapons, ammunition, and equipment
- Ability to perform-MOS specific and common soldiering tasks

G. Psychological Effects of Combat
- Reaction to stress associated with shooting and killing, losing a unit/team leader, seeing others wounded or killed, waiting for orders between engagements, etc.
- Ability to perform duties with little or no decrement under emotionally stressful situations

H. Initiative
- Ability and willingness to take the appropriate action at the appropriate time without being told to do so

Figure III.9 Revised set of combat performance dimensions.

Acceptable agreement was defined as greater than 50% of the judges sorting an example into the same dimension. Of the 231 examples, 108 did not meet this criterion and were placed in an "Other" category (Table III.24).

Table III.24

Agreement[a] by Discriminability Item Distribution

Dimension	t-Value				Total
	5.5 & Below	5.6-7.0	7.1-9.0	9.1 & Above	
Cohesion/Commitment	12	4	9	9	34
Intelligence/Common Sense	1	0	0	3	4
Self-Discipline/Responsibility	3	3	3	16	25
Physical/Medical Condition	2	1	1	2	6
Mission Orientation	1	6	7	5	19
Technical/Tactical Knowledge	4	3	2	4	13
Psychological Effects	4	3	2	0	9
Initiative	1	4	4	4	13
Other[b]	32	24	28	24	108
Total	60 (26%)	48 (21%)	56 (25%)	67 (27%)	231

[a] Greater than 50% agreement among judges in placing items in dimensions.

[b] Items not reliably retranslated into the eight dimensions.

The same group of judges performed a scaling task that provides a way of determining which examples discriminate between "best" and "worst" performers. Each judge was asked to make two ratings. For one rating, judges were asked to think of the "best" soldier they had ever worked with in combat and to decide how likely it was that that soldier would have behaved like the soldier in each example. For the other rating, they performed the same task, but this time considered the "worst" soldier they had ever worked with. Half of the 16 judges rated the "best" soldier first and half rated the "worst" soldier first, using a 15-point scale ranging from very unlikely (1) to very likely (15).

A discriminability index was calculated by computing a dependent t-value for each of the 231 items. The t-value is a measure of the statistical significance of the difference in mean probability assigned by the raters to their "best" soldier performing the act described in the item versus that assigned their "worst" soldier performing the act. A t-value equal to or greater than 2.95 would be significant at the $p < .01$ level.

As shown in Table III.24, the 231 items were divided roughly into quartiles on the basis of t-values, as an aid in selecting items that had both high discriminability and high agreement. A total of 171 items had t-values of 5.6 or greater. Thus, a sufficient number of items discriminated "best" from "worst" combat soldiers at a very high level of statistical significance. However, 76 of those examples had been assigned to the Other category after retranslation because they had not been reliably retranslated into dimensions. Additionally, the categories of Intelligence/Common Sense, Physical/Medical Condition, and Psychological Effects contained very few items.

A factor analysis (unweighted least squares, with a Promax rotation) was performed to attempt to reduce the number of dimensions. The results provided guidance on how to combine dimensions. Examples placed in the Intelligence/Common Sense category were reassigned to either Self-Discipline/Responsibility or Technical/Tactical Knowledge based upon the frequency of judges' placement of the items. Items from Physical/Medical Condition were combined with items in the Self-Discipline/Responsibility category. Behavioral examples of Psychological Effects were placed in the Mission Orientation category.

In developing the final form of the Combat Performance Rating Scale, the goal was to select items that reflected good performance and poor performance to represent the domain of combat effectiveness. In a summated scale, the most important criterion is the items' ability to discriminate between performance extremes. Consequently, reducing the agreement criterion for dimensional agreement among judges in order to increase the number of items does not violate good construction practice for a summated scale. To make sure that we considered a maximum set of discriminating items, we redefined the dimension agreement criterion (initially "greater than 50%") to "equal to or greater than 50%").

The agreement by discriminability item distribution for the reduced dimensional set and redefined agreement criterion is shown in Table III.25. It should be noted that at this point five items were viewed as too sensitive and were deleted from further consideration. Following these changes, 113 items had t-values of 5.6 or greater and were reliably retranslated into one of the five dimensions. This represented the item pool from which the items were selected for further development of the combat prediction rating scale.

Item Selection

Selection of items for the field test version of the scale was to be based primarily on discriminability, with consideration also given to dimension agreement, and with an approximate balance between positive and negative examples. Allowing for time constraints in testing, and eliminating poor items, the goal was to select 80 items--the 16 best discriminating items from each of the five dimensions. However, when items were rank ordered on the basis of t-values within each dimension by positive and negative items, some t-values were too low to allow the item to be included. Also, the Initiative dimension contained only 13 items. Therefore, in addition to the five dimensions, items from the Other category were selected for inclusion in

the scale. The items selected for field testing are shown in Table III.26. Including the Other category resulted in a more balanced coverage of the dimensions and of the positive/negative split, and a set of items all having t-values greater than 5.6.

Table III.25

Combat Prediction Agreement[a] by Discriminability Item Distribution for Reduced Dimensional Set and Redefined Agreement Criteria

	t-Value				
Dimension	5.5 & Below	5.6-7.0	7.1-9.0	9.1 & Above	Total
Cohesion/Commitment	14	4	9	10	37
Self-Discipline/Responsibility	8	4	5	25	42
Mission Orientation	9	12	11	8	40
Technical/Tactical Knowledge	8	6	5	3	22
Initiative	2	3	4	4	13
Other[b]	21	16	22	13	72
Total	62 (27%)	45 (20%)	56 (25%)	63 (28%)	226

[a] Equal to or greater than 50% agreement among judges.

[b] Items not reliably retranslated into one of the five dimensions.

Table III.26

Items Selected for Field Test of Combat Performance Prediction Scale

Dimension	Positive	Negative	Total
Cohesion/Commitment	12	3	15
Self-Discipline/Responsibility	6	10	16
Mission Orientation	8	7	15
Technical/Tactical Knowledge	4	7	11
Initiative	10	0	10
Other	9	4	13
Total	49	31	80

To reduce the administrative burden on any one rater, two forms (Form A and Form B) were developed. Each contained 60 items--40 common to both forms and 20 unique to each form.

Review and Rescaling

The two proposed 60-item forms of the Combat Performance Prediction Scale were reviewed by three company grade Army officers and three ARI scientists. As a result of that review, three items common to both forms were deleted and a large proportion of the remaining 77 items were reworded. The rewording was extensive enough to render questionable the discriminability indexes previously computed for each item. Therefore, the 77 items were subjected to a rescaling.

The rescaling workshop was conducted using the same procedures as for the original scaling. Eight officers and one civilian (seven of the nine were combat veterans) made the "best" and "worst" combat soldier ratings for each of the 77 items. New t-values were computed for each item. In general, the rescaled values were of lower statistical significance than the original t-values. However, in only one case did the rescaled item result in a non-significant t-value. This item (in Form A) was deleted.

Field Test Version of Combat Effectiveness Prediction Scale

For the field test, Part I of Form A contained 56 items and Form B contained 57 items. In Part II of both forms, raters were asked to respond to three additional questions: how confident they were, overall, in the ratings they had just completed; how many of the items made sense to them; and which items least applied to the soldiers whom they had just rated.

The field test version of the Combat Effectiveness Prediction Scale thus consisted of 76 items split between two forms, with 3 additional items designed to capture reactions to the scale itself. The instructions for raters and two sample items from the field version are shown in Figure III.10. Because the development of this scale followed a different schedule than the other criterion measures, it was field tested with only a portion of the Batch B sample. The field test results are reported in Section 14 of this report.

COMBAT PERFORMANCE PREDICTION SCALE

INSTRUCTIONS

On the following pages, you will find examples that describe activities of soldiers in combat. Assume that the soldiers you are rating were placed in the combat situation described and had the opportunity to behave as the soldier in each example behaved. Then, using the scale shown below, indicate the likelihood of each soldier you are rating performing as the soldier in the example performed.

Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely		
○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Please darken the circle under the point on the scale that gives the likelihood that each soldier would behave in the way described in the example. For example, if you think that there was absolutely no chance that the soldier you are rating would do what the soldier in the example did, then you would darken the first circle under the Very Unlikely part of the scale. If you think that the soldier you are rating would absolutely certainly do what the soldier in the example did, then darken the last circle under the Very Likely part of the scale. If the likelihood is between the two extremes, darken the appropriate circle.

Please evaluate each soldier's likelihood of doing every activity. Do not leave any blanks.

COMBAT PERFORMANCE PREDICTION RATING SCALE ITEMS

1. This soldier volunteered to lead a team to an accident scene where immediate first aid was required before an order was given.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely		
Line up the names	1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
of the soldiers	2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
you are rating	3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
with the rows	4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
to the right.	5	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

2. Near the end of a movement, when soldiers were ordered to prepare fighting positions, this soldier prepared his position quickly and then assisted other squad members.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely		
Line up the names	1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
of the soldiers	2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
you are rating	3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
with the rows	4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
to the right.	5	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Figure III.10. Sample of Combat Performance Prediction Scale

Section 7

ADMINISTRATIVE/ARCHIVAL RECORDS AS ARMY-WIDE PERFORMANCE MEASURES[1]

A major activity within the overall program of performance criterion development is to explore the use of the archival administration records as first-tour job performance criteria and in-service predictors of soldier effectiveness. The Enlisted Master File (EMF), the Official Military Personnel File (OMPF), and the Military Personnel Records Jacket (MPRJ) are the Army records sources that contain administrative actions that could be used to form measures of first-tour soldier effectiveness.

A serious difficulty in using administrative records for evaluation purposes is that the material in the records very often reflects only exceptionally good or exceptionally poor performance. Measures of performance based on personnel actions that appear infrequently could have very little variance. A strategy for dealing with the skewness and lack of variability in records data that result from low base rates is to combine records of different kinds of events and actions into more general indexes. When scores on administrative measures that reflect the same underlying constructs are combined, the base rate might improve to a level where significantly higher correlations with other variables would be possible. Accordingly, project staff undertook a detailed examination of the three archival data sources and an analysis of the feasibility of developing first-tour and in-service predictors from them.

Identification of Administrative Indexes

A preliminary 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 III.27. A description of the detailed investigation into each of the three records sources follows.

Enlisted Master File (EMF)

The EMF is an automated inventory of personal data, enlistment conditions, and military experience for every enlisted individual currently on the U.S. Army payroll. It contains a large number of variables for each individual, ranging from pay grade to Skill Qualification Test (SQT) scores to the Army's operational performance appraisal ratings in the form of the Enlisted Efficiency Report (EER).

---

[1]This section is based primarily on an ARI Technical Report 754, The Development of Administrative Measures As Indicators of Soldier Effectiveness, by Barry J. Riegelhaupt, Carolyn DeMeyer Harris, and Robert Sadacca.

Table III.27

Preliminary List of Administrative Measures Indicative of Soldier Effectiveness

- Reason for Separation From the Army
- Reenlistment Eligibility
- Reenlistment Eligibility Bar
- Enlisted Evaluation Report (EER)
- Promotion Rate
- Number and Duration of AWOL/Desertions
- Number and Type of Articles 15
- Number and Type of Courts-Martial
- Number and Type of Awards/Badges
- Number and Type of Letters of Appreciation/Commendation
- Number and Type of Letters of Reprimand/Admonition
- Number and Type 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

An initial examination of the EMF identified four variables as potentially useful: (a) reason for separation, (b) reenlistment eligibility, (c) reenlistment eligibility bar, and (d) EER score.

In theory, the EER, which is a weighted average of a soldier's last five performance ratings, should be a very useful variable. As a practical matter, however, for Project A purposes its value may be limited. EER ratings are obtained only for soldiers in grades E5 and above, so not more than a small percentage of first-tour enlisted personnel is likely to have had even one EER at the time of the Project A data collection. Also, for understandable reasons, EER ratings have tended to cluster near the maximum score.

Information relevant to two additional variables is available from the EMF. First, it is possible to compute a promotion rate, defined as grades advanced per year, for each soldier. Second, while neither the number of times an individual has been AWOL nor the duration of each AWOL is available from the EMF, it is possible to assign soldiers to the dichotomous variable, "Has or Has Never Been AWOL."

Information on awards, badges, letters and certificates of appreciation, achievement, and commendation, Articles 15, and so forth is not contained on the EMF. Information of this type exists only in the individual soldier's Official Military Personnel File (OMPF) or Military Personnel Records Jacket (MPRJ)

Official Military Personnel File (OMPF)

The OMPF is the permanent, historical, and official record of a member's military service. The information for enlisted personnel is maintained on microfiche records located at the Enlisted Records and Evaluation Center (EREC), Fort Benjamin Harrison, Indiana.

Depending upon their purpose, documents are filed in one of three sections:

- The performance (P) fiche - the portion of the OMPF where performance, commendatory, and disciplinary data are filed.
- The service (S) fiche - the OMPF section where general information and service data are filed.
- The restricted (R) fiche - the OMPF section for historical data that may be biased against the soldier when viewed by selection boards or career managers. For this reason release of information on this fiche is controlled.

The usefulness of the microfiche records for project purposes was examined systematically via a pilot study.

Sample Selection. A random sample of 25 enlisted personnel from each of the 19 MOS being studied in Project A was selected from the FY82 Enlisted Master File. The list of 475 names and corresponding social security numbers (SSN) was forwarded to the Enlisted Records and Evaluation Center, with a request to have the 475 records available for a project data collection team that would examine them on site.

Data Collection and Analysis. A data collection form for recording the administrative measures listed in Table III.27 was developed. Upon arrival at Fort Benjamin Harrison, the data collection team was handed 414 microfiche packets. This represented 89% of the 466 packets that EREC personnel attempted to locate (nine names had been omitted in the transmission of the request). Each of the microfiche records in the packets was examined by a staff member and information was entered on the records collection form.

After examining the microfiche and the regulations governing their composition, as well as interviewing knowledgeable officials, the team reached a number of conclusions, which are expressed below in terms of optimal and actual outcomes:

Optimal Outcomes -

(1) Performance data for 475 soldiers would be available.

(2) All 475 soldiers would be new, first-time soldiers in FY81.

(3) No Enlisted Evaluation Reports (EER) would be found.

(4) All authorized documents would appear on microfiche.

(5) Recorded information would be timely.

Actual Outcomes -

(1) Performance data were available for only 136 soldiers--29% of the projected sample.

(2) Of the 136 soldiers for whom performance information was available, 44 (32%) were prior service members.

(3) Since it had been assumed that the sample was comprised of new, first-term soldiers, individuals would not have been in the Army long enough to have had an EER. However, 26 EERs were found among the records for 20 soldiers, all of whom were prior service members.

(4) While many documents are authorized to appear in the OMPF performance section, a recent change to Army Regulation 640-10 requires written filing instructions if certain documents are to be entered. For example, a letter of commendation will not routinely be forwarded for filming; it will be sent to EREC only if it is specifically directed to the OMPF.

    Thus, it is possible for soldiers to have a number of documents in their Military Personnel Records Jacket that are authorized to appear on OMPF microfiche, but that may not be there because they were not directed to the OMPF.

(5) For grades below E5 (the grade levels of enlisted personnel in the first major Project A data collection), there is a backlog of 8 to 12 months from the time a personnel action is taken until it appears on microfiche at EREC. The primary reason for this backlog is that, for the grades E5 and above, microfiche are used by central promotion boards. Documents submitted for filming for these individuals take precedence over documents received for soldiers below the grade of E5.

Because of these aspects of the microfiche records, the next step was to determine the feasibility of developing criterion indexes from the Military Personnel Records Jacket, known as the 201 File.

Military Personnel Records Jacket (MPRJ)

The MPRJ, or 201 File, is the primary mechanism for storing information about an individual's service record. Updates/additions/corrections to the file are made at the time of the action. The MPRJ physically follows the individual wherever he or she goes and is normally located at the Military Personnel Office (MILPO) that serves the soldier's unit.

The feasibility of using data from the 201 File for Project A evaluations was examined in much the same way as for the microfiche records. To develop a data collection form that could be used to record information from 201 Files, detailed reviews of relevant Army Regulations and interviews with knowledgeable Army personnel were conducted. An expanded list of potential indexes was compiled (Table III.29) and a records collection form was developed for use in a pilot study.

Table III.28

Expanded List of Administrative Measures Indicative of Soldier Effectiveness

- Comparison of Skill Level of Primary to Duty MOS
- Existence of Secondary MOS
- Existence of Skills Qualification Identifier (SQI)
- Existence of Additional Skill Area (ASI)
- Existence of Language Identifier (LI)
- Record of Skill Qualification Test (SQT) Score Within Past 12 Months
- Type of Reenlistment Eligibility
- Type of Military Education Leadership Course
- Level of Highest Civilian Education
- Promotion Rate
- Existence of Promotion Packet at E4
- Number and Type of Awards/Badges
- Record of Requalification Weapons Score Within Past 12 Months
- Number and Type of Certificates of Achievement/ Appreciation/Commendation
- Number and Type of Letters of Appreciation/Commendation
- Number and Type of Letters of Reprimand/Admonition
- Number of Additional Military Training Courses Completed
- Number and Type of Correspondence Courses Completed
- Number of Additional Civilian Education Classes Completed
- Course Summary and Abilities Ratings - Service School
- Professional Competence and Standards Ratings and Summary Score of Enlisted Efficiency Report
- Type, Sentence, Suspension, Vacation of Court-Martial
- Existence of Court-Martial Proceedings in Action Pending
- Reason for Bar to Reenlistment
- Number and Duration of AWOL
- Number of Violations and Reason for Article 15
- Reason for FLAG Action
- Number of and Reason for Disposition - Block to Promotion

Sample Selection. The plan was to collect records data from the MPRJ for a sample of 750 soldiers, 150 in each of five MOS at five Army posts. To achieve this sample size while allowing for unavailability of some records, the records of 200 soldiers at each post were requested.

To increase the likelihood that findings from the records collection could be generalized, MOS choice was based on diversity. Each MOS represented a different Career Management Field (CMF), a different ASVAB area composite, and a different cluster where "clusters" refer to the job groupings derived from the Project A MOS clustering (Rosse, Borman, Campbell, & Osborn, 1983). The selected MOS and the corresponding incumbent populations are shown in Table III.29.

Table III.29

MOS x Post Populations in Study of Military Personnel Records Jackets

Post	MOS[a] 05C	11B	64C	71L	91B	Total
A	42	149	111	108	98	508
B	182	505	199	252	207	1,345
C	125	193	198	226	165	907
D	53	359	112	91	73	688
E	56	196	134	74	82	542
Total	458	1,402	754	751	625	3,990

[a]MOS: 05C Radio TT Operator
11B Infantryman
64C Motor Transport Operator
71L Administrative Specialist
91B Medical Care Specialist

Data Collection Procedure. Data were collected by teams of two research staff members in 2-day visits to each of five posts. Table III.30 indicates the number of MPRJs from which data were collected at each post.

Table III.30

Number of Military Personnel Records Jackets Requested and Received at Each Post

Post	Number of MPRJ Requested	Received	Percent Received
A	200	133	67
B	200	153	77
C	200	156	78
D	200	159	80
E	200	146	73
Total	1,000	747	75

Frequency distributions were generated for each data field. Based upon these frequencies, a set of 38 variables was created (Table III.31). Variables 11-13 and 21-24 were created based upon the model of soldier effectiveness dimensions that had been previously developed (Borman et al., 1985a). This research identified the following performance dimensions as relevant to all soldiers, regardless of their MOS:

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

B. Adhering to regulations, orders, and 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/mission/assignment

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.

Specifically, in addition to counting the number of Articles 15 that a soldier received, for example, we recorded the reason for the disciplinary action and mapped these reasons onto the model's dimensions. This allowed for the creation of variables based on the content of administrative actions as well as on a count of those actions. This was consistent with the Project A construct validation approach.

The original request for 1,000 MPRJs specified a Basic Active Service Date (BASD) window of 17 months. At this point, the 17-month window was reduced to 13 months to more accurately reflect the time that soldiers in the actual FY83/84 cohort first-tour data collection would be in the service. Only those soldiers who entered the Army between 1 July 1981 - 31 July 1982 at an initial grade of PFC or less were retained. 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.

Table III.31

List of Created Variables in Study of Administrative Measures

Variable Number	Description
01	Has SQI, ASI, or Language Identifier
02*	Is working at skill level DMOS higher/lower than PMOS
03	Is eligible to reenlist
04*	Highest grade attained
05*	Current grade
06	Never demoted
07	Number of awards
08	M16 rating
09	Has EXP grenade rating
10	Number of letters/certificates
11	Cited for exhibiting technical knowledge and skill (Constructs H and J)[a]
12	Cited for physical and mental self development (Constructs E and K)[a]
13	Cited for constructs other than E, H, J, and K[a]
14*	Has had special military education
15	Number of military training courses
16*	Years of civilian education
17*	Has high school diploma
18*	Has earned civilian education credits
19	Number of Articles 15/FLAG actions
20	Has been AWOL
21	Cited for failure to adhere to rules and regulations and disrespect for authority (Construct B)[a]
22	Cited for failure to control own behavior (Construct A)[a]
23	Cited for Construct violations other than Constructs A and B[a]
24	Number of times cited for construct violations (Variable 21 + 22 + 23)[a]
25	Number of times assigned extra duty
26	Has had punishment suspended
27	Has forfeited pay
28	Has been restricted
29	Has been confined
30*	Initial grade
31*	Change in grade (Variables 05, 30)
32*	Time period in years between first and last grade change
33	Promotion rate (number of grades advanced per year -- Variables 31/32)
34	Has received punishment
35	Has received Army Achievement Medal (AAM)
36	Has received air assault badge
37	Has received parachute badge
38	Has received other award

* Indicates an interim variable used only to define the actual variable. The interim variable was not used in subsequent analyses.

[a] See construct list in text. Construct definitions appear in Borman et al. (1987).

Comparison of Availability of Information

Military Personnel Records Jacket (MPRJ) - Official Military Personnel File (OMPF) Comparison. Using the records collection form developed to extract records data from the MPRJ, three research staff members spent 2 days collecting records data from the OMPFs of 292 soldiers. The 292 individuals represented a random sample of the 650 soldiers from whose MPRJs administrative records data had previously been collected. Thus, the amount of information available from the records sources could be compared.

The frequency distributions of selected administrative variables available from the MPRJ and the OMPF are compared in Table III.32. As can be seen, the MPRJ was found to be a much richer source of information on the administrative actions of interest in Project A. In the extreme case, information relevant to a soldier's reenlistment eligibility was not even available from the OMPF.

Military Personnel Records Jacket (MPRJ) - Enlisted Master File (EMF) Comparison. Presented in Table III.33 are frequency distributions of selected variables collected from the MPRJ that are also available from the EMF. As can be seen, unlike the MPRJ-OMPF comparison, a rather high degree of correspondence exists between the MPRJ and the EMF. It should be noted that the EMF was an FY83 end-of-year tape. The MPRJ data were collected during the second and third weeks in October 1983. Thus, the MPRJ information was being compared to EMF entries that were, at most, 3 weeks behind the information in the field. Even in light of the 3-week difference, the correspondence between sources is impressive and highlights the benefits of having current EMF information available.

Results of Analysis of MPRJ Data

Analyses were conducted in two stages:

(1) Identification of administrative variables potentially useful in Project A measures

(2) Examination of the relationships of the identified variables with selected nonadministrative variables (e.g., Post, MOS, Moral Waiver

Variable Selection

A first step in determining the usefulness for Project A purposes of the administrative variables collected from MPRJs (201 Files) was to select those measures with an acceptable amount of variance. The frequency distributions for each administrative measure are presented in Table III.34. The product moment correlations among the administrative variables are presented in Table III.35.

Table III.32

Frequency Distributions for Selected Variables in MPRJ-OMPF Comparison (N = 292 soldiers)

Variable	Category	MPRJ (201 File)	OMPF (Microfiche)
Number of Letters/Certificates	0	218	287
	1	45	4
	2 or More	29	1
Number of Awards	0	209	262
	1	69	27
	2 or More	14	3
Has Received Article 15	No	258	278
	Yes	34	14
Has Been AWOL	No	286	290
	Yes	6	2
Has Had Special Military Education	No	270	288
	Yes	22	4
Is Eligible to Reenlist	Blank	41	292
	No	29	--
	Yes	222	--
Highest Grade Attained	PV1	1	237
	PV2	13	20
	PFC	156	17
	SP4/CPL	116	18
	SP5/SGT	1	--
	SP6/SSG	5	--
Change in Grade	-1	1	--
	0	19	278
	1	56	3
	2	135	2
	3	77	9
	4	2	--
	5	2	--

Table III.33

Frequency Distributions for Selected Variables in MPRJ/EMF Comparison (N = 650 soldiers)

Variable	Category	MPRJ (201 File)	EMF (FY83 End)
Has Been AWOL	No	631	633
	Yes	19	17
Has Had Special Military Education	No	620	623
	Yes	30	27
Is Eligible to Reenlist	Blank	76	71
	No	57	52
	Yes	517	527
Initial Grade	Blank	1	2
	PV1	497	516
	PV2	76	68
	PFC	76	64
Current Grade	PV1	13	7
	PV2	32	14
	PFC	309	341
	SP4/CPL	290	282
	SP5/SGT	6	6
Promotion Rate	0	40	41
	1	136	112
	2	375	401
	3	98	96
	4	1	0

Based upon the information presented in Tables III.34 and III.35, and the regulations governing reenlistment and promotion criteria, six variables were selected as potentially useful criteria and in-service predictors for Project A. The six measures were:

- Eligible to Reenlist
- Number of Letters/Certificates
- Number of Awards

Table III.34

Frequency and Percentage Distributions for Administrative Variables

Variable Number	Variable	Category	Frequency	Percent
01	Has SQI/ASI/LI	No	518	79.7
		Yes	132	20.3
03	Is Eligible to Reenlist	Blank	76	-
		No	57	9.9
		Yes	517	90.1
06	Never Demoted	No	25	3.9
		Yes	625	96.1
07	Number of Awards	0	436	67.1
		1	169	26.0
		2 or more	37	6.9
08	M16 Rating	Blank	37	-
		MKM	290	47.3
		SP5	183	29.9
		EXP	140	22.8
09	Has EXP Grenade Rating	No	490	75.3
		Yes	160	24.6
10	Number of Letters/Certificates	0	461	70.9
		1	113	17.4
		2 or more	76	11.7
11	Cited for Technical Knowledge and Skill (Constructs H and J)	0	525	80.8
		1	83	12.8
		2 or more	42	6.5
12	Cited for Physical and Mental Self-Development (Constructs E and K)	0	609	93.7
		1 or more	41	6.3
13	Cited for Constructs Other Than E, H, J, and K	0	582	89.5
		1 or more	68	10.5
15	Number of Military Training Courses	0	484	74.5
		1	128	19.7
		2 or more	38	5.9
19	Has Received Article 15/FLAG Action	No	576	88.6
		Yes	74	11.4
20	Has Been AWOL	No	631	97.1
		Yes	19	2.9

(Continued)

Table III.34 (Continued)

Frequency and Percentage Distributions for Administrative Variables

Variable Number	Variable	Category	Frequency	Percent
22	Cited for Failure to Control Own Behavior (Construct A)	0	620	95.4
		1 or more	30	4.6
23	Cited for Construct Violations other than A and B	0	625	96.1
		1 or more	25	3.9
24	Number of Times Cited for Construct Violations	0	554	85.2
		1	61	9.4
		2 or more	35	5.4
25	Has Received Extra Duty	No	595	91.5
		Yes	55	8.5
26	Has Had Punishment Suspended	No	611	94.0
		Yes	39	6.0
27	Has Forfeited Pay	No	583	89.7
		Yes	67	10.3
28	Has Been Restricted	No	610	93.9
		Yes	40	6.1
29	Has Been Confined	No	638	98.1
		Yes	12	1.9
33	Promotion Rate (Grades Advanced/Year)	0	40	6.1
		1	136	20.9
		2	375	57.7
		3	98	15.1
		4	1	.1
34	Has Received Punishment	No	574	88.3
		Yes	76	11.7
35	Has Received AAM	No	582	89.5
		Yes	68	10.5
36	Has Received Air Assault Badge	No	618	95.1
		Yes	32	4.9
37	Has Received Parachute Badge	No	559	86.0
		Yes	91	14.0
38	Has Received Other Award	No	584	89.95
		Yes	66	10.1

Table III.35

Means, Standard Deviations, and Correlation Coefficients of Administrative Variables

Var. No.	Variables	Mean	SD	Correlational Coefficient[a]																										
				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
01	Has SQI/ASI/LI	.20	.40	-																										
03	Eligible to Reenlist	.90	.30		-																									
06	Never Demoted	.96	.19		.30	-																								
07	Number of Awards	.40	.62	.45	.08	.65	-																							
08	MI6 Rating	1.76	.80	.23		.30	.20	-																						
09	EIP Grenade Rating	.25	.43	.31			.25	.08	-																					
10	No. of Letters/Certificates	.41	.69				.18			-																				
11	Cited: Tech Knowl. & Skill	.26	.57				.17			.80	-																			
12	Cited: Phys & Ment Self Deve	.06	.24				.09			.44	.13	-																		
13	Cited: Other Constructs	.10	.31							.58	.18	.20	-																	
15	Military Training Courses	.31	.58	.49			.51	.18	.22	.11	.08	.10		-																
19	Received Article 15/FLAG	.11	.32		-.45	-.46	-.10								-															
20	Has Been AWOL	.03	.17		-.32	-.30									.46	-														
21	Cited: Failure Adhere to Regs	.06	.24		-.27	-.34									.71	.18	-													
22	Cited: Failure to Control Behavior	.05	.21		-.17	-.15									.38		.15	-												
23	Cited: Other Construct Violation	.04	.19		-.30	-.08			-						.23				-											
24	No. Times Cited: Construct Violation	.20	.52		-.50	-.39	-.09								.84	.39	.71	.54	.45	-										
25	Received Extra Duty	.08	.28		-.39	-.37									.80	.41	.55	.28	.20	.66	-									
26	Had Punishment Suspended	.06	.24		-.30	-.22									.64	.23	.54	.38	.15	.59	.55	-								
27	Forfeited Pay	.10	.30		-.42	-.38	-.08			-.08			-.08		.87	.45	.65	.38	.20	.78	.77	.66	-							
28	Been Restricted	.06	.24		-.31	-.35							-.09		.69	.37	.51	.25	.18	.60	.75	.42	.67	-						
29	Been Confined	.02	.13		-.34	-.21									.31	.18	.20	.19	.15	.39		.25	.37		-					
33	Promotion Rate	1.82	.76		.16	.36									-.22	-.16	-.20			-.19	.18		-.19	-.20	-.12	-				
34	Received Punishment	.12	.32		-.46	-.45									.94	.45	.70	.40	.23	.84	.84	.69	.93	.70	.38	-.22	-			
35	Received AAM	.10	.31				.53			.10	.10																	-		
36	Received Air Assault Badge	.05	.22				.25							.23	-.08													-.08	-	
37	Received Parachute Badge	.14	.35	.74			.61	.22	.33	.10		.08		.61																-
38	Received Other Award	.10	.30	.15	.08		.59	.11		.17	.15		.12	.17									-.08			.08		.08		.20	-

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

- Number of Military Training Courses
- Has Received Article 15/FLAG Action
- Promotion Rate (Grades Advanced/Year)

Based on the frequency distributions shown in Table III.34, the Number of Letters/Certificates, Number of Awards, and Number of Military Training Courses variables were transformed into dichotomous variables--Has Received Letters/Certificates, Has Received Awards, and Has Had Military Training Courses.

Relationships of Administrative Measures With Other Variables

Each of the six administrative measures and a combined "Has Received Letter/Certificate/Award" variable were subjected to a series of analyses. These included an examination of MOS and Post differences; stepwise multiple regressions, in which AFQT, Moral Waiver, Sex, and Race were entered after controlling for Post and MOS effects; and univariate analyses, in the form of chi-square tests, for those variables entered into the regression equation with a significant F value at the time of first entry.

The findings from this analysis are summarized in Table III.36. The asterisked cells indicate which of the other available variables (Post, MOS, AFQT, Moral Waiver, Sex, and Race) were significantly related to each of the administrative measures in both the univariate and multivariate analyses. The pattern of significant and nonsignificant relationships found was encouraging.

First, there was no evidence that a soldier's race was a significant determiner of his/her Reenlistment Eligibility, Number of Awards, or any other of the Army-wide administrative measures. Second, although a soldier's sex was related to Awards (males received more) and to Letter/Certificate (females received more), when the two variables were combined into the Letter/Certificate/Award measure, sex differentials were no longer statistically significant.

Third, AFQT score or mental category was related to successfully completing Military Training Courses and to Number of Awards, indicating the possible usefulness of the ASVAB in predicting aspects of Army-wide performance. Fourth, both Reenlistment Eligibility and Promotion Rate, which may be related to noncognitive as well as cognitive factors, do not appear to be dependent on the soldier's location (Post), MOS, or demographic group (i.e., these measures seem to be fairly even-handedly administered Army-wide).

Finally, there are distinct MOS and post differences in average scores for most of the measures. For example, Administrative Specialists (71L) received more Letters/Certificates and Infantrymen (11B) more Awards than soldiers in other MOS. Soldiers at one of the five posts visited received more letters, certificates, and awards, and more extra training than soldiers at the other posts. Care will have to be exercised in pooling performance measurement data across MOS and posts to try to weed out sources of criterion contamination (e.g., differences in local filing practices) while maintaining valid distinctions.

Table III.36

Summary of Univariate[a] and Multivariate[b] Analyses of Administrative Variables

Administrative Measure	Post	MOS	AFQT	Moral Waiver	Sex	Race
Reenlistment Eligibility						
Letter/Certificate	*	*			*	*
Awards	*	*	*		*	*
Letter/Certificate/Award	*	*	*			*
Military Training Courses	*	*	*			*
Article 15/FLAG Action		*				*
Promotion Rate						

* $p \leq .05$, in both univariate and multivariate analyses. In the multivariate analysis the significance level refers to the F value obtained when the variable was first entered into the prediction equation (see footnote [b] for order of entry).

[a]Univariate analyses consisted of chi-square tests and, where appropriate, analyses of variance.

[b]Multivariate analyses consisted of stepwise multiple regressions. Control variables, consisting of four dichotomous Post variables and four dichotomous MOS variables, were entered first, followed by AFQT, Moral Waiver, Sex, and Race, in turn.

Criterion Field Test: Self-Reports of Administrative Actions

While the use of administrative measures is consonant with the Project A multimethod approach to performance measurement, and while these indexes hold promise as criteria of first-tour soldier performance and in-service predictors of second-tour performance, it must be asked whether the effort and expense of collecting these indexes from the 201 Files are justified by the outcome. Also, while there was a high degree of correspondence between information on the EMF computerized file and information collected from the individual 201 Files, a number of the most promising variables are not available from the EMF.

Accordingly, a self-report instrument, the Personnel File Information Form, was developed and administered during the Batch A field testing. The self-report information could then be compared to the information in actual 201 Files, obtained by the project team during the field test period. Information on the field test results and subsequent modifications of the administrative measures is contained in Section 15.

Section 8

CRITERION FIELD TESTS: SAMPLE AND PROCEDURE[1]

The initial development of the Project A criterion measures has been described in Sections 2-7. These measures were revised on the basis of experience from the criterion field tests. This section describes the sample and procedures that were used in the field tests. Results from the field tests of specific measures are reported in the sections that follow.

The objectives of the criterion field tests were to:

- Provide item/scale analyses for the subsequent revision of the criterion measures to be used in the major validation samples.
- Provide data on the reliabilities and factor structures of the performance ratings, job sample measures, and job knowledge tests.
- Provide data to estimate the interrelationships among the major kinds of criterion measures.
- Evaluate the data collection procedures for use in the subsequent large-scale Concurrent Validation.

The Sample

The sample for the field tests was drawn from nine different jobs, or Military Occupational Specialties (MOS), and from six different locations. The nine jobs and their MOS designation--the now familiar Batch A and Batch B--were as follows:

11B Infantryman
13B Cannon Crewman
19E Armor Crewman
31C Radio Teletype Operator
63B Light Wheel Vehicle Mechanic
64C Motor Transport Operator
71L Administrative Specialist
91A Medical Specialist
95B Military Police

Tables III.37 and III.38 provide a breakdown of the criterion field test sample sizes by MOS and location, and by race and sex, respectively. USAREUR refers to the data collection site just outside Frankfurt, Germany.

---

[1]This section is based primarily on a paper, Criterion Reduction and Combination via a Participative Decision-Making Panel, by John P. Campbell and James H. Harris, in an ARI Research Note (in preparation) which supplements this Annual Report.

Table III.37

Field Test Sample Soldiers by MOS and Location

Location	MOS									Total
	11B	13B	19E	31C	63B	64C	71L	91A	95B	
Fort Hood	--	--	--	--	--	--	48	--	42	90
Fort Lewis	29	--	30	16	13	--	--	24	--	112
Fort Polk	30	--	31	26	26	--	60	30	42	245
Fort Riley	30	--	24	26	29	--	21	34	30	194
Fort Stewart	31	--	30	23	27	--	--	21	--	132
USAREUR	58	150	57	57	51	155	--	58	--	596
Total	178	150	172	148	156	155	129	167	114	1,369

Table III.38

Field Test Sample Soldiers by Sex and Race

Race	Sex		Total
	Male	Female	
Black	330	58	388
Hispanic	37	3	40
White	789	104	893
Other	43	5	48
Total	1,199	170	1,369

The Criterion Measures

As described in the earlier sections, the general procedure for criterion development in Project A was to follow a basic cycle of a comprehensive literature review, conceptual development, test and scale construction, pilot testing, test and scale revision, field testing, and Proponent (management) review.

Criterion Measurement Goals

The primary goals of criterion measurement in Project A were to (a) make a state-of-the-art attempt to develop job sample or hands-on measures of job task proficiency, (b) compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multitrait, multimethod approach), (c) develop rating scale measures of performance factors that are common to all first-tour enlisted MOS (Army-wide measures), (d) develop standardized measures of training achievement to determine the relationship between training performance and job performance, and (e) evaluate existing archival and administrative records as possible indicators of job performance.

The overall criterion development effort focused on three major methods: hands-on samples, multiple-choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in developing the rating methods.

Field Test Criterion Battery

The complete array of specific criterion measures used at the field test sites is given below. Again, the distinction between MOS-specific and Army-wide is that the Army-wide measures are the same across all MOS; that is, the same questionnaire or the same rating scale is used for all examinees. The content of the MOS-specific measures, regardless of whether they are job samples, knowledge tests, or ratings, is specific to a particular job and is based on the task content of that job. Also, the judgment (i.e., rating) of "NCO potential" refers to a first-tour enlisted soldier's potential, assuming the individual would reenlist, for being an effective noncommissioned officer, with supervisory responsibilities, during the second tour of duty.

A. MOS-Specific Performance Measures

1) Paper-and-pencil tests of achievement during training, consisting of job-relevant knowledge tests of 100 to 200 items per MOS.

   - Individual item scores
   - Mean test scores

2) Paper-and-pencil tests of knowledge of task procedures consisting of an average of about nine items for each of 30 major tasks for each MOS. Item scores can be aggregated in at least four ways.

- Sum of item scores for each of the 30 tasks.
- Total score for 15 tasks also measured hands-on.
- Total score for 15 tasks not measured hands-on.
- Total score on all 30 tasks.

3) Hands-on measures of proficiency on tasks for each MOS, measured on 15 tasks selected from the 30 tasks measured with the paper-and-pencil test.

- Individual task scores.
- Total score for all 15 tasks.

4) Ratings of performance, using a 7-point scale, on each of the 15 tasks measured via hands-on methods by:

- Supervisors
- Peers
- Self

5) Behaviorally anchored rating scales of 6-12 performance dimensions for each MOS by:

- Supervisors
- Peers
- Self

6) A general rating of overall MOS task performance by:

- Supervisors
- Peers
- Self

7) A job history questionnaire administered to incumbents to determine the frequency and recency of task performance on the 30 tasks being measured.

B. Army-Wide Measures

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

- Technical Knowledge/Skill
- Initiative/Effort
- Following Regulations/Orders
- Integrity
- Leading and Supporting
- Maintaining Assigned Equipment
- Maintaining Living/Work Areas
- Military Appearance
- Physical Fitness
- Self-Development
- Self-Control

2) A rating of general overall effectiveness as a soldier by:

- Supervisors
- Peers
- Self

3) A rating of noncommissioned officer (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 77-item summated rating scale measure of expected combat effectiveness.[2]

6) A 14-item self-report measure (the Personnel File Information Form) of certain administrative indexes such as awards, letters of commendation, and reenlistment eligibility.

7) The same administrative indexes taken from 201 Files (by project staff).

8) The Environmental Questionnaire, a 44-item descriptive questionnaire completed by both incumbents and supervisors for the purpose of describing 14 factors pertaining to organizational climate, structure, and practices.[3]

9) A Leader Behavior Questionnaire designed to permit incumbents and supervisors to describe leadership policies and practices in the unit.[4]

10) A Measurement Method Questionnaire administered at the end of the testing sessions to obtain soldiers' reactions to the various types of testing.

---

[2]Administered only to MOS in Batch B at Fort Riley.

[3]See Olson, Borman, Robertson, and Rose (1984).

[4]See White, Gast, Sperling, and Rumsey (1984).

Procedure

For the purpose of data collection in the field tests, the criterion measures were divided into four major blocks corresponding to:

1. Hands-on (job sample) measures (HO).

2. Rating measures (R) - both Army-wide and MOS-specific.

3. Paper-and-pencil measures of job knowledge (K_5).

4. Paper-and-pencil measures of training achievement (K_3).

Each block comprised one-half day of participant time and each participant was tested for a 2-day period.

During the week preceding data collection at each research site, the scorers for the hands-on (job sample) measure were given 2 days of training on scoring procedures, test standardization, and the overall design and objectives of Project A.

Advance Preparation on Site

This activity required approximately 3 days per test site for:

- Briefings to Commanders of the units supplying the troops to clarify the test objectives, activities, and requirements.

- Examination of the test site, equipment, supplies, and special requirements for the data collection and set-up of the hands-on test stations.

- Training of the test administrators and scorers.

- A "dry run" of the test procedures.

An officer and two NCOs from one of the supporting units were assigned to support the field test. The officer provided liaison between the data collection team and the tested units; the NCOs coordinated the flow of equipment and personnel through the data collection procedures. Each test site had a test manager (TSM) who supervised all of the research activity and maintained the orderly flow of personnel through the data collection points.

The logistics plan and test schedule were reviewed with the unit's administrative staff, and civilian and military scorers and other data personnel were trained. In the training phase, a dry run of the procedures followed the data collection schedule and used the personnel and locations designated for the test. The training focused on the handling of problem situations, particularly those requiring remediation by the scientific staff.

Training for scorers for the hands-on measures for each MOS was conducted by two project staff members. After an orientation session, staff members reviewed five HO tasks with the scorers by describing the equipment/material requirements, the procedures for setting up testing stations, and

the specific instructions for administering and scoring each HO test. The scorers then alternated evaluating each other performing the tasks; this provided experience both in administering the HO tests and scoring the performance measures of each. Project staff coached the "performers" to make unusual, as well as common, incorrect actions in order to give scorers practice in detecting and recording errors. The above procedure also identified any steps where local standard operating procedures (SOP) differed from the test; allowances for such differences were made in the test instructions. The second day of training was devoted to a dry run of the test procedures, with all scorers simultaneously evaluating a staff member performing a task. Problems arising from the instructions, test procedures, or task steps were identified and corrected.

Administration of the Measures

The administration schedule for a typical site (Fort Stewart, Georgia) is shown in Figure III.11. The field test proceeded as follows: Thirty MOS 31C and 30 MOS 19E soldiers arrived at the test site Thursday, 21 February 1985 at 0745. Each MOS was divided randomly into two groups of 15 soldiers each, identified as Groups A, B, C, or D. Each group was directed to the appropriate area to begin the administration appropriate for that group. They rotated under the direction of the test site manager through the scheduled areas according to the schedule shown in Figure III.11. The sequence was repeated for 30 MOS 91A and 30 MOS 63B soldiers beginning Monday (25 Feb 85) and for 30 MOS 11B soldiers on Wednesday (27 Feb 85). The order of administration of the measures was counterbalanced among the groups.

Before any instruments were administered to any soldier, each was asked to read a Privacy Act Statement, DA Form 4368-R. The Background Information and Job History forms were then administered, with 30 minutes allowed for completion.

Administration of Job Samples (15 tasks measured hands-on). Depending on the task being measured, the location was outside (e.g., vehicle maintenance, weapons cleaning) or inside (e.g., measure distance on a map). Scorers assigned to each test station ensured that the required equipment was on hand and the station was set up correctly, and followed the procedures for administering and scoring the tests. As each soldier entered the test station, the scorer read the instructions aloud and began the measure. The length of time a soldier was at the test station depended on both the individual's speed of performance and the complexity of the task.

MOS-Specific Job Knowledge Tests (30 tasks, half of them also in HO testing). The MOS-specific knowledge tests are grouped into four booklets of about seven or eight tasks per booklet. Each booklet took about 45 minutes to complete. The order of the booklets and the order of the tasks in each booklet were rotated. There was a 10-15 minute "smoke and stretch" break between booklets. The purpose of the grouping into booklets was to try to control the effects of fatigue and waning interest.

Training Achievement Tests. The training knowledge test for each MOS was in three booklets. The sequence of the booklets was alternated so that soldiers sitting next to each other had different booklets. Again, the

Group*	31C A	31C B	19E C	19E D	91A E	91A F	63B G	63B H	11B I	11B J
Tuesday 19 Feb 85	----------Scorer Training (All Scorers)----------									
Wednesday 20 Feb 85	----------Scorer Training (All Scorers)----------									
Thursday AM	PH	PK_5	PK_5	PK_3						
21 Feb 85 PM	K_3	H	R	R						
Friday AM	RS	K_3S	H	K_5						
22 Feb 85 PM	MK_5	MR	MK_3S	MHS						
Monday AM					PH	PK_5	PK_5	PK_3		
25 Feb 85 PM					K_3	H	R	R		
Tuesday AM					RS	K_3S	H	K_5		
26 Feb 85 PM					MK_5	MR	MK_3S	MHS		
Wednesday AM									PH	PK_5
27 Feb 85 PM									K_5	H
Thursday AM									K_3S	R
28 Feb 85 PM									MR	MK_3S

* Each group equals 15 soldiers in the same MOS.

Code: P = Personal and Job History forms
K_3;K_5 = Task 3 or Task 5 Knowledge Measures
H = Hands-on Measures
R = Peer Ratings
S = Supervisor (rater and endorser) Ratings
M = Measurement Method Questionnaire

Figure III.11. Typical field test administration schedule.

purpose of using booklets was to try to control the effects of fatigue and waning interest. Soldiers had 45 minutes for each booklet and a 10-15 minute "smoke and stretch" break between booklets.

Rating Scales. The supervisory, peer, and self ratings are designed around "rating units." Each rating unit consists of the individual soldier to be evaluated, four identifiable peers, and two identifiable supervisors. A peer is defined as an individual soldier who has been in the unit for at least 2 months and has observed the ratee's job performance on several occasions. A supervisor is defined as the individual's first or second line supervisor (normally his rater and endorser).

The procedure for assigning ratees to raters (both peers and supervisors) consists of two major steps: (a) a screening step that determines which raters could potentially rate which ratees; and (b) a computerized random assignment procedure that assigns raters to ratees within the constraints that 1) the rater indicated he/she could rate the ratee; 2) ratees with few potential raters are given priority in the randomized assignment process; 3) the number of ratees assigned is equalized as much as possible across raters; and 4) the number of ratees any given rater is asked to rate does not exceed a preset maximum.

The potential raters were given an alphabetized list of the ratees. They were told the purpose of the ratings within the context of the research, and the criteria (e.g., minimum length of period of working together) they should use in deciding who they could rate. They were told the maximum number of people they would be asked to rate and that assignments of ratees to raters would be made randomly. The importance of their careful and comprehensive examination of the list of ratees was emphasized.

The rating scale administrator, using the training guide, then discussed the content of each effectiveness category, and urged raters to avoid common rating errors. A major thrust of this training was an attempt to standardize the rating task for the raters. With the lack of control to be expected, an important concern was that all raters face the same (or a very similar) rating task. A serious potential confounding involves rating unit and administrator; lower average ratings for some rating units might be a result of different sets (i.e., "rate more severely") provided by administrators handling those rating units rather than true performance deficiencies. Standardization of the administration helps reduce this potential problem. A second major thrust of the rater training was to minimize the amount of reading the raters had to do. This was, as much as possible, an oral administration; the rating program was not dependent on raters' reading large amounts of material.

Planned Analysis

The general analytic steps were straightforward and consisted of the following:

1. Item analysis for each job knowledge test for each MOS.

2. Item analysis for the training achievement tests for each MOS. An analysis of item responses was done for a sample of 50 trainees as well as for the incumbent samples in the field tests.

3. An item analysis summary table for each knowledge test for each MOS. The table for each MOS summarized item discrimination indexes, item difficulties, and the frequency of items that were flagged for various kinds of potential keying errors (e.g., negative correlation with total score, high frequency of response for incorrect answer).

4. An item (where task = item) analysis for each hands-on (job sample) test.

5. Frequency distribution and scale statistics for each rating scale for each MOS.

6. Interrater reliabilities for the individual rating scales.

7. Split-half correlations (Spearman-Brown estimates) for the knowledge tests and hands-on measures, test-retest coefficients for the hands-on measures, and internal consistency indexes where applicable.

8. A complete intercorrelation matrix of all the criterion variables for each MOS down to the scale score and task score level (i.e., the matrix included all the variables listed in the previous sections).

9. A set of reduced intercorrelations matrixes that included subsets of the total array of variables.

10. Factor analyses for selected matrixes, primarily those having to do with the rating scale measures.

11. For a selected number of variable pairs, correction of the intercorrelation for attenuation in an attempt to estimate the correlation between the true scores.

Interpretation and Use of the Field Test Results

The results of the above analyses were prepared in a master data book for each MOS. Each data book contained item and scale analyses, intercorrelations down to the scale and subscale level, and factor analyses of selected data sets.

These data were then carefully scrutinized by a designated criterion analysis group. The group included the principal investigator for each of the criterion measures; consequently, for each variable there was at least one committee member with a strong vested interest. The other members of the committee consisted of the principal scientist for the project, the Army

Research Institute's chief scientist for the project, and one hapless individual (the assistant project director) who had to serve as chair--10 people in all.

The objectives of the group were to review the results of the field tests and to agree upon the specific revisions that were to be made in each criterion measure before the criterion array was declared the set of criterion measures that would be used for the Concurrent Validation. The mode of operation was for the principal investigator responsible for each criterion to review carefully the relevant field test data and propose specific revisions, additions, or deletions aimed at maximizing the reliability, acceptability, and construct validity of the measure. A general discussion then followed, continuing until the investigator's proposal was accepted or a consensus was reached on what specific changes should be made.

The obvious disadvantage of the committee approach to data interpretation is the time involved. More than once the membership wished for a good dose of totalitarian power. On the positive side, all the major benefits of participative decision making seemed to manifest themselves. Everyone concerned always knew what was being done, crucial issues tended not to get lost, investigators could exercise veto power if the integrity of their product was being threatened, and considerable commitment seemed to have been generated. On balance, the time investment seemed worth it. In truth, on such a large, multifaceted project it probably is not possible for one "expert" to make these decisions unilaterally. If the Project A model is used in the future with any frequency, applied psychologists must learn how to "manage" data interpretation as well as data collection.

The following sections summarize the major findings generated by the above analyses and outline the revisions made in the performance measures as a result. Results pertaining to the self ratings are not included in these summaries; initial analyses indicated that the self ratings suffered from relatively more halo, central tendency, and leniency error than did supervisor and peer ratings, and self ratings were not considered further.

Section 9

FIELD TEST RESULTS: TRAINING ACHIEVEMENT TESTS[1]

Descriptive Statistics for Field Tests

The descriptive statistics for the training achievement tests administered in the field tests to job incumbents are shown in Table III.39. Test scores are based on the items judged relevant for the job or for the job and for training content. Those few items judged relevant only for training content (see Table III.10) are not included because the respondents being tested were job incumbents rather than trainees.

These data are for Batches A and B only (nine MOS) since Batch Z MOS were not field tested. Mean values for the previous trainee figures based on 19 MOS (Section 2) have been recomputed including only the nine MOS that participated in the field tests; trainee and field test job incumbent results match closely. Mean trainee alpha for the nine MOS was .882, and mean incumbent alpha is .877. Mean correct for trainees was 53.9%, compared to 54.5% for job incumbents.

Revisions to Training Achievement Tests

Reduction in Number of Items for Concurrent Validation

Because of time constraints, the length for the Concurrent Validation versions of the training tests would be limited to approximately 150 items. To reduce the size of the item pool, any item that had been rated not relevant to the job and also not relevant to training was dropped first. To reduce test length further, items were dropped that had been rated lowest in importance and/or highest in difficulty. Because the training performance domain was assumed to be multidimensional, items were not usually eliminated solely because of a negative biserial correlation with the total test score. However, some items were dropped that exhibited the three characteristics of (a) low pass rate, (b) negative biserial, and (c) a distractor or distractors with a high positive biserial. During the revision of the item pools, the relative frequency of items in each job task duty area was maintained as it had been previously.

Tables III.40, III.41, and III.42 report the number of items remaining on each test after the revisions had been made. The versions to be used for the Concurrent Validation contained the number of items shown in the columns on the far right. The tables for Batches A and B differ slightly from the table for Batch Z because many of the Batch A and B item reductions were made using field test data, which are not available for Batch Z.

---

[1]Development of the training achievement tests was described in Section 2, Part III. Section 9 is based primarily on ARI Technical Report 757, Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS, by Robert M. Davis, Gregory A. Davis, John M. Joyner, and Maria Veronica de Vera, and the supplementary ARI Research Note in preparation, which contains the report appendixes.

Table III.39

Results From Training Achievement Field Tests Administered to Incumbents

MOS	Number of Subjects	Number of Items	Mean Number Correct	SD	Range	Alpha	Mean Percent Correct
Batch A							
13B	149	133	49.2	16.5	74	.90	44.5
64C	155	137	70.3	17.2	75	.91	51.3
71L	129	97	50.5	9.9	51	.83	52.1
95B	112	131	77.3	10.2	51	.76	59.0
Batch B							
11B	166	162	86.4	20.0	98	.93	53.3
19E	169	193	112.9	21.0	142	.93	58.5
31C	143	176	99.6	20.1	120	.92	55.6
63B	155	205	106.9	19.4	107	.90	52.1
91A	155	115	72.9	10.3	76	.82	63.4

Review by TRADOC Proponent Agencies

Before being administered to job incumbents as part of the Concurrent Validation, each item pool was submitted to the appropriate TRADOC Proponent for review. The number of items sent out for review and the number of items eliminated, added, or modified as a result of this review are also summarized in Tables III.40, III.41, and III.42.

Comparison of Initial Item Pool and Concurrent Validation Version

When initial item pool and Concurrent Validation versions are compared, there is a small increase in the percentage of items rated very important and a small decrease in the proportion of items rated of little importance on both the combat scenario (Very Important, 33.1 to 34.0%; Of Little Importance, 22.8 to 20.6%) and the garrison scenario (Very Important, 43.1 to 46.5%; Of Little Importance, 11.2 to 8.3%). These changes are all in the expected direction, given the procedures that were used to revise the initial item pools.

Mean importance ratings across MOS for item pool and Concurrent Validation versions of the tests for each scenario were also compared. All changes are in the expected direction (i.e., higher importance on the Concurrent Validation test version than on the item pool), and two are significant when compared using the Wilcoxon Matched Pairs test: combat scenario (Initial Item Pool vs. Concurrent Validation version), $Z = 1.73$, $p = .08$; combat readiness scenario (Initial Item Pool vs. Concurrent Validation version), $Z = 2.01$, $p = .04$; garrison scenario, $Z = 2.86$, $p = .004$.

Table III.40

Number of Items in Training Achievement Tests at Each Stage of Development: Batch A

MOS	Initial Item Pool No.[a]	No. of Cuts by Category		No. of Items Sent to Proponent Review[b]	Proponent Review		No. of Items Remaining[c]
		Not Relevant	Low Importance or Poor Item Characteristics		Cut/Added	Modified	
13B (SP)	163/68[d]	18	75	138	2/0	0	136
13B (T)	163/67[d]	14	55	161	5/0	0	156
64C	228	2	86	140	12/0	70	128
71L	130	1	28	101	6/10	12	105
95B	223	11	72	140	6/5	9	139

[a] Items field tested.

[b] Reflects one or more Proponent reviews.

[c] No budget balancing was needed for Batch A tests.

[d] There were 163 items common between the SP & T versions; 68 items were unique to SP and 67 unique to T.

Table III.41

Number of Items in Training Achievement Tests at Each Stage of Development: Batch B

MOS	Item Pool No.		No. of Cuts by Category		No. of Items Sent to Proponent Review	Proponent Review		Items Added to Rebalance Budget	No. of Items Remaining
	School	Batch B Field Test	Not Relevant	Low Importance or Poor Item Characteristics		Cut/Added	Modified		
11B	200	199	5	13	181	35/0	13	4	150
19E	214	202	2	21	179	17/0	8	0	162
31C	192	199	5	15	179	4/0	7	0	175
63B	238	217	2	47	168	24/23a	81	5	172
91A	299	260	5	46	209	34/0b	17	0	175

a Reflects two Proponent reviews.

b Reflects an additional cut made in conjunction with Proponent review to bring the test down to 175 items.

Table III.42

Number of Items in Training Achievement Tests at Each Stage of Development: Batch Z

MOS	No. of Items Sent to Proponent Review	Proponent Review		Additional Cuts Based on Low Importance or Poor Item Characteristics	Items Added to Rebalance Budget	No. of Items Remaining
		Cut/Added	Modified			
12B	211	0/0	5	35	0	176
16S	202	49/0	35	5	1	149
27E	205	2/0	9	28	0	175
51B	202	8/0	5	31	0	163[a]
54E	207	63/0	22	5	6	145
55B	212	0/0	5	31	0	181
67N	207	6/0	0	26	0	175
76W	195	1/0	0	19	0	175
76Y	188	2/0	11	11	0	175
94B	187	3/0	4	8	0	175

[a] Reduced to this number to eliminate time-consuming math items.

For the version of the tests administered as part of the Concurrent Validation, the distribution across relevance categories is nearly the same as for the original item pool.

Some Lessons Learned

Since this was such a large-scale test development effort conducted over a relatively short period of time, a number of things were learned in addition to the psychometric properties of the scales. We summarize a few of these below.

Item Tracking

Developing more than 200 test items for each of the 20 different MOS required keeping track of data on more than 4,000 test items, through several revisions. To do this, each draft item was assigned a master number, and a large table was constructed for each MOS showing, for each version of the item pool (version shown to incumbents, version used in the field test, etc.), the test booklet number of that item and the item in its revised form. Into these same tables were entered the AOSP (job task) duty area for the item, whether the item was judged relevant to training and to the job, whether the item was modified or dropped from the pool, and so forth.

Tracking items is further complicated by the fact that as items are reviewed, many are changed significantly. Judgments regarding relevance and importance refer, of course, to a particular item at a given point in time. After each item change, a judgment must be made as to whether or not the item is still the "same" item. If not, then the original item must be recorded as dropped, and a new item with a new master number (linked to the old) must be entered into the item pool. An automated database program running on a small (or large) computer appears necessary in an effort of this magnitude.

Evolution of Item Budgets

Item budgets were originally developed to help assure that the content domain for each test would be clear, representative, and relevant. Such budgets also serve the important functions of guiding and providing discipline to item writers who often do not understand the psychometric issues involved in test construction.

Since the original pool of items was larger than needed for the tests, it was possible to keep reworking the budgets, to ensure that the content domain was appropriately sampled. The important point to note here is that the original budgets were a starting point in the test development process. The SME and Proponent reviews provided important insights to ensure that the training achievement tests were indeed content valid.

A reasonable way to track budgets is to set up spread sheets that forecast the number of items needed in specific content areas as the item pools evolve into actual tests.

Summary and Discussion

The major objective for this part of Project A's criterion development was to create content-valid and reliable training achievement tests for measuring the cognitive component of training success. How successful were these efforts?

The tests in Batches A and B may be evaluated from three perspectives. First, since content validity is so crucial, one can examine the process by which the tests were developed and use some of the standards identified by Guion (1977) and others as criteria for evaluating that process. Second, one can consider the development process up to the point of the final Proponent review, which indeed was an added step in the process, and compare the tests before and after Proponent review. The assumption here is that if the tests undergo relatively little change (particularly fundamental change such as cutting items and/or adding new items) as a result of the final Proponent review, the development process as originally conceived was valid. Finally, one can look at descriptive psychometric indexes such as reliability and item distributional characteristics.

The development process did conform to the three criteria of domain definition, content representativeness, and content relevance. First, the domain was operationally identified and items were drawn from that content. The developmental model prescribed that the initial items would be drawn from published Army literature. Since the published literature inevitably lags behind practice (i.e., doctrine and equipment), some change was inevitable as subject matter experts examined items. Nevertheless, the changes were, in most cases, not dramatic and many concerned terminology or phrasing rather than content.

With respect to content representativeness, the proportions of items assigned to different duty areas on different versions of the test are similar. Inevitably, there were some modifications in the percentage of items in a given duty area, but radical changes in the distribution of items across duty areas were not necessary.

Elaborate procedures were used to determine content relevance. Items judged by experts as being not relevant to training and/or the job were eliminated. Moreover, relevance was judged in terms of importance; only those items judged to be very important on one or more of the three scenarios were retained in the item pool.

With respect to fairness, as procedures were being developed for review of items by subject matter experts, guidelines were developed and implemented to ensure that the groups reviewing the items were balanced for race and gender.

Next is the question of whether the Proponent review altered or changed the tests. The short answer is: With one or two exceptions, not very much. Proponents requested three types of changes. The mean percentages of these changes across all 19 MOS were as follows: cuts, 7.5%; additions, 1.4%; and modifications, 9.4%. When one considers the lengths of the tests, these percentages are not very great. Furthermore, modifications were in many cases relatively trivial and did not concern content so much as format or

phrasing. The distributions of these changes were, however, quite skewed. By consulting Tables III.40, III.41, and III.42, one can note that the most significant disagreements occurred for MOS 16B (cuts), 54E (cuts), 11D (cuts), and 63B (modifications). Items were added to these tests to rebalance their respective item budgets.

Finally, the tests can be evaluated in terms of more traditional psychometric measurements, particularly reliability. All of the tests have high reliability coefficients and reasonable distributal properties. In total they appear to possess considerable content validity for their intended purpose.

Section 10

FIELD TEST RESULTS: TASK-BASED MOS-SPECIFIC CRITERION MEASURES[1]

The analyses of the field test data for the task-specific performance measures used three general kinds of information. First, extensive item/scale analyses, including the calculation of reliability indexes, were carried out on each measure. Second, the intercorrelations among the different measures were examined. Finally, consideration was given to SME and staff judgments on the relative suitability of job sample vs. paper-and-pencil for assessing specific task performance.

Item/Scale Analyses

The basic psychometric properties of each measure are described in turn for the job knowledge tests, the job sample or hands-on tests, the task performance rating scales, and the Job History Questionnaire.

Job Knowledge Tests

The output generated by the item analysis procedure for the knowledge tests included, for each item, the number and percentage who selected each alternative, and for each item alternative, the Brogden-Clemans item-total correlation (in which the total score represents the sum of all the items used to assess knowledge of a specific task, excluding the item being correlated with the total). Recall that there were about 30 task total scores in each job knowledge test.

Although items with extremely high or low difficulties provide relatively few discriminations, some such items might still be retained to enhance test acceptability (e.g., because of the importance of the content) or to preserve a measure of a task to be tested across several MOS. Also, items with low item-total correlations might be deficient in some respect or might simply be increasing test content heterogeneity. Since neither type of information provided conclusive evidence regarding an item's utility, both were applied in a judicious and cautious manner.

Those items that were particularly easy (more than 95% pass) or particularly difficult (fewer than 10% pass), or that had low or negative item-total correlations were examined first for keying errors or obvious sources of cueing. Deficient items that could not be corrected were then

---

[1]Development of these measures was described in Section 3, Part III. Section 10 is primarily based on ARI Technical Report 717, Development and Field Test of Task-Based MOS-Specific Criterion Measures, by Charlotte H. Campbell, Roy C. Campbell, Michael G. Rumsey, and Dorothy C. Edwards, and the supplementary ARI Research Note in preparation, which contains the report appendixes.

deleted, and the item analysis was produced again. The process was iterative; various sets of items were included in the analysis, and the set that produced the highest coefficient alpha for the entire knowledge task test with an acceptable pass rate (between 15% and 90%) was retained.

Revisions were made on between 14 and 18% of the items in each MOS set, and between 17 and 24% of the items were dropped. Following item deletions, the distributions of items with regard to difficulty and item-total correlations for each of the nine MOS were as summarized in Table III.43. The median difficulty levels were 55 to 58% for five of the MOS, with the MOS 63B, 91A, 19E, and 95B tests having medians of 65 to 74%. Although some skew in item difficulties was observed, it was not extreme.

Table III.43

Summary of Item Difficulties (Percent Passing) and Item-Total Correlations for Knowledge Test Components in Nine MOS

MOS		Number of Items		Mean	Median	Min	Max
13B	Cannon Crewman	236	Difficulty(%)	59.2	55.5	13.4	97.2
			Item-Total(r)	.38	.38	-.06	.88
64C	Motor Transport Operator	166	Difficulty(%)	60.7	58.0	03.6	94.3
			Item-Total(r)	.31	.32	-.00	.91
71L	Administrative Specialist	170	Difficulty(%)	57.4	56.5	04.7	96.1
			Item-Total(r)	.30	.31	-.19	.84
95B	Military Police	177	Difficulty(%)	66.4	74.0	00.0	100.0
			Item-Total(r)	.33	.32	.00	.82
11B	Infantryman	228	Difficulty(%)	57.3	55.4	05.3	97.1
			Item-Total(r)	.30	.31	-.39	.88
19E	Armor Crewman	205	Difficulty(%)	64.6	66.8	13.4	96.9
			Item-Total(r)	.32	.31	-.26	.95
31C	Single Channel Radio Operator	211	Difficulty(%)	58.0	57.1	11.3	95.4
			Item-Total(r)	.31	.31	-.09	.84
63B	Light Wheel Vehicle Mechanic	197	Difficulty(%)	65.1	64.5	07.8	97.4
			Item-Total(r)	.30	.30	-.13	.92
91A	Medical Specialist	236	Difficulty(%)	66.9	69.0	08.6	98.7
			Item-Total(r)	.32	.32	-.25	.78

The item-total correlation distributions were also highly similar across the nine MOS, with most items exhibiting correlations of .21 to .40 in each MOS. Pruning items on the basis of low correlations was done very conservatively. As a result, there remained in each knowledge test items with low or negative correlations with the task total score. These ranged from 9% of the items in the MOS 13B (Cannon Crewman) tests to 29% of the items in the MOS 19E (Armor Crewman) tests that had correlations below .20. Negative correlations were found in no more than 8.8% of the items in any of the nine MOS. The average of the item-total correlations in the various knowledge components ranged from .30 to .38.

The means, standard deviations, and reliabilities for the total test score in each MOS are shown in Table III.44. The reliabilities are split-half coefficients, using 15 task tests in each half, corrected to a total length of 30 task tests.

Table III.44

Means, Standard Deviations, and Split-Half Reliabilities for Knowledge Test Components for Nine MOS

MOS	Mean (%)	Standard Deviation	Split-Half Reliability[a]
13B - Cannon Crewman	58.9	12.6	.86
64C - Motor Transport Operator	60.3	10.1	.79
71L - Administrative Specialist	55.8	10.4	.81
95B - Military Police	66.4	9.2	.75
11B - Infantryman	56.0	10.5	.81
19E - Armor Crewman	64.0	10.1	.90
31C - Single Channel Radio Operator	57.7	9.6	.84
63B - Light Wheel Vehicle Mechanic	64.4	9.1	.86
91A - Medical Specialist	69.8	8.1	.85

[a] Fifteen task tests in each half, corrected to a total length of 30 tests.

For all MOS, the majority of individual task test means were between about 35 and 85%; total score means were from 55 to 70%. The standard deviations were also similar across the nine MOS, and although coefficient alphas were variable across tasks, split-half reliabilities were in the .70s and .90s for total job knowledge score.

The reliabilities (coefficient alpha) of task tests appearing in multiple MOS are shown in Table III.45. The magnitude of the correlations is, for some individual task tests, disappointing. However, a number of the subtests were very short, containing no more than 3-5 items.

Table III.45

Coefficient Alpha of Knowledge Tests Appearing in Multiple MOS

Test	13B	64C	71L	95B	11B	19E	31C	63B	91A
Perform CPR	.31	.34		.38	.33	.38	.41		.55
Administer nerve antidote to self	.55		.39		.36				
Prevent shock	.22					.12			.31
Put on field dressing		.34	.39	.39	.19	.15	.31	.16	.31
Administer nerve agent antidote to buddy		.58						.32	
Load, reduce stoppage, clear M16	.56	.46	.47	.52			.51	.32	.43
Perform operator maintenance on M16		.31	.38		.39		.44	.22	
Load, reduce, clear M60		.30		.40	.47				
Perform operator maintenance .45				.45		.36			
Determine azimuth using a compass			.81	.84				.74	
Determine grid coordinates		.23	.53	.57		.79	.74	.70	.74
Decontaminate skin	.71	.42		.48			.47		.47
Put on M16 mask	.50	.49	.44	.56	.49			.33	
Put on protective clothing	.56	.55	.31		.40	.31	.52	.39	.40
Maintain M17 mask			.38	.53			.28		
Challenge and Password	.46	.48						.41	
Know rights as POW			.48			.45	.44		
Noise, light, litter discipline			.38				.12		.07
Move under fire				.59	.56				
Identify armored vehicles	.62			.64	.68	.75	.57		.58
Camouflage equipment	.31	.31						.08	
Camouflage self			.06	.47	.48				
Report information - SALUTE		.76			.84	.74			
Operate vehicle in convoy		.40		.36					

Step/Scale Analyses for Hands-On Tests

For each hands-on step, the number and percentage who scored GO and NO-GO were determined. The Brogden-Clemans biserial was computed for hands-on steps just as for knowledge test items; that is, the step was correlated with the total task score minus that step. Recall that for the hands-on tests there were 15 task scores.

Steps that had low or negative correlations with the total task score were reviewed to identify situations where performance scored as NO-GO was in fact prescribed by local practices, and was as correct at that site as doctrinally prescribed procedures. Instructions to scorers and to soldiers were revised as necessary to ensure consistent scoring.

Use of step difficulty data to revise hands-on tests was limited by a number of considerations. First, a task test usually represents an integrated procedure. Each individual step must typically be performed by someone in order for the task to continue. Removal of a step from a scoresheet, regardless of its psychometric properties, might only confuse or frustrate the scorer. Second, removal of a step which the Soldier's Manual specifies as a part of the job may result in deleting a doctrinal requirement and undermining the acceptability of the hands-on measure to management.

Because of these considerations very few performance measures were dropped from scoring on hands-on tests, regardless of their difficulty level. However, under certain limited circumstances, exceptions were made. On a very few hands-on tasks, the test steps represent a sample of performance from a large domain (e.g., Identify Threat Aircraft, Recognize Armor Vehicles, Give Visual Signals). Individual steps could be deleted without damaging task coverage or test appearance.

Three types of reliability data were explored for hands-on task tests: test-retest, interscorer reliability, and internal consistency. For reasons discussed below, only the internal consistency data were systematically used in test revision.

So that test-retest reliability could be computed, all soldiers in the Batch A MOS were retested on a subset of the same tasks they had been tested on initially. Due to scheduling and resource constraints, the interval between first and second testing was only 2 to 7 days. Thus, memory of initial testing was a probable contaminant of retesting performance. Soldiers were aware that they would be retested and some were found to have trained to improve their performance between the two testing sessions. Thus, training was not consistent across soldiers, but varied partly as a function of motivation and partly as a function of the extent to which special duties restricted training opportunities. Scores improved on second testing for many soldiers. On the other hand, some soldiers resented having to repeat the test; some told the scorer that they were unfamiliar with the task, when in fact they had scored very high on initial testing. Thus, retest scores were contaminated widely and variably by motivational factors. Overall, test-retest was of limited utility and was not collected for Batch B soldiers.

The use of alternate forms of a test offers an approximation of test-retest reliability. However, development and large-scale administration of alternate forms in either hands-on or knowledge mode was beyond the resources of the project.

An attempt was made in Batch B to acquire interscorer reliability estimates by having a Project A staff member score the soldier at the same time the NCO was scoring. Two factors limited the feasibility of this approach. First, sufficient personnel were not available to monitor all eight stations within an MOS for the length of time required to accumulate sufficient data. The problem was exacerbated when, for whatever reason, it was necessary to test two MOS simultaneously. Second, for some MOS, particularly those performed in the radio-teletype rig for MOS 31C and in the tank for MOS 19E, it was difficult or even impossible to have multiple scorers without interfering

with either the examinee or the primary scorer. Because of these factors, insufficient interscorer reliability data were available to systematically affect the process of revising task measures.

By process of elimination, the reliability measure of choice for the hands-on test was an internal consistency estimate, using split-half. Table III.46 shows, for each MOS, the means, standard deviations, and split-half reliability estimates of the hands-on components across revised task tests. The mean task scores tend to fall between 40 and 80%, with a few very difficult tasks and a few very easy tasks for most soldiers in each MOS. The standard deviations for task tests are in many cases high relative to the means. This is at least in part an artifact of the sequential nature of many of the hands-on tests: If soldiers cannot perform early steps, the test stops and remaining steps are failed. For many MOS, the overall split-half, calculated using seven scores against eight scores (odd-even), is rather low, but these may be underestimates since it is difficult to conceive of parallel forms arranged from tests of such heterogeneous tasks.

Table III.46

Means, Standard Deviations, and Split-Half Reliabilities for Hands-On Test Components for Nine MOS

MOS	N	Mean (%)	Standard Deviation	Split-Half Reliability[a]
13B - Cannon Crewman	146	64.5	14.0	.82
64C - Motor Transport Operator	149	72.9	9.1	.59
71L - Administrative Specialist	126	62.1	9.9	.66
95B - Military Police	113	70.8	5.8	.30
11B - Infantryman	162	56.1	12.3	.49
19E - Armor Crewman	106	81.1	11.8	.56
31C - Single Channel Radio Operator	140	80.1	10.7	.44
63B - Light Wheel Vehicle Mechanic	126	79.8	8.7	.49
91A - Medical Specialist	159	83.4	11.4	.35

[a] Calculated as 8-test score correlated with 7-test score, corrected to 15 tests.

Task Performance Rating Scales

Inspection of the rating data revealed level differences in the mean ratings provided by two or more raters of the same soldier. Therefore, all raters' responses were adjusted to eliminate these level differences. Additionally, a small number of raters were identified as outliers, in that their ratings were severely discrepant overall from the ratings of other raters on the same soldiers; their rating data were excluded from the analyses. (The procedures for adjusting the ratings for level effects and for identifying outliers are described in more detail in the discussion of measures in Section 11.)

Means and standard deviations were computed on the adjusted ratings for each 7-point scale. Interrater reliabilities were computed as intraclass correlations, and the estimates were adjusted so as to represent the reliability of two supervisors per soldier and four (Batch A) or three (Batch B) peer raters for each soldier. The adjustment was made because numbers of raters varied for each soldier, and it seemed reasonable to expect that ratings could be obtained from two supervisors and four peers during the Concurrent Validation. However, further experience in Batch B data collection suggested that three peers per soldier was more reasonable. These issues are discussed more fully in the sections on rating scale results.

Summary statistics on the task performance rating scales across the 15 tasks in each MOS are presented in Table III.47 (more detailed results can be found in ARI Technical Report 717). The distributions for the rating scales are surprisingly free of leniency and skewness, with means between 4.3 and 4.9 on the 7-point scale and standard deviations largely between .55 and .75.

Reliabilities varied widely across the tasks. In MOS such as 71L (Administrative Specialist) and 63B (Light Wheel Vehicle Mechanic) where soldiers work in isolation from each other or with only one or two others, few peer ratings were obtained on each soldier and reliabilities are correspondingly lower. The mean number of peer ratings among 11B (Infantryman) was much higher, and many of the soldiers being tested comprised training cohorts that had been together since their earliest Army training. Some tasks that soldiers rarely perform were also characterized by lower numbers of ratings and lower reliabilities.

During the Batch A field tests, it was observed that supervisors and peers confronted with only the task title, might not have been entirely clear on the scope of tasks they were rating. Low interrater reliability supported this observation. Consequently, for the Batch B data collection for two MOS (31C and 19E), the task statements were augmented with the brief descriptions of the tasks that had been developed for the task clustering phase of development. However, this modification did not appear to affect results from these MOS and it was not given further trial.

Job History Questionnaire

Job history responses were analyzed to determine whether task experience as captured by the Job History Questionnaire is related to performance on the

Table III.47

Means, Standard Deviations, Number of Raters, and Interrater Reliabilities of Supervisor and Peer Ratings Across 15 Tasks for Nine MOS

MOS	Group	Mean Raters	Mean[a]	Standard Deviation[a]	Split-Half Reliability[b]
13B - Cannon Crewman	Sup.	1.5	4.99	.72	.67
	Peer	2.5	4.85	.60	.87
64C - Motor Transport Operator	Sup.	1.8	4.35	.64	.69
	Peer	2.0	4.26	.58	.70
71L - Administrative Specialist	Sup.	1.0	4.97	.70	.75
	Peer	1.9	4.97	.64	.60
95B - Military Police	Sup.	1.9	4.51	.49	.64
	Peer	3.4	4.53	.46	.82
11B - Infantryman	Sup.	1.8	4.45	.59	.74
	Peer	3.0	4.50	.55	.77
19E - Armor Crewman	Sup.	1.7	4.69	.62	.76
	Peer	3.0	4.71	.45	.67
31C - Single Channel Radio Operator	Sup.	1.7	4.68	.68	.81
	Peer	2.5	4.68	.58	.74
63B - Light Wheel Vehicle Mechanic	Sup.	1.8	4.72	.68	.76
	Peer	2.1	4.68	.63	.81
91A - Medical Specialist	Sup.	1.6	4.97	.75	.69
	Peer	3.1	4.95	.60	.81

[a] Computed on adjusted ratings.

[b] Computed on adjusted ratings; corrected to reliabilities for two supervisors and four (Batch A) or three (Batch B) peers.

task-specific criterion measures. The questionnaire asked soldiers to estimate the recency and frequency of performance of each task. If sufficient relationships were found, job history data would also be collected during the Concurrent Validation. Because the Job History Questionnaire data analyses were performed solely to inform the decision on whether to continue collecting job history information, attention was focused on detailed analyses of one Batch A MOS (13B) and three Batch B MOS (11B, 19E, and 63B).

Recency and frequency were summed with frequency reverse scored prior to summing, so that high scores indicate greater recency and/or frequency of task experience. For the Batch A MOS 13B (Cannon Crewman), this summated experience score was significantly related, in the positive direction, with test scores for 9 of the 15 hands-on tests, and for 9 of the 30 knowledge tests. For six tasks, experience was significantly related to performance on both knowledge and hands-on tests. These findings certainly support the continued examination of job experience effects.

For the three Batch B MOS, frequency and recency were treated separately. For MOS 11B (Infantryman), recency and frequency or both correlated significantly, and in the appropriate directions, for 7 of the 15 hands-on tests, and for 15 of the 30 knowledge tests. For six tasks, one or both experience indexes were related to both hands-on and knowledge performance.

For MOS 19E (Armor Crewman), experience indexes were related to only three hands-on tests and two knowledge tests. For one 19E task, experience was significantly related to both knowledge and hands-on scores. For MOS 63B (Light Wheel Vehicle Mechanic), experience was significantly related to only two hands-on tests and five knowledge tests, with none of the tasks having significant relationships with experience measures for both types of tests. For soldiers in these two MOS, experience differences appear to have less influence on performance.

Intercorrelations Among Task Performance Measures

For each of the nine MOS, performance on 15 tasks was assessed by four methods: hands-on tests, knowledge tests, supervisor ratings, and peer ratings. Thus, a 60 X 60 correlation matrix could be generated for each of the MOS, as in a multimethod-multitrait matrix (where traits are tasks). For purposes of simple examination each MOS matrix was collapsed, by averaging correlations across tasks, to a 4 X 4 matrix (see Figure III.12). For each pair of methods, the 15 correlations between the two methods on the same tasks (heteromethod-monotrait) were averaged and entered above the diagonals. The 210 correlations between each pair of methods on different tasks (heteromethod-heterotrait) were averaged and entered below the diagonals. Finally, the 105 correlations between pairs of tasks measured by the same method (monomethod-heterotrait) were averaged and are shown in the diagonals of each matrix.

In general, there are three considerations in examining a full multimethod-multitrait matrix: (a) The heteromethod-monotrait correlations (above the diagonals) are indications of convergent validity among the methods, the extent to which the different methods measure the same trait

13B - Cannon Crewman

	HO	K	R-S	R-P
HO	82			
K	41	79		
R-S	34	24	67	
R-P	47	18	46	87

11B - Infantryman

	HO	K	R-S	R-P
HO	49			
K	55	91		
R-S	46	39	74	
R-P	36	30	61	77

64C - Motor Transport Operator

	HO	K	R-S	R-P
HO	59			
K	59	68		
R-S	32	23	69	
R-P	22	10	70	70

19E - Armor Crewman

	HO	K	R-S	R-P
HO	56			
K	39	90		
R-S	09	19	76	
R-P	10	16	50	67

71L - Administrative Specialist

	HO	K	R-S	R-P
HO	66			
K	52	67		
R-S	23	12	75	
R-P	16	-02	77	60

31C - Single Channel Radio Operator

	HO	K	R-S	R-P
HO	44			
K	37	84		
R-S	17	21	81	
R-P	18	20	71	74

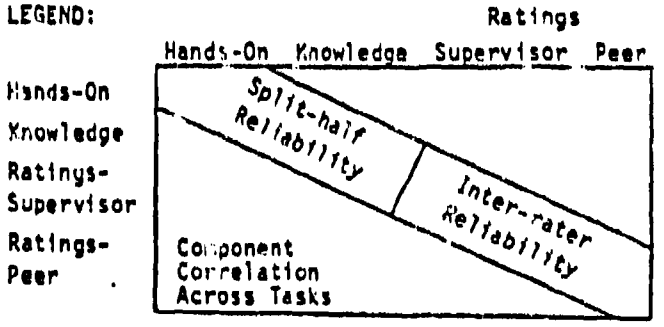
95B - Military Police

	HO	K	R-S	R-P
HO	30			
K	11	63		
R-S	27	17	64	
R-P	31	14	65	82

63B - Light Wheel Vehicle Mechanic

	HO	K	R-S	R-P
HO	49			
K	31	86		
R-S	18	08	76	
R-P	12	23	59	81

LEGEND:



91A - Medical Specialist

	HO	K	R-S	R-P
HO	35			
K	21	85		
R-S	16	00	69	
R-P	19	-03	61	81

Figure III.12. Average correlations between task measurement methods on same tasks and different tasks for nine MOS.

(here, the traits are proficiency on tasks). (b) These same validity coefficients (above the diagonals) should be greater than the corresponding values in the heteromethod-heterotrait triangle (below the diagonals), as an indication that the method-trait relationships are not all a reflection of some other unspecified factor. (c) The monomethod-heterotrait correlations (in the diagonals) should be lower than the coefficients above the diagonals, as evidence of discriminate validity--that is, the methods of measuring tasks are not overshadowing differences among tasks.

Without exception, the average correlations are highest both between and within peer and supervisor ratings, with method variance (different tasks) in general higher than variance accounted for by tasks. For hands-on and knowledge tests, the average of same-task correlations between the two methods (above the diagonal) is higher than either of the single-method different-task average correlations (in the diagonal), which in turn are usually higher than the average correlation between the two methods on different tasks (below the diagonal). The lower correlations between the task tests and task ratings, even on the same tasks (above the diagonal), further evidence the influence of method variance in the ratings.

The correlations among the methods obviously tend to be higher when results are aggregated across tasks to the total score level (see Figure III.13). Again, the correlations between the two ratings methods are highest, and correlations between rating methods and test methods are in general lowest. The exceptions are among MOS 95B (Military Policy) where the hands-on/knowledge correlation is particularly low, and MOS 11B (Infantryman) where ratings and test results are correlated nearly as highly as the two test methods.

Table III.48 shows overall correlations between hands-on and knowledge tests for MOS grouped by occupational category. The categories used correspond to the Aptitude Area composites identified by McLaughlin, Rossmeissl, Wise, Brandt, and Wang (1984) based on which Armed Services Vocational Aptitude Battery (ASVAB) tests were most predictive of future training performance success for particular Army MOS. These composites were labeled as clerical, operations, combat, and skilled technical. The correlations were clearly lowest in the skilled technical category; otherwise, there were no major differences between groupings.

To know whether these correlations are high or low, some frame of reference is needed. Rumsey, Osborn, and Ford (1985) reviewed 19 comparisons between hands-on and job knowledge tests. For 13 of the 19 comparisons using work samples classified as "motor" because the majority of tasks involved physical manipulation of things (see Asher & Sciarrino, 1974, for a distinction between "motor" and "verbal" work samples), they found a mean correlation of .42 prior to correction for attenuation and .54 following such correction. Results were further divided into occupational categories based primarily on which aptitude areas in the ASVAB best predicted performance for that category, as follows: skilled technical, operations, combat arms, and electronics. Table III.49 shows corrected and uncorrected correlations in each of these categories. An additional category, clerical, was identified, but no investigations using a motor work sample had reported any results in this category.

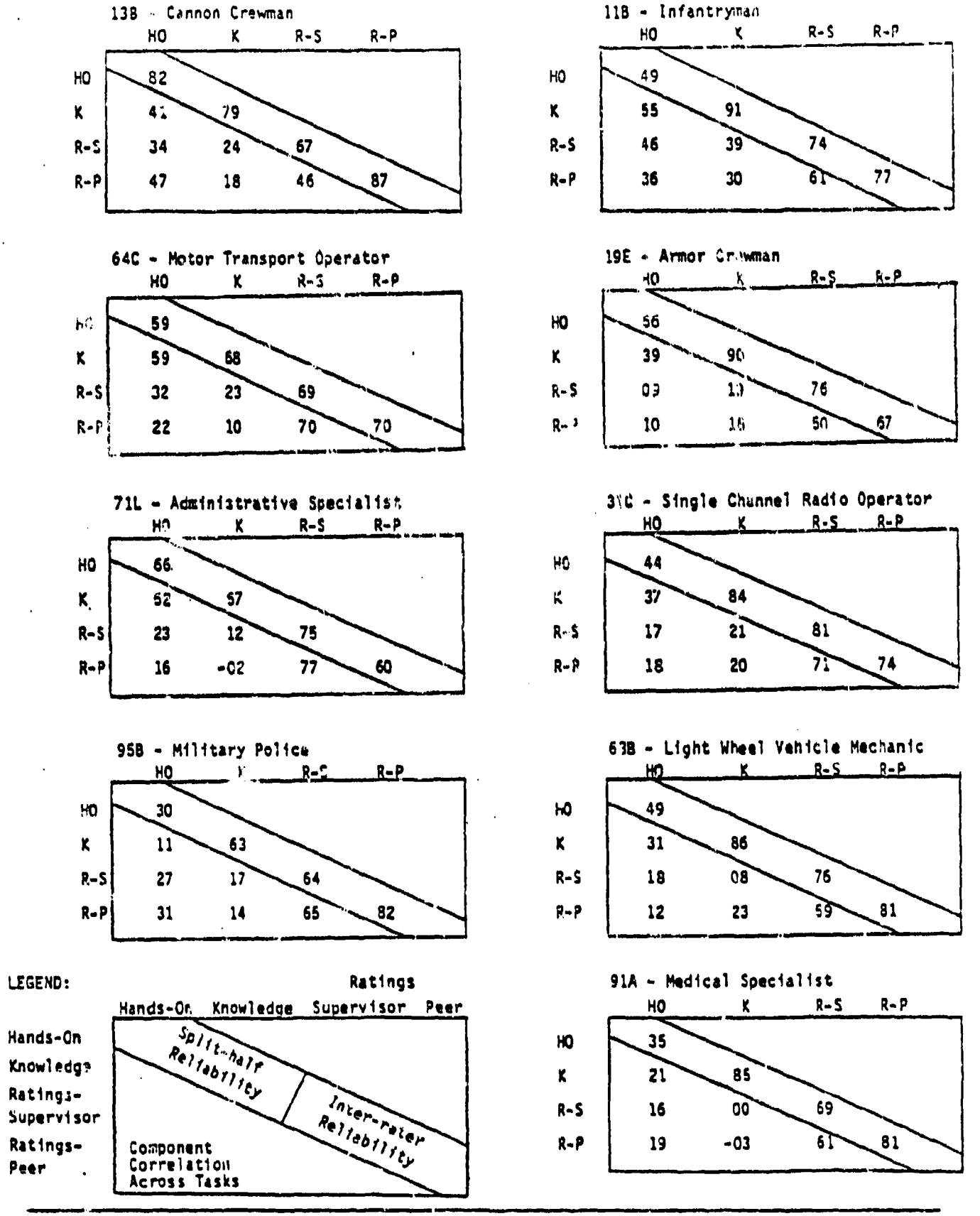


Figure III.13. Reliabilities and correlations between task measurement methods across task for nine MOS.

Table III.48

Correlations Between Hands-On and Knowledge Test Components for MOS Classified by Type of Occupation

Type of Occupation (MOS)	Total Sample Size	Correlation Between Knowledge and Hands-On	
		r[a]	Corrected r[b]
Clerical (71L - Administrative Specialist)	126	.52	.76
Operations (63B - Light Wheel Vehicle Mechanic; 64C - Motor Transport Operator; 31C - Single Channel Radio Operator)	393	.43	.71
Combat (11B - Infantryman; 13B - Cannon Crewman; 19E - Armor Crewman)	414	.46	.67
Skilled Technical (95B - Military Police; 91A - Medical Specialist)	250	.17	.35
OVERALL	1,183	.39	.62

[a] Correlation between knowledge and hands-on test scores averaged across samples.

[b] Correlation between knowledge and hands-on test scores average across samples and corrected for attenuation.

Table III.49

Reported Correlations Between Hands-On (Motor) and Knowledge Tests

	Correlation	
	r[a]	Corrected[b]
Operations	.45	.60
Combat Arms	.47	.62
Skilled Technical	.58	.67
Electronics	.27	.34
ALL	.42	.54

[a] Correlation between knowledge and hands-on test scores averaged across samples.

[b] Correlation between knowledge and hands-on test scores averaged across samples and corrected for attenuation.

In general, the correlations observed in the Project A field tests were at a level consistent with those found in the literature. They were particularly consistent for three MOS, Motor Transport Operator, Infantryman, and Administrative Specialist, that represented three separate occupational groupings. They were low in two skilled technical occupations, Military Police and Medical Specialist. This pattern in the skilled technical groupings does not correspond to findings reported in the literature (Rumsey et al., 1985). Since the Military Police and Medical Specialist occupations were also the MOS for which qualifying scores on the Armed Forces Qualifying Test (AFQT) were higher than in any of the other Project A MOS, there is some reason to believe that restriction in range may have been a factor contributing to the lower correlations.

How reliable are the measures? The weighted average of the split-half reliability estimates shown in Table III.44 for the 30 knowledge tests is .80; this average does not substantially deviate from an average reliability of .83 reported in the literature for job knowledge tests (Rumsey et al. 1985).

The average of the split-half reliability estimates shown in Table III.46 for the 15 hands-on tasks is .52. Ultimately, a 30-task test will be generated for each MOS based on the 15 tasks for which both types of measures have been developed and the 15 tasks for which only job knowledge tests have been developed. Using the Spearman-Brown formula, it can be estimated that the reliability of a 30-task hands-on test would have been .68, relative to an average value of .71 reported in the literature (Rumsey et al., 1985). While the estimates found here clearly were not higher than those previously reported, it should be noted that the overall test development strategy in Project A placed more emphasis on comprehensiveness than on content homogeneity.

Revision of Task-Specific Performance Measures

In revising the hands-on and knowledge tests, the goals for each MOS was a reduction in knowledge test items of 25 to 40% (depending on the MOS), and a set of between 14 and 17 hands-on task tests. Field test experience indicated that reductions of this magnitude would meet the time allotments for Concurrent Validation.

To make these reductions, both the field test results and additional systematic judgments of the "suitability" of hands-on measurement were used.

Hands-On and Knowledge Measures

Developing an appropriate hands-on, or job sample, measure of a particular job task is not always feasible, since it may be difficult to standardize conditions or obtain the necessary equipment. Compromises may be necessary and the question arises as to what effect the compromises have on content validity. To assess such effects during the field tests, the project staff involved in developing and administering the hands-on measures rated the entire pool of hands-on tests according to their suitability for hands-on measurement. The points considered were standardization of conditions, reliability of scorers, and quality of task coverage.

The suitability ratings were then used with other data to further refine the task test sets. First, if the hands-on set was too long (more than 17 tests, or likely to run over 3 hours) after revisions, the developers dropped the hands-on tests that were judged least suitable for hands-on measurement, or that were judged suitable but had very high correlations with the corresponding job knowledge task test, or that had correlations with a similar hands-on task test. However, if dropping such tests would not effect a savings, because the tests were not time-consuming or resource intensive, the tests were often retained. When the hands-on set comprised 14-17 of the best available hands-on tests, the set was considered final.

If, after revisions, the knowledge test set had 60 to 70% as many items as before, the tests were considered feasible for the 2-hour time slot. The knowledge test was then accepted as complete, and finalized for Proponent review.

However, if there were too many items, the strengths and weaknesses of each knowledge and hands-on test were analyzed by means of a procedure used to assign "flaw" points to each test. The flaw point procedure considered whether the test was or was not revised after the field test, test difficulty, variability in scores, reliability, and hands-on suitability. The points assigned were considered in conjunction with an analysis of the specific content overlap between hands-on and knowledge tests. Knowledge tests were reduced to items that tended not to overlap with hands-on tests by considering first the more flawed knowledge tests, and then the knowledge tests that were found to be redundant with hands-on tests. The steps in reduction are described more completely in Appendix N in the ARI Research Note in preparation.

The extent of the changes made on the tests, considering both obtained data and informed judgments, was small. Among common task tests, judgments of hands-on suitability resulted in deleting five tests (Recognize Armored Vehicles, Visually Identify Threat Aircraft, Decontaminate Skin, Move Under Direct Fire, Collect and Report Information). Additionally, in each MOS two to five MOS-specific tasks were dropped as not suitable for hands-on testing. For most MOS, the set of hands-on tests, including those field tested in MOS and tests judged not suitable, comprised 19 to 23 tasks; after suitability judgments were made, the hands-on sets were reduced to 15 to 19 tasks in each MOS. Appendix T in the ARI Research Note in preparation lists the full set of hands-on tests that were developed for all MOS, and indicates, for common tasks, the other MOS for which the tasks were selected and where, therefore, they might also be tested hands-on.

By following the adjustment steps described, each MOS was covered by a set of 15-17 hands-on tests, and a set of knowledge items that was 60 to 70% of the set field-tested. The array of hands-on and knowledge tests for each MOS is summarized in Table III.50; a list of the tests offered for Proponent agency review is presented in Appendix U of the ARI Research Note in preparation.

Table III.50

Summary of Testing Mode Array for MOS Task Tests Before Proponent Review

MOS	Hands-On and Knowledge	Hands-On and Reduced Knowledge	Hands-On Only	Knowledge Only	Total Hands-On	Knowledge Items
13B	8	9	0	13	17	177,181[a]
64C	8	6	2	14	16	168
71L	5	5	5	15	15	148
95B	15	0	0	15	15	210
11B	10	4	1	16	15	198
19E	10	4	1	15	15	196
31C	15	0	0	0	15	215
63B	15	0	0	0	15	196
91A	15	0	0	0	15	234

Job Task Ratings

The high correlations among rating scales, relative to their correlations with other methods, are consistent with previous literature. At this point the question still remains as to whether the "method" variance inherent in the ratings represents relevant or extraneous components of performance. Interrater reliabilities are sufficiently high to warrant retention of the rating scales for the Concurrent Validation.

Also, findings reported by Borman, White, and Gast (1985) using this same field test data set reveal that, for some MOS, overall performance ratings were more closely related to hands-on and job knowledge tests than the task-based ratings examined here. This raises additional questions about whether raters adequately understood and appropriately used the task-based scales.

Job History Questionnaire

The results from the Job History Questionnaire, although far from conclusive, provided sufficient indication that job experience may be an important factor to warrant further consideration of this variable. As a consequence, the Job History Questionnaire is being retained in the Concurrent Validation data collection. Those data, with much larger sample sizes, will be used to identify which, if any, task measures should be corrected for the contaminating effects of differential experience. Furthermore, the relationship between experience and performance may vary as a function of the aptitude being validated and the difficulty of the task. Therefore, care will be taken regarding the possibility of interaction effects as well as covariance effects.

Proponent Agency Review

The final step in the development of hands-on and knowledge tests was Proponent agency review. This step was consistent with the procedure of obtaining input from Army subject matter experts at each major developmental stage.

The Proponent was asked to consider two questions: (a) Do the measures reflect doctrine accurately, and (b) do the measures cover the major aspects of the job? A Proponent representative was given copies of the measures; staffing of the review was left to the discretion of the Proponent agent.

Item changes generally affected fewer than 10% of the items within an MOS and most such changes involved the wording, not the basic content, of the item. Changes affecting the task list occurred in only three MOS. Proponent comments and resulting actions taken are summarized below for each of these MOS.

11B - Infantryman. The Infantry Center indicated that the primary emphasis for infantry should be nonmechanized. To that end, they advised dropping three tasks: Perform PMSC on Tracked or Wheeled Vehicle, Drive Tracked or Wheeled Vehicle, and Operate as a Station in a Radio Net. Two

tasks field tested in other MOS were substituted: Move Over, Through, or Around Obstacles, and Identify Terrain Features on a Map. The Center also concurred with the addition of a hands-on test of the task, Conduct Surveillance Without Electronic Devices; the hands-on test of Estimate Range was dropped in exchange.

71L - Administrative Specialist. The Soldier Support Center, Proponent for MOS 71L, recommended that Post Regulations and Directives be eliminated from the 71L task list. They also recommended that four tasks originally to be tested only in the knowledge mode be tested in the hands-on mode as well: File Documents/Correspondence, Type a Joint Message Form, Type a Military Letter, and Receipt and Transfer Classified Documents. To allow testing time for additions, the following tasks, originally to be tested in both the hands-on and knowledge modes, will now be tested only in the knowledge mode: Perform CPR; Put On/Wear Protective Clothing; Load, Reduce Stoppage, and Clear M16A1 Rifle; and Determine Azimuth with Lensatic Compass. The 71L test set was then composed of 29 tasks, 14 tested in a hands-on mode.

95B - Military Police. The Military Police School, Proponent for MOS 95B, indicated that the role of the military police was shifting toward a more combat-ready, rear area security requirement, rather than the domestic police role emphasized by the tasks selected for testing. They recommended that five tasks be added. Three of these--Navigate from One Position on the Ground to Another Position, Call for and Adjust Indirect Fire, and Estimate Range--had previously been field tested with MOS 11B soldiers. Both hands-on and knowledge tests for these tasks were added to 95B. Another, Use Automated CEOI, had been field tested with MOS 19E soldiers; this task was added to the list of knowledge tests only. The final task, Load, Reduce a Stoppage, and Clear a Squad Automatic Weapon, not previously field tested, was also added to the list of knowledge tests only. Four tasks were dropped. Two--Perform a Wall Search, and Apply Hand Irons--had initially been proposed for both hands-on and knowledge testing. The remaining two--Operate a Vehicle in a Convoy, and Establish/Operate Roadblock/Checkpoint--had been on the knowledge only task list. The 95B test set then consisted of 31 tasks, 16 tested in a hands-on mode.

In determining whether any of these task list changes constituted a major shift in content coverage, special consideration was given to the principle applied in the initial task selection process that every cluster of tasks be represented by at least one task. What impact did the Proponent changes have with respect to this principle? For MOS 71L and MOS 95B, each cluster was still represented after the Proponent changes had been implemented. For MOS 11B, the deletion of Perform PMCS on Tracked or Wheeled Vehicle and Drive Tracked or Wheeled Vehicle left one cluster, consisting of tasks associated with vehicle operation and maintenance, unrepresented. However, since it was the Infantry School's position that tasks in this cluster did not represent the future orientation of the 11B MOS, this omission was considered acceptable.

A second condition in which strict adherence to Proponent suggestions was not necessarily advisable was where the suggestions could not be easily reconciled with documented Army doctrine. Where conflict with documentation emerged, the discrepancy was pointed out; if the conflict was not resolved,

items were deleted. Finally, if Proponent comments seemed to indicate a misunderstanding of the intended purpose or content of the test items, clarification was attempted. The basic approach was to continue discussions until some mutually agreeable solution could be found.

Copies of all tests, reflecting revisions based on field test data adjustments to fit constraints of Concurrent Validation and changes recommended by Proponent agencies, are presented in Appendix V in the ARI Research Note in preparation.

Summary and Discussion

The results of the task-based MOS-specific development effort, from the first perusals of the MOS task domains, through task selection, test development, and field test data collection, to the Proponent review and the final production of criterion measures for Concurrent Validation, are satisfying at several levels. More than 200 knowledge tests and more than 100 hands-on tests were developed and field tested, and the field test experience was applied to the production of criterion measures of more than 200 tasks for the nine MOS. The tests provide broad coverage of each MOS in a manner that is both psychometrically sound and appealing to MOS proponents.

Initial predictions of the capability of Army units to support hands-on tests, and of the ability of SL1 soldiers to comprehend the knowledge tests and rating scales, were largely borne out during data collection. Where serious misadjustments had been made, it was possible to make corrections to eliminate the problems encountered.

The several methodologies developed for defining the task domains, obtaining SME judgments of task criticality and difficulty, selecting tasks for testing, assigning tasks to test modes, and reducing test sets to manageable arrays proved both comprehensive and flexible. The peculiarities of each MOS required that the methods be adapted at various points, yet for every MOS all vagaries were dealt with to the satisfaction of both developers and proponents.

In general, means and standard deviations revealed a reasonable level of performance variability on hands-on and knowledge tests. In one MOS where the variability of hands-on tests was most limited, Military Police, there have been Proponent-directed changes that may result in increased variability in Concurrent Validation testing.

The developmental activities that have been described resulted in the preparation of performance measures to be administered concurrently with predictor measures in a large-scale validation. As this effort is completed, a new set of task-based measures will be developed to measure performance of soldiers in their second tour. It is anticipated that many of the procedures used in developing first-tour measures will be appropriate for this new purpose as well, although some revisions will be needed to accommodate the expanded responsibilities associated with second-tour jobs. Work on developing these revised procedures is already under way.

Section 11

FIELD TEST RESULTS: MOS-SPECIFIC RATINGS (BARS)[1]

This section presents the field test reliability data and scale characteristics for the job (MOS)-specific rating scales that were developed via the BARS method. Before those analyses were carried out, the ratings for the MOS-specific BARS, as well as the other rating scale criterion measures, were adjusted for certain between-rater differences. After describing these adjustments, and the results of the analyses, the section closes with a discussion of the BARS modifications for use in the Concurrent Validation study.

Rating Scale Adjustments

Differences in Rater Levels

One problem with ratings is that although raters might agree on a particular ratee's strengths and weaknesses across different performance dimensions, differences between raters in the level of mean ratings sometimes appear. Consequently, for the Project A ratings measures we decided to compute adjusted scores that would reduce or eliminate the level differences between scores provided by two or more raters for a single ratee. The procedures developed to compute adjusted ratings or scores were as follows:

- Consider the score provided for one enlistee across all other raters. For example, Rater 1 gave Enlistee A a score of 4.0 and Enlistee B a score of 5.0 on Dimension X.
- For each enlistee, compute the mean rating across all other raters. For example, if Raters 2, 3, and 4 evaluated Enlistee A on Dimension X, we computed the mean rating for Enlistee A across these three raters.
- Compare the score for the target rater-enlistee pair with the mean computed for the same enlistee across all other raters. Use these values to compute a mean difference score for the target rater-enlistee pair.
- Repeat this procedure to compute a difference score for each rater-enlistee combination on each performance dimension.

---

[1]The development of these tests is described in Section 4, Part III. Section 11 is based primarily on an ARI Technical Report in preparation, Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS, by Toquam et al., and the supplementary ARI Research Note in preparation, which contains the report appendixes.

- Weight each difference score by the number of other raters evaluating each enlistee.

- For each rater, compute a weighted average difference score for each performance dimension.

- Finally, compute an average difference score across all performance dimensions for the target rater. Use the average difference score to adjust the ratings provided by that rater.

These procedures were used to compute adjusted scores for all raters. Ratings supplied by peers and supervisors were pooled to compute adjusted scores.

Identification of Outliers

The next step involved identifying ratings that appeared unrealistic. Two criteria for identifying questionable raters were developed. First, the correlation between performance dimension ratings for a target rater-enlistee pair and the mean performance dimension ratings provided by all other raters evaluating that enlistee was computed. If this correlation was -.2 or lower for any ratee, all of that rater's ratings were deleted from the data set. Second, any rater who generated an average difference score of 2.0 or greater was deleted from the sample.

All ratings made by any rater whose adjusted scores exceeded one or both of the above criteria were deleted. These ratings were omitted because a negative correlation (-.2 or lower) or a relatively high adjustment score (2.0 in absolute terms) indicated that this rater's data did not correspond to information provided by other raters evaluating the same enlistee(s). The goal for eliminating outliers was to be as conservative as possible by deleting only the most extreme ratings. For each of the MOS by rater type (supervisors or peers) combinations, the number of ratees deleted ranged from zero to seven. Across all MOS and rater types, data were eliminated for only 22 ratees.

For all remaining analyses, we analyzed ratings provided by supervisors separately from ratings provided by peers, using the adjusted scores computed for each rater.

Differences Between Batch A and Batch B Data Sets

When the "raw" ratings were adjusted for level differences between raters, using the procedure described above, some adjusted scores fell outside the actual range of rating scale values. For example, the rating scores for one performance dimension ranged from 0.49 to 7.17. In the analyses conducted for Batch A MOS, the adjusted ratings were allowed to exceed the actual scale point range. For Batch B MOS, adjusted scores were modified by truncating scores that exceeded 7.0 or that fell below 1.0. Thus, in the MOS data the ratings for Batch A exceed the range of 1 to 7, whereas ratings for Batch B fall within this range.

Rater/Ratee Ratio

Assumptions concerning the number of raters evaluating each soldier affect the resulting reliability estimate. Generally, the more raters evaluating a soldier, the higher the estimate. For each group of ratings, the ratio of the number of raters to the number of ratees was calculated. These data are reported in Table III.51. For comparison purposes, the table includes the ratios for rating data computed before and after the ratings were adjusted and screened. Note that these ratios changed very little following the screening process.

Table III.51

Ratio of Raters to Ratees, Before and After Screening, for Supervisors and Peer Ratings on MOS-Specific BARS

MOS	Supervisors		Peers	
	Before	After	Before	After
13B - Cannon Crewman	1.47	1.47	2.83	2.52
64C - Motor Transport Operator	1.84	1.82	2.77	2.57
71L - Administrative Specialist	1.04	1.04	1.90	1.89
95B - Military Police	1.94	1.88	3.67	3.39
11B - Infantryman	1.81	1.81	2.99	2.99
19E - Armor Crewman	1.68	1.68	2.95	2.95
31C - Radio Teletype Operator	1.73	1.73	2.49	2.50
63B - Light-Wheel Vehicle Mechanic	1.77	1.77	2.08	2.09
91A - Medical Specialist	1.59	1.59	3.10	3.10

For a majority of enlistees in each MOS, there were ratings from two supervisors. For the Administrative Specialist, however, only one supervisor rating for each enlistee was obtained. This reflects the job content for most administrative specialists, which makes it difficult to obtain very many supervisor or peer ratings. These specialists do tend to work alone and for one boss only.

For peer ratings, the ratio of raters to ratees ranged from 1.89 for Administrative Specialist (71L) to 3.39 for Military Police (95B) with a median value of 2.57. We obtained at least two peer ratings for every enlistee with the exception of Administrative Specialists. For enlistees in four of the MOS--Military Police (95B), Infantryman (11B), Armor Crewman (19E), and Medical Specialist (91A)--there were about three peer ratings for each.

For the reliabilities presented below, estimates were adjusted so that reliability computed for peer ratings provided for Batch A MOS samples can be interpreted as the expected correlation between the mean ratings provided by equivalent groups of four peers. For Batch B MOS peer ratings, however, three rather than four peers were assumed. Interrater reliability estimates computed for supervisors can be interpreted as assuming that all soldiers were rated by two supervisors.

Descriptive Statistics for MOS BARS Ratings

Supervisor and Peer Ratings

Table III.52 presents the means, standard deviations, ranges, and reliability (interrater agreement) of the rating scores on each of the individual MOS BARS scales.

Supervisor and peer ratings yielded similar levels of reliability estimates. Across all MOS, median reliability estimates for supervisor ratings range from .53 for Infantryman (11B) to .66 for Medical Specialist (91A) with a median value of .57. For peer ratings, median values range from .43 for Armor Crewman (19E) to .65 for Military Police (95B) with a median value of .55. The median values indicate that for single-item scales, interrater reliability estimates are at a respectable level. The reliability estimates reported were adjusted for the number of raters for each ratee. Given equal numbers of supervisor and peer raters for each ratee, the data indicate that the supervisor ratings would be somewhat more reliable than the peer ratings.

Supervisors and peers also provided similar information about the mean level of performance. Across the nine MOS, peers provided slightly higher grand mean values than supervisors in two MOS, Administrative Specialist (71L) and Infantryman (11B). Supervisors provided slightly higher grand mean values than peers in two MOS, Motor Transport Operator (64C) and Military Police (95B). Mean ratings by the two groups were nearly identical for the remaining MOS.

Scale Intercorrelations

A summary of the average scale intercorrelations for supervisors, for peers, and between supervisors and peers is shown in Table III.53. Average intercorrelations among performance dimension ratings for supervisors and peers are similar. The greatest difference between mean intercorrelations for supervisors and peers occurs for Military Police (95B), with the mean value for supervisors at .39 and mean value for peers at .58.

Revision of the MOS-Specific BARS for Administration to the Concurrent Validation Sample

Prior to the administration of the MOS-specific rating scales in the Concurrent Validation study, the scales were submitted to a Proponent review to verify that critical first-term job requirements were represented in the

Table III.52

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

A. Cannon Crewman (13B)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Loading Out Equipment	141	5.02	1.17	0.65 - 7.67	.58	140	4.83	0.85	2.19 - 6.62	.54
B. Driving and Maintaining Vehicles, Howitzers and Equipment	141	5.14	1.17	1.67 - 7.67	.63	141	5.05	1.00	1.76 - 7.06	.59
C. Transporting/Sorting/Storing and Preparing Ammunition for Fire	141	4.98	1.09	0.65 - 6.86	.61	141	5.04	0.90	2.28 - 6.69	.59
D. Preparing for Occupation/ Emplacing Howitzer	141	4.89	1.14	1.48 - 6.87	.45	141	4.71	0.98	2.21 - 6.87	.53
E. Setting Up Communications	141	4.81	1.04	2.48 - 7.67	.43	141	4.95	0.80	2.19 - 6.59	.42
F. Gunnery	141	4.48	1.26	1.42 - 7.76	.44	141	4.47	1.22	0.76 - 7.37	.54
G. Loading/Unloading Howitzer	141	5.19	1.10	0.65 - 7.67	.35	141	5.02	1.05	2.05 - 8.87	.66
H. Receiving and Relaying Communications	141	4.84	1.11	1.48 - 6.87	.40	141	4.94	0.83	2.05 - 6.87	.40
I. Recording/Record Keeping	140	4.64	1.31	0.65 - 7.35	.61	141	4.61	1.03	2.05 - 6.87	.61
J. Position Improvement	141	4.95	1.03	1.48 - 7.67	.33	141	4.92	0.93	2.19 - 7.87	.51
K. Overall Cannon Crewman Performance	141	4.92	1.15	0.65 - 7.67	.70	141	4.91	0.95	2.19 - 7.87	.61
Mean Ratings	141	4.89	0.81	1.88 - 6.81	.73	141	4.85	0.71	2.49 - 6.97	.80

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

B. Motor Transport Operator (64C)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Driving Vehicles	147	5.06	1.01	1.39 - 6.97	.54	154	4.92	1.08	0.17 - 8.49	.63
B. Vehicle Coupling	149	4.80	1.12	1.39 - 7.17	.57	154	4.67	0.99	1.17 - 7.49	.53
C. Checking and Maintaining Vehicles	149	4.91	1.12	0.49 - 6.97	.48	154	4.98	0.94	2.03 - 7.19	.58
D. Using Maps/Following Proper Routes	147	4.84	0.97	0.94 - 7.08	.59	154	4.69	1.05	1.76 - 6.91	.56
E. Loading Cargo and Transporting Personnel	147	4.91	0.98	1.32 - 7.02	.66	153	4.88	0.88	2.03 - 7.49	.54
F. Parking and Securing Vehicles	147	5.11	1.04	1.39 - 7.08	.47	154	5.39	0.85	2.49 - 7.69	.42
G. Performing Administrative Duties	147	4.91	0.97	1.89 - 7.94	.56	154	4.87	0.75	2.49 - 7.19	.32
H. Self-Recovering Vehicles	145	4.40	1.02	1.73 - 6.39	.53	154	4.23	0.93	1.13 - 6.32	.54
I. Safety-Mindedness	148	4.85	1.00	1.88 - 6.49	.66	154	5.00	0.92	2.01 - 7.49	.60
J. Performing Dispatcher Duties	138	4.16	1.10	1.52 - 6.47	.60	152	3.78	1.09	0.59 - 6.23	.36
K. Overall Motor Transport Operator Performance	147	4.82	0.97	1.82 - 6.60	.62	154	5.03	0.95	1.61 - 7.49	.58
Mean Ratings	154	5.07	0.73	2.80 - 6.70	.74	154	4.74	0.66	2.49 - 6.79	.82

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

C. Administrative Specialist (71L)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Preparing, Typing, and Proofreading Document	105	4.48	1.38	1.00 - 7.02	--	64	4.94	0.95	2.50 - 7.22	.50
B. Distributing and Dispatching Incoming and Outgoing Documents	104	4.52	1.23	1.00 - 7.13	--	63	4.49	0.86	2.00 - 6.28	.54
C. Maintaining Office Resources	107	4.66	1.20	2.00 - 7.02	--	63	4.73	0.91	1.96 - 7.31	.44
D. Posting Regulations	105	4.11	1.29	1.00 - 7.02	--	63	4.47	0.88	2.41 - 6.78	.48
E. Establishing and/or Maintaining Files in Accordance with TAFFS	106	4.27	1.44	1.00 - 7.17	--	63	4.55	1.04	1.56 - 7.31	.46
F. Keeping Records	107	4.42	1.25	2.00 - 7.63	--	63	4.77	1.00	2.00 - 6.28	.49
G. Safeguarding and Monitoring Security of Classified Documents	95	4.57	1.13	1.80 - 7.00	--	63	4.32	1.02	1.89 - 6.71	.55
H. Providing Customer Service	107	5.26	1.35	2.00 - 7.63	--	64	5.48	1.09	1.96 - 7.22	.37
I. Overall Administrative Specialist Performance	107	4.65	1.27	1.00 - 8.03	--	64	5.24	0.81	3.00 - 7.00	.55
Mean Ratings	107	4.52	0.94	1.86 - 6.65	--	64	4.72	0.64	2.84 - 5.90	.81

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

D. Military Police (958)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Traffic Control and Enforcement	113	4.67	0.91	2.07 - 6.31	.54	113	4.60	0.69	3.11 - 6.19	.68
B. Providing Security	112	4.77	0.82	2.59 - 6.68	.39	112	4.61	0.64	3.08 - 6.25	.39
C. Investigating Crimes and Making Arrests	113	4.50	0.89	2.59 - 6.98	.57	112	4.42	0.78	2.64 - 6.43	.63
D. Patrolling	111	4.55	1.02	1.59 - 6.67	.56	112	4.51	0.87	1.88 - 7.19	.67
E. Promoting the Public Image of the Military Police	113	4.29	1.03	2.01 - 7.19	.57	112	4.19	0.87	2.28 - 5.88	.66
F. Interpersonal Communications Skills	112	4.36	0.88	1.67 - 7.19	.49	112	4.36	0.81	2.16 - 6.16	.59
G. Responding to Medical Emergencies	112	4.12	0.87	2.17 - 6.09	.52	112	4.38	0.63	2.39 - 5.89	.60
H. Overall Military Police Performance	113	4.57	0.94	1.59 - 6.57	.74	113	4.75	0.76	2.36 - 6.19	.71
Mean Ratings	113	4.47	0.63	2.81 - 5.85	.76	113	4.43	0.60	2.74 - 5.74	.83

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

E. Infantryman (11B)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Maintaining Supplies, Equipment, and Weapons	149	4.53	1.03	2.08 - 6.90	.63	172	4.55	0.78	2.37 - 6.29	.48
B. Assisting and Leading Others	148	4.06	1.06	1.65 - 6.15	.48	172	4.34	0.88	2.10 - 6.34	.49
C. Navigation	149	4.00	1.02	1.50 - 6.30	.47	172	4.22	0.99	2.12 - 6.52	.64
D. Use of Weapons and Other Equipment	149	4.73	0.97	1.67 - 6.96	.55	172	4.61	0.75	2.58 - 6.28	.55
E. Field Sanitation, Personal Hygiene, and Safety	149	4.77	1.04	1.18 - 7.00	.54	172	4.80	0.84	1.76 - 7.00	.43
F. Fighting Position	149	4.40	1.05	1.39 - 7.00	.49	172	4.33	0.86	1.99 - 6.03	.56
G. Avoiding Enemy Detection	149	4.28	1.01	1.58 - 6.30	.42	172	4.50	0.74	1.87 - 6.24	.30
H. Operating a Radio	149	4.64	1.05	2.22 - 6.96	.61	172	4.57	0.88	1.76 - 6.70	.60
I. Reconnaissance and Patrol	149	4.28	0.98	1.39 - 6.90	.55	172	4.43	0.88	1.90 - 6.44	.56
J. Guard and Security Duties	149	4.31	1.10	1.28 6.51	.53	172	4.60	0.81	2.32 6.48	.50
K. Courage and Proficiency in Battle	149	4.77	0.85	2.13 6.96	.33	172	4.63	0.85	2.26 6.47	.57
L. Prisoners of War	149	4.51	0.89	2.22 6.51	.29	172	4.32	0.81	2.24 6.03	.37
M. Overall Infantryman Performance	149	4.63	0.93	1.22 6.67	.53	172	4.76	0.79	2.12 6.70	.58
Mean Ratings	149	4.45	0.70	2.55 - 5.98	.78	172	4.51	0.62	2.84 - 6.05	.81

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

F. Armor Crewman (19E)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Maintaining Tank, Tank Systems, and Associated Equipment	154	4.56	0.89	2.17 - 6.57	.55	163	4.57	0.78	1.85 - 6.46	.55
B. Driving and Recovering Tanks	153	4.67	1.07	1.15 - 6.57	.63	163	4.79	0.82	2.25 - 6.46	.42
C. Stowing Ammunition Aboard Tanks	154	5.08	0.72	2.57 - 6.76	.50	163	4.86	0.71	2.25 - 7.00	.29
D. Loading/Unloading Guns	154	5.23	0.90	2.01 - 7.00	.57	163	5.01	0.79	2.25 - 6.89	.37
E. Maintaining Guns	154	4.86	0.84	2.17 6.79	.46	163	4.83	0.79	2.83 6.49	.51
F. Engaging Targets with Tank Guns	146	4.35	1.15	1.57 - 6.60	.73	163	4.38	0.98	1.45 - 6.82	.51
G. Operating and Maintaining Communications Equipment	154	4.39	0.92	2.20 - 6.78	.57	163	4.51	0.80	1.81 - 6.81	.38
H. Preparing Tanks for Field Problems	154	4.79	0.94	1.22 - 6.78	.64	163	4.92	0.82	1.45 - 6.94	.43
I. Overall Armor Crewman Performance	153	4.85	0.82	2.17 - 7.00	.68	163	4.96	0.85	1.95 - 7.00	.65
Mean Ratings	154	4.75	0.58	2.56 - 6.11	.87	163	4.76	0.56	3.19 - 5.87	.76

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

G. Radio Teletype Operator (31C)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Inspecting and Servicing Equipment	125	4.26	1.08	1.00 - 6.43	.68	120	4.38	0.89	1.00 - 7.00	.52
B. Installing and Repairing Equipment	126	4.76	1.01	2.00 - 7.00	.63	121	4.64	0.85	2.78 - 7.00	.64
C. Operating Communications Device	125	4.88	1.07	1.50 - 7.00	.57	121	4.84	0.86	2.00 - 5.43	.56
D. Preparing Reports	125	4.36	1.01	1.68 - 7.00	.58	121	4.41	1.03	1.76 - 7.00	.64
E. Maintaining Security	125	4.93	1.12	1.71 - 7.00	.67	121	4.91	0.88	2.78 - 7.00	.60
F. Providing Safe Transportation	125	4.83	1.11	1.00 - 7.00	.63	121	4.55	0.89	1.00 - 6.28	.53
G. Overall Radio Teletype Operator Performance	125	4.80	1.16	1.78 - 6.71	.70	121	4.80	1.02	1.26 - 6.76	.69
Mean Ratings	126	4.68	0.86	2.43 - 6.25	.80	121	4.66	0.69	2.78 - 6.54	.80

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

H. Light-Wheel Vehicle Mechanic (63B)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Inspecting, Testing, and Detecting Problems with Equipment	141	4.31	1.14	1.00 - 6.96	.64	128	4.39	1.02	1.67 - 7.00	.68
B. Troubleshooting	141	4.19	1.16	1.00 - 7.00	.63	128	4.26	1.05	1.00 - 6.83	.63
C. Performing Routine Maintenance	142	4.86	1.03	1.43 - 7.00	.43	128	4.85	1.04	1.43 - 7.00	.70
D. Repair	142	4.71	1.12	1.00 - 7.00	.57	128	4.69	1.04	1.99 - 7.00	.66
E. Using Tools and Test Equipment	142	4.45	1.21	1.00 - 7.00	.59	128	4.40	1.01	1.00 - 6.43	.40
F. Using Technical Documentation	142	4.58	1.23	1.00 - 7.00	.66	128	4.37	0.94	1.33 - 7.00	.42
G. Vehicle and Equipment Operation	141	4.92	1.12	1.43 - 7.00	.50	128	4.92	1.00	1.00 - 7.00	.63
H. Safety Mindedness	142	4.60	1.08	1.98 - 7.00	.64	128	4.32	0.95	2.25 - 6.43	.49
I. Administrative Duties	143	4.03	1.21	1.08 - 7.00	.57	128	4.25	1.12	1.00 - 7.00	.64
J. Planning/Organizing Jobs	143	3.96	1.17	1.00 - 7.00	.61	128	4.11	1.05	1.33 6.28	.42
K. Recovery	137	4.51	1.17	1.00 - 7.00	.62	127	4.34	0.98	1.17 6.42	.35
L. Overall Light-Wheel Vehicle Mechanic Performance	142	4.69	1.14	1.00 - 7.00	.67	128	4.76	1.04	1.08 7.00	.54
Mean Ratings	143	4.48	0.87	1.33 - 6.67	.85	128	4.47	0.73	2.42 - 6.20	.86

Table III.52 (Continued)

Means, Standard Deviations, Ranges, and Reliability Estimates for MOS-Specific BARS, by MOS

I. Medical Specialist (91A)

Scale	Supervisors					Peers				
	N	Mean	SD	Range	Rxx	N	Mean	SD	Range	Rxx
A. Maintaining and Operating Army Medical Vehicles and Equipment	154	4.40	1.24	1.00 - 6.72	.70	148	4.61	0.95	1.40 - 7.00	.45
B. Maintaining Accountability of Medical Supplies and Equipment	156	4.48	1.22	1.63 - 7.00	.62	157	4.54	0.91	1.45 - 7.00	.65
C. Keeping Medical Records	144	4.53	1.11	1.13 - 6.69	.75	158	4.62	1.03	1.00 - 6.80	.66
D. Arranging for Transportation and/or Transporting Injured Personnel	148	4.89	1.00	1.72 - 7.00	.48	155	4.73	0.94	1.70 - 7.00	.60
E. Dispensing Medications	138	4.79	0.97	2.17 - 7.00	.61	155	4.86	0.89	1.40 - 6.90	.55
F. Preparing and Inspecting Field Site or Clinic Facilities	146	4.39	1.11	1.77 - 7.00	.68	150	4.45	0.84	1.64 - 6.80	.44
G. Providing Routine and Ongoing Patient Care	152	5.00	1.05	2.22 - 7.00	.45	158	4.89	0.94	1.45 - 7.00	.64
H. Responding to Emergency Situations	154	5.17	1.14	1.22 - 7.00	.72	158	4.93	0.94	1.45 - 7.00	.54
I. Providing Health Care and Health Maintenance Instructions to Army Personnel	155	4.54	1.12	1.13 - 7.00	.63	158	4.45	0.98	1.40 - 6.70	.68
J. Overall Medical Specialist Performance	156	4.89	1.04	1.22 7.00	.69	158	4.88	0.91	1.45 7.00	.67
Mean Ratings	156	4.71	0.79	1.82 - 6.50	.85	158	4.71	0.72	1.70 - 6.49	.84

Table III.53

Average Intercorrelations of MOS-Specific BARS for Supervisors, for Peers, and Between Supervisors and Peers

MOS		Statistic	Within Supervisors	Within Peers	Between Supervisors and Peers[a]	
					Scale in Common	Scale not in common
13B	Cannon Crewman	$\bar{r}$	.46	.50	.39	.33
		SD	.12	.07	.10	.09
64C	Motor Transport Operator	$\bar{r}$	.48	.42	.43	.38
		SD	.12	.16	.10	.09
71L	Administrative Specialist	$\bar{r}$	.42	.36	.38	.37
		SD	.14	.11	.11	.12
9B	Military Police	$\bar{r}$	.39	.58	.44	.41
		SD	.15	.07	.08	.09
11B	Infantryman	$\bar{r}$	.42	.50	.41	.34
		SD	.10	.08	.07	.08
19E	Armor Crewman	$\bar{r}$	.29	.35	.28	.25
		SD	.11	.13	.10	.09
31C	Radio Teletype Operator	$\bar{r}$	.53	.49	.43	.38
		SD	.05	.09	.13	.07
63B	Light-Wheel Vehicle Maintenance	$\bar{r}$	.53	.43	.43	.40
		SD	.10	.13	.10	.10
91A	Medical Specialist	$\bar{r}$	.45	.53	.45	.38
		SD	.08	.09	.10	.09

[a] The first column is the average of supervisor x peer intercorrelation when both are using the same rating scale dimension. The second column is the average of all "off diagonal" intercorrelations; that is, when supervisor and peer are rating the same person but not using the same rating scale.

performance scales. Revisions were made in the MOS-specific behaviorally anchored rating scales, using results from the field test as well as input supplied by the Proponent review committee.

Revisions Based on Field Test Data

For each MOS, the reliability estimates computed for performance dimension ratings provided by supervisors were compared with estimates for dimension ratings provided by peers to identify problem dimensions. (See Table III.54 for a summary of the median reliability estimates as well as the range of reliabilities for each MOS.)

For most MOS, there appears to be no consistent pattern when reliability estimates computed for supervisor ratings are compared with those computed for peer ratings. Within MOS 95B one performance dimension, Providing Security, appeared to present problems for both rater groups. The interrater reliability estimate computed separately for supervisors and peers is .39. Therefore, the definition as well as the behavioral anchors for this particular dimension were clarified.

For the remaining MOS-specific rating scales, performance dimensions with low reliability estimates for supervisor or peer ratings were identified and the rating scale definitions and anchors developed for these dimensions were reviewed. Anchors and definitions were revised if it seemed appropriate.

Table III.55 contains the adjusted and unadjusted grand mean values by MOS and by rater type. Grand mean values computed using adjusted scores correspond very highly with grand mean values computed using unadjusted scores. Since very little leniency or central tendency error is exhibited in Table III.55, no changes were made in the scales as the result of these data.

Revisions Based on Proponent Review

Following the Batch B field test administration, the nine MOS-specific behaviorally anchored rating scales were each submitted to a Proponent committee for review. Proponent committee members, who were primarily technical school subject matter experts from each MOS, studied the scales and made suggestions for modifications. For most MOS, suggestions made by committee members were minor wording changes. For example, they noted a problem with one of the anchors in one Administrative Specialist (71L) performance dimension, Keeping Records. The committee recommended deleting one anchor from this dimension because it described job duties typically required of second-term personnel only (i.e., Handle Suspense Dates). Therefore, this anchor was omitted.

For another MOS, Radio Teletype Operator (31C), the Proponent review committee noted that the job title had been changed, and the necessary changes were made on all Concurrent Validation rating forms. The current MOS-specific rating form for this MOS now reads "Single Channel Radio Operator--31C."

Table III.54

Summary of Reliability Estimates of MOS-Specific BARS for Supervisor and Peer Ratings

MOS	Supervisors			Peers		
	Median	Range		Median	Range	
13B Cannon Crewman	.54	.33	J. Position Improvement	.54	.40	H. Receiving and Relaying Communications
		.70	K. Overall Performance		.66	G. Load/Unload Howitzers
64C Motor Transport Operator	.57	.47	F. Parking and Securing Vehicles	.54	.32	G. Performing Administrative Duties
		.66	I. Safety Mindedness		.68	D. Using Maps and following proper Routes
			E. Loading Cargo/Tranporting Personnel			
71L Administrative Specialist	Not calculated - Insufficient Number of Pairs			.46	.37	H. Providing Customer Service
					.55	G. Safeguarding and Monitoring Security of Classification Documents
						I. Overall Performance
95B Military Police	.55	.39	B. Providing Security	.65	.39	B. Providing Security
		.74	H. Overall Performance		.71	H. Overall Performance
11B Infantryman	.53	.29	L. Prisoners of War	.55	.30	G. Avoiding Enemy Detection
		.63	A. Maintaining Supplies, Equipment and Weapons		.64	C. Navigation
19E Armor Crewman	.57	.46	E. Maintaining Guns	.43	.29	C. Stowing Ammunition Aboard Tanks
		.73	F. Engaging Targets with Tank Guns		.65	I. Overall Performance
31C Radio Teletype Operator	.63	.57	C. Operating Communication Devices	.60	.52	A. Inspecting and Servicing Equipment
		.70	G. Overall Performance		.69	G. Overall Performance
63B Light-Wheel Vehicle Mechanic	.62	.43	C. Performing Routine Maintenance	.59	.35	K. Recovery
		.67	L. Overall Performance		.70	C. Performing Routine Maintenance
91A Medical Specialist	.66	.45	G. Providing Routine and Ongoing Patient Care	.62	.44	F. Preparing and Inspecting Field Site or Clinic Facilities
		.75	C. Keeping Medical Records		.68	I. Providing Health Care Instruction to Army Personnel

Table III.55

Summary of Grand Mean Values for Unadjusted and Adjusted BARS Ratings by MOS[a]

MOS			Supervisors		Peers	
			Unadjusted	Adjusted	Unadjusted	Adjusted
13B	Cannon Crewman	Mean Median	4.39(1.13) 4.90	4.89(0.81) 4.92	4.89(0.84) 4.97	4.85(0.71) 4.92
64C	Motor Transport Operator	Mean Median	4.92(1.02) 5.00	5.07(0.73) 4.85	4.66(0.83) 4.78	4.74(0.66) 4.88
71L	Administrative Specialist	Mean Median	4.56(1.13) 4.57	4.52(0.94) 4.52	4.75(0.81) 4.79	4.72(0.64) 4.73
95B	Military Police	Mean Median	4.59(0.75) 4.59	4.47(0.63) 4.58	4.43(0.66) 4.41	4.43(0.60) 4.47
11B	Infantryman	Mean Median	4.39(0.91) 4.44	4.46(0.70) 4.51	4.56(0.70) 4.60	4.51(0.60) 4.55
19E	Armor Crewman	Mean Median	4.89(0.78) 4.91	4.78(0.68) 4.79	4.75(0.60) 4.84	4.76(0.56) 4.83
31C	Radio Teletype Operator	Mean Median	4.46(0.93) 4.57	4.68(0.86) 4.80	4.88(0.86) 4.87	4.66(0.69) 4.64
63B	Light-Wheel Vehicle Mechanic	Mean Median	4.34(0.98) 4.41	4.48(0.87) 4.59	4.64(0.81) 4.54	4.47(0.73) 4.38
91A	Medical Specialist	Mean Median	4.71(0.83) 4.70	4.71(0.79) 4.67	4.72(0.76) 4.70	4.71(0.72) 4.68

[a] Standard deviations are shown in parentheses.

For one MOS, Military Police (95B), the committee asked for more extensive changes. Committee members noted that because critical incident workshops were conducted only in CONUS locations, a few requirements of the Military Police job were missing. Incumbents in this MOS are required to provide combat and combat support functions. Therefore, four performance dimensions describing these requirements were added to the Military Police MOS-specific rating scales: Navigation (Dimension H); Avoiding Enemy Detection (Dimension I); Use of Weapons and Other Equipment (Dimension J); and Courage and Proficiency in Battle (Dimension K). Definitions and behavioral anchors for these scales had been developed for the Infantryman (11B) performance dimensions rating scales. Proponent committee members reviewed these definitions and anchors and authorized including the same information in the Military Police performance rating scales.

Project Review and Revision

Following the Batch B field test sessions, Project A staff members reviewed the final set of rating scales. This group, the Criterion Measurement Task Force, was composed of project personnel responsible for developing task-oriented and behavior-oriented measures.

Most members of the Task Force had participated in administering criterion measures during the Batch A and Batch B field tests. They reported that some of the rating scales, the behaviorally anchored scales in particular, required considerable reading time, and they felt that some raters were not reading the scales thoroughly before making their ratings. The panel recommended that the length of the behavioral anchors be reduced to ensure that all raters would review the anchors thoroughly before using them to evaluate incumbents.

The performance dimension definitions and scale anchors were modified accordingly. The goal was to retain the specific job requirements and depiction of ineffective, adequate, or effective performance in each anchor while eliminating unnecessary information or lengthy descriptions. Figure III.14 shows an example of the anchors for one performance dimension in the Military Police (95B) rating scales, as they appeared for the Batch B administration and as they appear for the Concurrent Validation study.

A complete description of the rating scales administered in the Concurrent Validation study is given in the MOS appendixes in the ARI Research Note in preparation.

Before - BATCH A Field Test Administration

A. TRAFFIC CONTROL AND ENFORCEMENT

Controlling traffic and enforcing traffic laws and parking rules.

1 2	3 4 5	6 7
Below Standard	Adequate/Mid-Range	Superior
• Often uses hand/arm signals that are difficult to understand, at times resulting in unnecessary accidents; often fails to wear reflectorized gear; overlooks hazardous traffic conditions; sleeps on duty; pays excessive attention to things unrelated to the job. • May display excess leniency or harshness when citing offenders, allowing their military rank, race, and/or sex to influence his/her actions; makes many errors when filling out citations.	• Usually does a reasonable job when directing traffic by using adequate hand/arm signals and/or wearing reflectorized gear. • Makes few errors when filling out citations; usually does not allow an offender's race, sex, and/or military rank to interfere with good judgment.	• Consistently uses appropriate hand/arm signals; always wears reflectorized gear; generally monitors traffic from plain-view vantage points; consistently refrains from behaviors such as reading and prolonged conversation on non-job related topics. • Always uses emergency equipment (e.g., flares, barricades) to highlight unsafe conditions and ensures that hazards are removed or otherwise taken care of.

III-150

After - Concurrent Validity Study

A. TRAFFIC CONTROL AND ENFORCEMENT

How effective is each soldier in controlling traffic and enforcing traffic laws and parking rules?

① ②	③ ④ ⑤	⑥ ⑦
Often uses hand/arm signals that are difficult to understand or fails to wear reflectorized gear; overlooks many hazardous traffic conditions and violations. May allow the military rank, race, or sex of traffic offenders to influence his/her enforcement of traffic laws; makes many errors when filling out citations.	Generally uses adequate hand/arm signals and wears reflectorized gear when directing traffic; usually pays attention to traffic conditions and enforces traffic laws. Usually does not allow a traffic offender's military rank, race, or sex to influence his/her enforcement when filling out citations.	Always uses adequate hand/arm signals and wears reflectorized gear when directing traffic; monitors traffic carefully from plain-view vantage points and enforces all traffic laws Never allows a traffic offender's military rank, race, or sex to influence his/her enforcement of traffic laws; rarely or never makes errors when filling out citations.

Figure III.14. Sample Performance Rating Scale before and after modifications for Military Police (95B) MOS-Specific BARS.

Section 12

FIELD TEST RESULTS: ARMY-WIDE RATING MEASURES[1]

Analyses of the field test data from the Army-wide rating measures focused on (a) distributions of the ratings, (b) interrater reliabilities, and (c) intercorrelations among the rating scale dimensions.

Prior to these analyses, the same rater adjustments and outlier analyses were conducted as were done for the MOS-specific ratings (see Section 11). A relatively small number of raters (9 supervisors and 46 peers out of the total sample of 904 supervisors and 1,205 peers) were identified as outliers. Because these raters' ratings were so severely discrepant from other raters' ratings of the same target soldiers, their data were excluded from further analyses.

Statistics From Field Test

Distributions of Ratings

Table III.56 presents frequency distributions of ratings made on each of the seven points on the 7-point Army-wide rating scales. Table III.57 then depicts the means and standard deviations of selected composite ratings as well as the Overall Effectiveness and NCO Potential scales. Taken together, findings from the two tables suggest that raters did not succumb to excessive leniency (overly high ratings) or restriction-in-range (rating everyone at about the same level). The modal rating of 5 on a 7-point scale and means generally between 4 and 5 seem reasonable in that we would expect the first-term performer to be a little above average, because some percentage of the poor performers will have already left the Army.

Interrater Reliability

Interrater reliability results appear in Table III.58. In general, the levels of the reliabilities are encouraging. Intraclass correlations for the composites of the Army-wide behavioral dimensions are almost uniformly in the 80s (median = .84). Reliabilities of the individual behavioral scales are lower (.51-.68, median = .58) but still respectable. The Overall Effectiveness and NCO Potential reliabilities are likewise reasonably high (.47-.82, median = .66). Regarding the Army-wide common task ratings, interrater reliabilities for the dimension composites are satisfactory (.55-.84, median = .71), but not as high as the behavioral dimension composites. Individual common task scale interrater reliabilities are lower (.33-.60, median = .44).

---

[1]The development of the Army-wide rating measures was described in Section 5, Part III. Section 12 is primarily based on ARI Technical Report 716, Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program, Elaine D. Pulakos and Walter C. Borman (Eds.), and the supplementary ARI Research Note 87-22, which contains the report appendixes.

Table III.56

Frequency Distributions (Percent) of Ratings Across the Seven Points of the Army-Wide Measures[a]

MOS	Scale Points 1	2	3	4	5	6	7
			Individual Army-wide Behavioral Dimensions				
11B	4/2	8/6	13/11	20/23	25/31	18/18	11/8
13B	3/4	7/6	13/11	17/18	24/30	23/23	13/9
19E	2/2	8/7	13/15	22/24	28/30	18/15	10/6
31C	3/3	9/6	14/12	19/18	30/32	16/19	9/10
63B	2/2	10/7	16/13	20/21	27/31	15/17	9/9
64C	4/3	9/6	12/12	19/21	26/34	18/18	12/6
71L	2/1	6/4	14/12	17/25	26/30	20/19	15/8
91A	4/3	9/8	12/13	19/22	27/28	18/18	12/9
95B	2/2	6/6	15/16	25/26	29/30	16/17	8/4
			Overall Effectiveness				
11B	2/1	6/2	17/12	24/26	30/37	16/19	4/3
13B	2/3	6/5	16/8	25/15	24/36	17/26	10/8
19E	0/1	4/4	10/12	25/28	42/33	15/19	3/3
31C	1/1	6/4	15/8	25/21	36/37	12/21	4/8
63B	2/1	8/4	14/12	29/25	30/38	14/15	4/4
64C	3/2	9/5	18/7	18/16	30/42	17/24	6/3
71L	0/0	7/3	15/9	27/31	29/32	20/23	2/2
91A	2/1	5/4	16/13	24/26	27/35	21/17	5/4
95B	1/1	4/3	14/11	23/27	32/38	20/19	7/2
			NCO Potential				
11B	9/4	18/12	15/18	17/19	22/25	15/17	5/4
13B	11/17	5/7	10/8	18/12	27/32	17/21	10/13
19E	3/6	11/10	13/17	21/24	27/26	20/13	5/4
31C	7/5	15/6	14/10	16/18	27/25	15/27	5/9
63B	6/4	13/10	17/14	21/19	21/29	15/17	6/7
64C	8/8	6/10	13/13	21/20	27/30	16/15	8/4
71L	2/0	4/3	11/12	18/24	38/39	19/15	10/7
91A	8/7	14/10	14/15	15/20	22/25	18/16	9/8
95B	6/7	5/7	10/15	21/23	25/24	18/18	14/7
			Individual Army-wide Common Task Dimensions				
11B	2/1	6/4	13/10	18/20	27/29	20/23	15/13
13B	3/3	5/3	9/8	19/15	28/28	23/29	13/14
19E	1/1	3/3	9/10	19/22	33/28	22/24	13/13
31C	1/1	3/3	9/7	20/19	29/30	22/24	16/16
63B	2/2	5/3	12/12	20/20	30/29	21/23	10/12
64C	3/3	5/6	12/11	20/20	28/34	25/19	7/6
71L	2/3	6/6	12/15	20/19	25/31	26/23	8/4
91A	2/3	4/4	10/10	18/20	27/27	25/21	16/14
95B	0/2	2/4	11/12	26/25	31/32	20/20	10/4

[a] In each cell, the percentage for supervisors is on the left and the percentage for peers is on the right. The scale values range from Poor (1) to Excellent (7).

Table III.57

Means and Standard Deviations of Selected Army-Wide Rating Measures[a]

	MOS								
	11B	13B	19E	31C	63B	64C	71L	91A	95B
Average Army-Wide Behavioral Dimensions									
Supervisors	4.50 (.82)	4.76 (.90)	4.46 (.65)	4.59 (.88)	4.42 (.87)	4.52 (.91)	4.73 (.77)	4.56 (.95)	4.44 (.79)
Peers	4.53 (.68)	4.67 (.75)	4.47 (.59)	4.54 (.76)	4.49 (.76)	4.56 (.71)	4.78 (.65)	4.57 (.81)	4.46 (.66)
Overall Effectiveness									
Supervisors	4.47 (1.02)	4.59 (1.23)	4.48 (.94)	4.55 (1.08)	4.38 (1.14)	4.36 (1.22)	4.39 (1.16)	4.57 (1.18)	4.56 (1.10)
Peers	4.62 (.76)	4.85 (.99)	4.67 (.76)	4.71 (.95)	4.52 (.95)	4.75 (.95)	4.73 (.93)	4.63 (.93)	4.64 (.91)
NCO Potential									
Supervisors	3.97 (1.37)	4.34 (1.55)	4.26 (1.23)	4.28 (1.42)	4.14 (1.36)	4.30 (1.37)	4.76 (1.27)	4.23 (1.48)	4.59 (1.35)
Peers	4.14 (1.08)	4.66 (1.27)	4.23 (1.06)	4.56 (1.24)	4.31 (1.18)	4.14 (1.26)	4.76 (.93)	4.29 (1.27)	4.35 (1.13)
Average Army-Wide Common Task Dimensions									
Supervisors	4.87 (.66)	4.97 (.70)	5.02 (.55)	5.07 (.61)	4.87 (.65)	4.53 (.63)	4.53 (.81)	4.91 (.71)	4.70 (.53)
Peers	4.96 (.61)	4.99 (.68)	4.93 (.47)	5.12 (.61)	4.84 (.77)	4.54 (.56)	4.75 (.69)	4.95 (.68)	4.63 (.57)

[a] The mean is based on a 7-point scale ranging from Poor (1) to Excellent (7). The standard deviation is shown in parentheses. The means, standard deviations, interrater reliabilities, and intercorrelations appear in Appendix E of ARI Research Note 87-22, for each individual Army-wide behavioral dimension, and in Appendix F for each individual Army-wide common task dimension.

Table III.58

Intraclass Correlation Coefficients for Selected Army-Wide Rating Measures

	MOS								
	11B	13B	19E	31C	63B	64C	71L[a]	91A	95B
ICCs for Average Behavioral Dimensions									
Supervisors	.82	.81	.86	.83	.84	.84	--	.81	.85
Peers	.80	.83	.78	.86	.84	.85	.82	.86	.88
Mean ICCs Across Individual Behavioral Dimensions									
Supervisors	.58	.58	.46	.60	.60	.58	--	.60	.63
Peers	.55	.61	.55	.60	.57	.58	.51	.67	.68
ICCs for Overall Effectiveness									
Supervisors	.64	.62	.54	.70	.63	.72	--	.74	.82
Peers	.47	.60	.48	.65	.71	.66	.70	.68	.79
ICCs for NCO Potential									
Supervisors	.74	.61	.53	.71	.63	.68	--	.64	.68
Peers	.57	.63	.59	.74	.66	.69	.60	.69	.68
ICCs for Average Common Tasks									
Supervisors	.77	.70	.74	.55	.55	.60	--	.71	.74
Peers	.78	.72	.67	.64	.84	.65	.57	.79	.82
Mean ICCs Across Individual Common Tasks									
Supervisors	.42	.48	.38	.38	.42	--	--	.46	.41
Peers	.51	.47	.46	.41	.51	.34	.33	.60	.57

[a] ICCs were not computed for 71L supervisor raters because almost all of the ratees were evaluated by only one supervisor.

Supervisor and peer ratings have very similar levels of interrater reliability. Median reliabilities were computed for supervisors and peers separately, first for all of the behavioral dimension entries in Table III.58 and then for the common task scale reliability values. The peer ratings are slightly higher in average reliability than those of the supervisors (supervisors: BARS median = .66, task median = .55; peers: BARS median = .68, task median = .59).

It should be noted that the data in Table III.58 are intraclass correlation coefficients (ICC) representing the reliabilities of mean ratings across supervisors or peers and, accordingly, are dependent on the average number of raters per ratee. Just as adding items to a test increases its reliability, larger rater/ratee ratios yield higher reliabilities as a function of the Spearman-Brown formula. Considering the present rater/ratee ratios (about 2.8 for peers vs. 1.8 for supervisors), the supervisor ratings would have been somewhat more reliable than peer ratings if each source had had the same number of raters per ratee.

However, the coefficients appearing in the table provide the appropriate reliability estimates (of the mean supervisor and mean peer ratings), because correlations between the rating data and other variables were calculated using the mean supervisor and mean peer rating for each ratee. That is, ratings of a given ratee were averaged across supervisors and across peers, and all of the correlations reported here were computed on these means. The sample size for each correlation is the number of ratees on which it was calculated.

Rating Scale Intercorrelations

The intercorrelations and the cross-correlations of the individual scales for the Army-wide ratings (supervisor) and the MOS-specific BARS ratings (supervisor) are shown in Table III.59 for all MOS. The average of the within-measure scale correlations and the average of the scale cross-correlations are also shown.

Overall, the scale intercorrelations are perhaps not as high as are usually found for rating scale intercorrelations and they are certainly lower than the individual scale reliabilities. This is particularly significant because the scale reliabilities (i.e., the intraclass r) incorporate rater differences as error while the scale intercorrelations do not (i.e., all correlations are based on the same set of raters).

In general, the correlations between scales taken from the different measures (the cross-correlations) are slightly lower than the within-measure scale intercorrelations). However, the differences are not as great as one might expect, given the different objectives of the two measures.

Revision of the Army-Wide Scales

As with the MOS-specific BARS scales, experience administering the Army-wide rating scales during Batch A indicated that some soldiers had difficulty with the amount of reading required. It thus seemed prudent here

Table III.59

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

A. Cannon Crewman (13B)

	BARS	MOS-Specific BARS												Army-Wide BARS											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
MOS-Specific																											
1	Average Rating	--							Among MOS-Specific BARS, $\underline{r}$ = .46																		
2	Lead Out Equipment	77	--																								
3	Drive/Maintain Howitzers	70	61	--																							
4	Prepare Ammo for Fire	80	69	62	--															Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .37							
5	Prepare to Emplace Howitzer	72	46	44	59	--																					
6	Set up Communications	67	50	58	48	41	--																				
7	Gunnery	66	43	40	33	20	46	--																			
8	Load/Unload Howitzer	73	55	36	51	41	42	66	--																		
9	Receive/Relay Comm.	72	41	36	53	70	32	30	43	--																	
10	Record Keeping	66	40	23	46	54	19	42	39	60	--																
11	Position Improvement	73	44	44	58	44	56	46	48	47	40	--															
12	Overall Job Performance	68	53	43	49	31	54	66	61	30	35	65	--														
Army-Wide																											
13	Average Rating	72	56	60	56	48	53	53	43	49	44	56	67	--													
14	A - Technical Skill	65	48	51	61	52	41	38	42	54	34	44	46	69	--					Among Army-Wide BARS, $\underline{r}$ = .45							
15	B - Effort	72	60	57	60	49	53	45	40	48	46	55	47	69	64	--											
16	C - Following Regs	55	37	40	47	51	31	28	27	54	37	42	45	75	60	54	--										
17	D - Integrity	48	38	47	34	32	44	41	29	23	15	46	55	78	45	54	64	--									
18	E - Leadership	61	43	49	56	61	36	27	27	64	36	41	32	66	63	65	58	49	--								
19	F - Maintenance Equipment	65	55	64	60	37	55	37	38	43	32	47	55	77	55	63	54	56	54	--							
20	G - Maintenance Areas	40	31	26	29	26	36	33	26	23	34	24	44	68	29	21	44	49	25	50	--						
21	H - Military Appearance	37	31	25	19	18	27	39	20	24	34	27	41	68	21	27	39	46	27	41	65	--					
22	I - Physical Fitness	31	23	41	23	07	30	33	23	05	11	29	41	55	29	22	19	29	17	37	36	42	--				
23	J - Self-Development	53	40	42	33	18	45	57	41	20	31	56	70	76	48	52	39	60	31	52	45	51	56	--			
24	K - Self-Control	36	30	29	20	22	22	34	22	23	30	24	47	72	31	24	48	55	23	41	61	61	42	60	--		
25	Z - Overall Effectiveness	57	43	51	37	21	52	61	39	22	27	55	69	77	53	54	48	66	39	54	50	53	52	76	55	--	
26	Y - NCO Potential	65	49	52	59	57	28	38	30	53	51	45	47	63	56	54	58	44	68	46	32	32	21	41	37	49	--

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

B. Motor Transport Operator (64C)

BARS	MOS-Specific BARS												Army-Wide BARS											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
MOS-Specific																										
1 Average Rating								Among MOS-Specific BARS, $\underline{r}$ = .48																		
2 Drive Vehicles	24	--																								
3 Vehicle Coupling	20	51	--																							
4 Check/Maintain Vehicle	30	61	59	--															Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .43							
5 Use Maps/Follow Routes	22	62	46	46	--																					
6 Lead Cargo/Transport Pers	18	59	63	64	61	--																				
7 Park/Secure Vehicle	18	39	53	65	36	57	--																			
8 Perform Adm. Duties	03	46	33	56	41	57	58	--																		
9 Self-Recover Vehicles	10	41	41	53	48	46	45	43	--																	
10 Safety-Mindedness	15	59	46	59	40	59	64	64	47	--																
11 Perform Dispatcher Duties	14	26	23	33	21	22	31	50	47	45	--															
12 Overall Job Performance	22	61	61	71	53	73	63	67	48	73	42	--														
Army-Wide																										
13 Average Rating	25	58	51	75	48	59	54	67	46	62	48	77	--													
14 A - Technical Skill	29	59	58	69	56	59	54	58	55	59	43	74	83	--					Among Army-Wide BARS, $\underline{r}$ = .55							
15 B - Effort	20	48	48	68	53	54	48	50	37	50	35	62	82	70	--											
16 C - Following Regs	21	46	48	57	46	48	39	48	28	52	35	63	84	67	67	--										
17 D - Integrity	22	49	48	63	46	51	45	55	41	54	33	64	82	71	70	74	--									
18 E - Leadership	28	40	42	57	39	46	36	47	46	46	42	63	80	69	63	64	62	--								
19 F - Maintenance Equipment	65	47	54	66	42	64	58	56	32	55	31	68	79	70	63	65	63	56	--							
20 G - Maintenance Areas	13	43	42	60	30	43	46	58	23	48	34	58	77	57	54	57	59	50	60	--						
21 H - Military Appearance	12	34	24	54	30	33	36	53	27	39	30	46	75	50	50	55	51	51	51	21	--					
22 I - Physical Fitness	12	30	09	40	15	18	24	38	34	32	29	31	53	31	32	35	33	45	25	27	42	--				
23 J - Self-Development	16	48	34	52	38	39	32	51	32	39	52	65	79	60	57	63	59	53	57	57	57	39	--			
24 K - Self-Control	09	41	34	55	28	40	36	56	28	51	35	51	68	52	51	61	57	44	51	54	61	41	54	--		
25 Z - Overall Effectiveness	20	46	47	67	32	47	54	53	37	53	45	64	81	70	69	66	66	66	63	61	56	47	66	64	--	
26 Y - NCO Potential	23	46	43	55	35	47	43	53	32	52	31	64	78	60	65	70	63	68	60	50	52	47	58	57	78	--

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

C. Administrative Specialist (71L)

	BARS	MOS-Specific BARS												Army-Wide BARS										
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
MOS-Specific																									
1	Average Rating								Among MOS-Specific BARS, $\underline{r}$ = .42																
2	Prepare Doc	76	--																						
3	Distribute Doc	85	61	--																					
4	Maintain Off. Resources	65	46	51	--																				
5	Post RegulationsRoutes	74	49	59	43	--																			
6	Est. Files and Tapes	79	51	66	48	57	--																		
7	Keep Records	74	52	61	37	41	51	--																	
8	Safeguard Classified Doc	48	18	34	15	25	22	31	--																
9	Provide Customer Service	65	45	42	28	38	38	48	19	--															
10	Overall Job Performance	81	69	68	46	62	56	62	30	63	--														
Army-Wide																									
11	Average Rating	68	53	59	43	41	52	56	31	47	66	--													
12	A - Technical Skill	75	65	60	42	56	56	54	39	50	71	68	--												
13	B - Effort	44	41	41	34	33	34	26	14	26	43	62	54	--											
14	C - Following Regs	46	35	43	28	28	35	41	25	34	43	72	48	47	--										
15	D - Integrity	48	34	41	32	28	48	38	21	29	43	71	38	39	54	--									
16	E - Leadership	58	41	43	41	45	60	33	21	41	48	59	58	55	37	41	--								
17	F - Maintenance Equipment	32	36	12	22	10	12	31	28	28	29	53	31	14	32	35	22	--							
18	G - Maintenance Areas	40	32	42	22	13	31	46	14	24	41	58	31	29	36	37	14	30	--						
19	H - Military Appearance	16	18	11	10	00	09	26	05	13	21	40	13	06	14	11	03	26	37	--					
20	I - Physical Fitness	01	07	08	06	06	09	04	10	06	07	37	03	13	10	14	02	09	01	35	--				
21	J - Self-Development	45	27	42	29	38	30	31	27	28	47	56	47	18	38	41	30	26	14	05	16	--			
22	K - Self-Control	29	25	32	13	15	22	32	01	22	33	62	18	22	43	42	21	21	38	24	28	31	--		
23	Z - Overall Effectiveness	60	52	53	36	45	48	45	17	42	64	69	65	65	54	42	58	24	37	17	14	33	37	--	
24	Y - NCO Potential	54	41	53	23	39	48	40	25	43	53	64	53	47	59	53	43	15	25	10	23	36	43	62	--

Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .28

Among Army-Wide BARS, $\underline{r}$ = .27

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

D. Military Police (95B)

BARS	MOS-Specific BARS												Army-Wide BARS										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
MOS-Specific																							
1 Average Rating								Among MOS-Specific BARS, $\underline{r}$ = .39															
2 Traffic Control	74	--																					
3 Provides Security	76	58	--																				
4 Investigate Crime	78	60	56	--															Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .36				
5 Patrolling	75	50	61	53	--																		
6 Promote Public Image	58	26	24	26	22	--																	
7 Communication Skills	64	29	36	38	37	48	--																
8 Respond to Emergencies	57	25	38	46	25	24	20	--															
9 Overall Job Performance	82	57	60	69	56	54	52	46															
Army-Wide																							
10 Average Rating	75	56	58	57	45	55	52	39	78	--													
11 A - Technical Skill	70	43	51	64	34	55	40	54	74	77									Among Army-Wide BARS, $\underline{r}$ = .44				
12 B - Effort	57	36	43	40	27	49	40	37	65	79	67	--											
13 C - Following Regs	61	53	43	51	38	39	52	17	64	77	52	60	--										
14 D - Integrity	55	28	47	36	36	41	47	21	56	81	55	64	61	--									
15 E - Leadership	70	51	60	50	39	49	45	48	73	78	74	65	56	60	--								
16 F - Maintenance Equipment	48	39	28	41	24	49	20	27	52	69	54	52	49	47	44	--							
17 G - Maintenance Areas	45	37	39	27	28	36	36	16	43	72	37	51	51	53	44	52	--						
18 H - Military Appearance	51	50	40	49	34	29	26	18	49	68	47	33	51	45	40	47	49	--					
19 I - Physical Fitness	14	25	07	05	25	60	17	00	07	22	05	06	11	08	04	15	18	23	--				
20 J - Self-Development	53	28	47	43	32	35	29	46	58	70	60	54	36	52	59	43	43	50	11	--			
21 K - Self-Control	53	31	40	40	35	40	44	24	54	75	48	50	59	64	50	36	54	44	11	52	--		
22 Z - Overall Effectiveness	75	62	61	60	52	45	48	34	76	83	71	66	73	62	74	51	52	58	15	61	53	--	
23 Y - NCO Potential	70	47	56	66	44	40	40	48	77	73	73	60	63	50	69	46	32	51	02	59	53	77	

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

E. Infantryman (11B)

	BARS	MOS-Specific BARS															Army-Wide BARS												
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
MOS-Specific																													
1	Average Rating	--							Among MOS-Specific BARS, $\underline{r}$ = .42																				
2	Overall Performance	81	--																										
3	Maint Sup. Equip. Weapons	77	68	--																									
4	Assist/Lead Others	78	64	65	--															Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .38										
5	Navigation	74	55	53	62	--																							
6	Use of Weapons/Equipment	75	56	59	54	56	--																						
7	Field Sanitation/Safety	63	46	49	40	33	45	--																					
8	Fighting Position	68	54	35	50	46	43	38	--																				
9	Avoid Enemy Detection	70	44	44	45	50	46	43	48	--																			
10	Operate Radio	57	51	43	40	44	37	23	30	36	--																		
11	Reconnaissance & Patrol	72	58	52	57	51	51	37	48	41	26	--																	
12	Guard/Security Duties	60	51	46	49	35	32	37	41	42	19	39	--																
13	Courage/Prof. in Battle	73	56	48	50	53	55	39	59	49	35	56	31	--															
14	Prisoners of War	54	39	32	27	30	47	26	31	43	30	35	19	44	--														
Army-Wide																													
15	Average Rating	84	73	70	74	61	61	55	59	51	49	57	52	57	38	--													
16	A - Technical Skill	69	63	66	64	57	54	36	46	33	45	47	36	51	31	77	--			Among Army-Wide BARS, $\underline{r}$ = .41									
17	B - Effort	63	60	62	62	47	42	24	41	41	40	44	45	41	22	71	65	--											
18	C - Following Regs	58	49	50	50	37	44	43	38	42	36	37	38	39	25	74	54	56	--										
19	D - Integrity	62	55	40	44	49	42	42	47	40	38	43	50	47	20	73	48	51	54	--									
20	E - Leadership	68	62	55	73	56	46	33	51	39	44	49	41	43	31	80	62	62	51	51	--								
21	F - Maintenance Equipment	62	58	57	51	46	49	25	38	42	48	33	40	41	40	71	55	49	42	54	57	--							
22	G - Maintenance Areas	52	34	40	41	24	43	60	42	34	14	38	30	36	24	62	30	20	37	40	42	43	--						
23	H - Military Appearance	51	42	43	37	31	42	58	34	27	25	30	21	34	37	66	43	24	44	37	41	36	62	--					
24	I - Physical Fitness	38	32	22	19	23	26	33	26	26	20	31	08	42	26	45	27	27	24	32	23	24	31	38	--				
25	J - Self-Development	54	45	41	57	47	36	25	42	32	37	41	31	34	18	71	54	45	47	42	65	52	42	41	22	--			
26	K - Self-Control	30	24	30	24	30	33	25	31	20	14	16	21	21	26	51	27	18	35	34	39	33	29	32	10	34	--		
27	Z - Overall Effectiveness	77	71	67	70	60	47	48	56	44	45	54	49	57	33	81	72	65	60	56	62	52	36	49	31	51	30	--	
28	Y - MOO Potential	73	63	62	75	52	48	49	54	47	39	53	49	51	28	83	67	64	62	58	70	50	43	47	29	54	32	76	--

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

f. Armor Crewman (19E)

	BARS	MOS-Specific BARS												Army-Wide BARS										
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
MOS-Specific																									
1	Average Rating								Among MOS-Specific BARS, $\underline{r}$ = .29																
2	Overall Perf	78	--																						
3	Maint. Tank	67	58	--																					
4	Drive/Recover Tanks	58	42	38	--															Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .23					
5	Stow Ammo Aboard Tanks	55	27	32	36	--																			
6	Load/Unload Guns	58	31	20	27	47	--																		
7	Maint. Guns	63	44	30	14	24	40	--																	
8	Engage Targets w/Tank Guns	67	50	45	33	23	09	40	--																
9	Operate/Maint. Comm. Equip.	55	42	15	10	11	27	36	24	--															
10	Prep Tank for Fld Problems	68	54	43	18	24	31	37	31	44	--														
Army-Wide																									
11	Average Rating	69	56	51	41	34	39	46	40	41	44	--													
12	A - Technical Skill	59	50	24	43	30	33	44	30	41	38	58	--							Among Army-Wide BARS, $\underline{r}$ = .25					
13	B - Effort	55	47	53	41	25	30	33	44	30	41	66	49	--											
14	C - Following Regs	28	26	13	14	18	11	16	18	16	22	54	17	31	--										
15	D - Integrity	41	34	34	11	24	29	30	17	28	33	56	31	43	35	--									
16	E - Leadership	48	31	31	40	23	25	31	29	27	32	63	46	41	17	27	--								
17	F - Maintenance Equipment	51	34	48	28	22	22	31	40	26	37	64	41	51	19	22	53	--							
18	G - Maintenance Areas	23	25	16	10	10	13	12	13	28	33	52	18	21	27	13	24	35	--						
19	H - Military Appearance	28	20	27	13	17	13	12	22	09	26	55	10	15	29	17	18	29	39	--					
20	I - Physical Fitness	22	21	13	06	15	19	20	19	05	13	44	19	08	13	13	12	12	23	49	--				
21	J - Self-Development	42	25	29	24	26	37	38	13	37	16	66	42	42	19	35	45	44	27	19	22	--			
22	K - Self-Control	05	11	06	03	09	08	10	05	00	09	37	10	05	44	24	06	07	11	30	23	12	--		
23	Z - Overall Effectiveness	60	57	36	41	21	35	38	35	39	39	82	53	54	31	39	53	54	47	38	33	52	18	--	
24	Y - NCO Potential	60	52	46	45	19	28	40	40	37	30	77	46	54	36	36	62	47	23	37	16	50	24	66	--

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

6. Radio Teletype Operator (31C)

	BARS	MOS-Specific BARS												Army-Wide BARS									
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
MOS-Specific																							
1	Average Rating								Among MOS-Specific BARS, $\underline{r}$ = .53														
2	Overall Perf	86	--																				
3	Inspect/Service Equip.	78	67	--																			
4	Install/Repair Equip.	81	69	58	--																		
5	Operate Comm. Device	78	58	54	65	--																	
6	Prepare Report	73	59	54	46	48	--																
7	Maint. Security	76	58	52	52	52	40	--															
8	Provide Safe Transportation	79	63	48	60	55	50	58	--														
Army-Wide																							
9	Average Rating	75	64	65	58	52	54	59	63	--													
10	A - Technical Skill	66	59	55	63	46	42	51	47	70	--												
11	B - Effort	64	55	59	51	47	45	46	51	80	61	--											
12	C - Following Regs	60	48	59	41	47	49	53	48	74	35	61	--										
13	D - Integrity	54	42	49	32	34	37	53	47	71	43	56	62	--									
14	E - Leadership	68	53	58	55	58	45	48	60	85	65	70	58	55	--								
15	F - Maintenance Equipment	61	49	60	47	44	43	46	49	70	51	60	55	43	61	--							
16	G - Maintenance Areas	44	37	40	34	31	33	32	39	72	32	47	50	43	27	46	--						
17	H - Military Appearance	40	31	36	30	28	31	39	34	63	33	32	46	35	44	38	70	--					
18	I - Physical Fitness	31	26	25	17	18	17	31	34	54	40	40	32	41	32	31	26	30	--				
19	J - Self-Development	49	39	43	36	35	37	39	39	66	44	51	45	39	59	28	37	28	41	--			
20	K - Self-Control	40	40	23	29	32	40	26	33	61	31	34	39	33	43	28	49	40	18	42	--		
21	Z - Overall Effectiveness	75	66	64	57	53	53	55	65	91	70	71	60	68	80	63	61	47	50	57	49	--	
22	Y - NCO Potential	61	55	53	53	43	43	44	47	83	57	70	57	51	75	54	52	39	33	55	50	80	

Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .41

Among Army-Wide BARS, $\underline{r}$ = .44

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

H. Light-Wheel Vehicle Mechanic (63B)

	BARS	MOS-Specific BARS													Army-Wide BARS												
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
MOS-Specific																												
1	Average Rating	--							Among MOS-Specific BARS, $\underline{r}$ = .53																			
2	Overall Performance	82	--																									
3	Inspect/Test w/Equipment	79	68	--																								
4	Troubleshooting	85	73	77	--															Between MOS-Specific and Army-Wide BARS, $\underline{r}$ = .47								
5	Perform Routine Maintenance	80	70	72	69	--																						
6	Repair	78	68	66	69	67	--																					
7	Use Tools/Test Equipment	74	61	46	63	49	50	--																				
8	Use Tech Document	79	55	55	60	58	54	62	--																			
9	Vehicle/Equipment Operation	71	50	57	50	63	47	44	47	--																		
10	Safety Mindedness	70	49	44	50	49	50	55	58	53	--																	
11	Administrative Duties	73	48	51	53	46	44	55	59	50	49	--																
12	Plan/Organize Jobs	81	65	53	68	56	61	64	70	45	52	64	--															
13	Recovery	59	52	43	48	43	46	31	33	43	35	36	40	--														
Army-Wide																												
14	Average Rating	90	73	74	76	76	69	67	71	64	61	69	72	47	--													
15	A - Technical Skill	78	72	73	70	69	62	47	56	60	52	46	57	49	77	--				Among Army-Wide BARS, $\underline{r}$ = .44								
16	B - Effort	68	57	58	60	58	56	42	48	44	44	54	52	46	74	55	--											
17	C - Following Regs	62	52	46	48	55	48	52	50	41	47	40	50	24	76	48	53	--										
18	D - Integrity	56	44	36	47	45	47	48	50	42	39	40	46	27	66	43	52	54	--									
19	E - Leadership	74	62	63	64	66	49	56	59	56	48	54	66	34	77	68	50	51	45	--								
20	F - Maintenance Equipment	65	38	49	54	44	44	53	61	43	38	68	53	37	72	45	59	50	51	50	--							
21	G - Maintenance Areas	57	40	38	49	40	47	56	50	27	43	49	52	19	68	38	39	54	40	45	52	--						
22	H - Military Appearance	54	39	49	45	48	37	44	43	38	48	42	44	19	70	49	30	55	34	49	45	62	--					
23	I - Physical Fitness	54	45	49	48	48	46	29	36	42	33	43	44	30	57	45	41	33	25	36	37	32	33	--				
24	J - Self-Development	70	55	57	62	55	48	54	63	45	46	49	57	41	76	57	53	51	40	61	53	52	49	38	--			
25	K - Self-Control	40	34	30	31	39	30	35	34	27	25	29	32	23	51	26	26	50	35	28	24	29	40	13	37	--		
26	Z - Overall Effectiveness	81	73	74	69	69	64	52	61	61	56	62	60	42	85	75	60	58	50	67	62	50	55	51	62	30	--	
27	Y - NCO Potential	72	65	61	58	60	58	49	44	63	46	56	51	49	78	59	62	49	50	62	46	41	49	48	49	32	67	

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

Table III.59 (Continued)

Intercorrelation Matrixes for MOS-Specific BARS and Army-Wide BARS by MOS

!. Medical Specialist (91A)

	BARS	MOS-Specific BARS											Army-Wide BARS											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
MOS-Specific																										
1	Average Rating								Among MOS-Specific BARS, r = .45																	
2	Overall Perf	88	--																							
3	Maint/Op Med Vehicle/Equip	72	58	--																						
4	Maint Med Supplies/Equip	73	64	58	--															Between MOS-Specific and Army-Wide BARS, r = .39						
5	Keep Med Records	75	65	47	57	--																				
6	Arrange/Transport Injured	68	58	47	40	39	--																			
7	Dispense Medications	72	58	46	35	52	45	--																		
8	Prepare/Inspect Field Clinic	69	54	53	50	36	56	46	--																	
9	Provide Patient Care	64	54	38	37	51	25	43	26	--																
10	Respond to Emergency	72	70	38	38	46	53	56	34	52	--															
11	Provide Health Care to Army	74	60	43	49	57	40	56	49	45	43															
Army-Wide																										
12	Average Rating	78	70	39	62	64	50	49	55	48	48	62	--													
13	A - Technical Skill	62	57	46	50	46	42	48	36	38	48	47	70	--						Among Army-Wide BARS, r = .45						
14	B - Effort	69	60	53	60	55	43	39	53	42	47	47	78	59	--											
15	C - Following Regs	67	60	58	50	48	54	47	46	37	38	49	80	51	62	--										
16	D - Integrity	59	48	48	49	51	33	38	46	34	36	47	75	43	59	63	--									
17	E - Leadership	70	62	50	58	57	49	38	51	36	45	60	83	62	64	60	61	--								
18	F - Maintenance Equipment	56	47	54	45	46	41	29	42	32	31	38	67	40	59	53	52	58	--							
19	G - Maintenance Areas	48	45	40	34	49	21	34	32	40	24	34	69	31	42	50	52	48	48	--						
20	H - Military Appearance	47	43	35	31	49	31	25	33	39	24	38	72	44	42	47	41	58	38	63	--					
21	I - Physical Fitness	32	30	27	14	20	26	30	29	17	25	22	46	36	30	23	13	36	19	21	41	--				
22	J - Self-Development	60	55	51	52	40	41	43	42	29	38	46	75	53	60	67	54	53	47	42	47	35	--			
23	K - Self-Control	38	35	26	35	36	18	27	19	25	17	39	55	27	24	54	46	39	22	50	33	05	32	--		
24	Z - Overall Effectiveness	68	62	46	56	55	41	36	52	43	43	55	89	62	75	67	66	71	56	55	63	33	63	45	--	
25	Y - NCO Potential	59	57	38	52	50	34	27	43	37	34	54	81	53	60	55	57	70	48	48	59	34	50	41	82	--

Note: All ratings shown are supervisor ratings. Decimal points have been omitted in correlations.

also to reduce the length of the behavioral anchors on the Army-wide behavior-based scales. This was accomplished by editing each behavioral statement to remove unnecessary language and reduce the reading difficulty.

In addition, it was felt that a few of the statements anchoring the different effectiveness levels were multidimensional. That is, the example behaviors contained in certain individual anchors were sufficiently different to cause raters potential confusion regarding the level at which a ratee should be evaluated. This potential problem was addressed by extrapolating more global performance information from the specific behaviors and writing the scale anchors to reflect these more general performance levels. The changes were similar to those illustrated for the MOS-specific BARS (see Section II).

Another revision between the Batch A and Batch B administrations was to drop 1 of the 13 common task scales. This was done simply because a 13th scale would have required an additional page on the printed version of the scales. The task dimension that had the lowest interrater reliability and seemed the most redundant with others was eliminated for Batch B and the Concurrent Validation effort. The final version of these scales, as well as the Army-wide BARs, is shown in Appendix C of ARI Research Note 87-22.

Finally, after the Batch B administration, the instruments were submitted to Proponent review. In this review, technical school subject matter experts studied the scales and made suggestions for minor wording changes on some of the anchors. Also, the dimension Maintaining Living/Work Areas was dropped to reduce the length of time required to complete the behavioral rating scales. Proponent review experts judged that dimension to be the least important and the most expendable.

In summary, only minimal changes were made to the Army-wide rating scales as a result of the field tests: first, eliminating one behavioral dimension and one common task dimension to improve administrative efficiency; second, making relatively minor wording changes and reducing the length of the scale anchors to lessen the reading difficulty as well as the time required to complete the scales.

Summary and Conclusions

Results of the field tests for the Army-wide measures are very encouraging. In particular: (a) rater participants seemed reasonably accepting of the rating program and appeared able to understand and comply with the instructions; (b) rating distributions were acceptable, with means a little above the scale midpoints and standard deviations comparable to those found in other research; and (c) interrater reliabilities were acceptably high, for both supervisor and peer raters.

Although results from both batch A and B field tests were on the whole positive, valuable information for improving the Army-wide scales was gleaned from these trial administrations. To obtain the best possible program, we requested that each batch B rating session administrator provide written feedback on his/her experiences, outlining any suggestions for possible

program improvement. While no major changes were required, several suggestions were made to facilitate program administration for Concurrent Validation and to prevent errors in completing the rating forms.

Because of the importance of rater orientation and training to the use of ratings as criteria, an experiment was conducted on certain parameters. Section 13 reports on that experiment.

Section 13

RATER ORIENTATION AND TRAINING[1]

The rater orientation and training program was seen as very important for reaching the objective of obtaining high-quality ratings. Recent reviews of research on rater training conclude that training is likely to improve performance appraisals (Landy & Farr, 1980; Zedeck & Cascio, 1982). Studies have shown that rating errors such as halo and leniency can be reduced by appropriate training (Borman, 1975, 1979; Brown, 1968; Latham, Wexley, & Purcell, 1975). Also, the accuracy of performance ratings has been enhanced using rater training programs (McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986).

Project staff experience with the training of raters suggested that even brief rater training sessions can result in ratings with reasonably good psychometric characteristics. For example, in research that employed 5-15 minutes of rater training, mean ratings have been between 5 and 6 on a 9-point scale, with standard deviations between 1.25 and 2.00. Interpretable factor analyses have resulted, suggesting that halo was not overly severe, and interrater reliabilities have been in the .55-.85 range (e.g., Borman, Rosse, Abrahams, & Toquam, 1979; Hough, 1984a; Peterson & Houston, 1980).

As a starting point for Project A, a rater training program that staff members had developed and revised over the past several years was adapted for use in this project.

Components of Rater Training

Initial Program for Batch A Testing

Components of the initial rater orientation and training program were as follows:

1. Rater selection guidelines were prepared. Where feasible, supervisors and four peers were identified for each first-tour soldier ratee. To be eligible to rate a soldier, the supervisor or peer must be familiar with the ratee's performance and have supervised or worked with the ratee for at least 2 months.

---

[1]This section is based primarily on ARI Technical Report 716, Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and the Training Program, Elaine D. Pulakos and Walter C. Borman (Eds.), and the supplementary ARI Research Note 87-22, which contains the report appendixes.

2. A briefing was prepared to acquaint participant raters with the main objectives of Project A and to explain where the performance ratings fit into the project.

3. An orientation to the behavior-based rating scales was developed. The principle of matching observed ratee performance with performance described in the scales' behavioral anchors was carefully explained and illustrated with several hypothetical examples.

4. A short program was aimed at avoiding three common rating errors: halo, stereotyping, and paying too much attention to one or two events relevant to the ratee's performance (heretofore labeled as the "one incident of performance error").

5. For practice, peer raters were asked to make self-ratings using the Army-wide behavior-based rating scales, to ensure that they became acquainted with the rating process before they began their evaluations of other soldiers.

The orientation and training program described above was developed for the Batch A field tests. The intent was to start with this program, evaluate its effectiveness in the Batch A tests, revise for Batch B based on Batch A experience, continue the tryout in Batch B, and finally revise as required for the large-scale Concurrent Validation effort.

Lessons Learned During the Batch A Field Tests

The Batch A rater training and orientation program seemed quite successful in that: (a) it appeared to flow well and be acceptable to both supervisor and peer raters (i.e., the soldiers were generally attentive to the program and appeared to complete their ratings responsibly); (b) the interrater reliabilities were very reasonable, especially in light of the fact that the peer raters were inexperienced at evaluating performance; (c) the rating distributions were very reasonable, with no drastic skewing; (d) the training effects did not seem to be trainer-bound in that at least seven trainers administered the program at one time or another during the Batch A field tests; and (e) the relationships between the ratings and other criterion variables showed some predictable patterns (e.g., correlations between MOS task scale ratings and hands-on test scores averaged about .25).

Although the program seemed effective, Batch A field test experience suggested some additions and modifications. First, some supervisor and peer raters did appear to be evaluating all of their ratees at approximately the same level of effectiveness on many of the dimensions (i.e., halo error). To counteract this tendency, raters were subsequently encouraged not only to tell us about each individual's strengths and weaknesses (thereby avoiding halo error) but also to indicate differences between soldiers who perform well in a particular rating category and those who perform less well in the category.

Second, although error reduction training is very important in yielding high-quality evaluations, recent research (McIntyre et al., 1984; Pulakos,

1984) has suggested that error training alone may be insufficient for increasing rating accuracy, the crucial criterion for evaluating performance rating quality (Ilgen & Feldman, 1983; Landy & Farr, 1980). Therefore, following the Batch A field tests we incorporated a more comprehensive accuracy training component into the program. We stressed that, although we did not want raters to make rating errors, the most important element was to rate each of their subordinates or co-workers accurately. Thus, if raters felt that their ratees actually performed at the same effectiveness level in a given performance category, or that a particular soldier performed at approximately the same level across several categories, then they were encouraged to rate those individuals in this way. However, it was emphasized that when real differences exist, the ratings should reflect these differences.

Finally, a question was raised as to the usefulness of self-ratings as an aid in familiarizing raters with the rating scales. Since less time was to be available for ratings during Concurrent Validation, it was important to consider which instruments and/or aspects of training might be eliminated. Toward this end, an empirical evaluation of the self-rating effects would be useful. No research to date has investigated the effects of self-ratings on subsequent ratings of others, so evaluating this aspect of the training had general as well as specific project implications. An experiment, described below, was designed to investigate the self-rating effect and was conducted as part of the Batch B field tests.

Batch B Rater Training Experiment

Two training treatments were evaluated for peers as raters: (a) rater orientation and error reduction training, including a brief refresher of the error training points prior to administering each new scale; and (b) this same program plus a self-rating warm-up for each scale.

Parallel training treatments were also evaluated for supervisors. However, because the rating scales were specifically developed for evaluating first-term soldier performance, having the supervisors use these scales to perform a self-rating task would have been inappropriate. Consequently, practice for the supervisors entailed rating a description of one hypothetical soldier prior to evaluating their subordinates. The two supervisor training treatments were thus: (a) rater orientation and error reduction training, including brief refresher training before each new instrument; and (b) this same program plus practice rating of one hypothetical soldier on the Army-wide BARS.

The training treatments for each peer and supervisor rater group were evaluated in terms of their effects on rating accuracy and three rating errors (halo, leniency/severity, and restriction of range).

Subjects

A total of 817 peer raters and 660 supervisor raters participated in the Batch B field tests. Each soldier represented one of the following five MOS: 11B (Infantryman), 19E (Armor Crewman), 31C (Single Channel Radio Operator),

63B (Light Wheel Vehicle Mechanic), and 91A (Medical Specialist). Data were collected from four CONUS locations and USAREUR.

Rating Instruments

Four of the rating instruments used during the Batch B field tests were relevant for the present study:

- Army-wide behavioral rating scales
- Army-wide common task scales
- MOS-specific behavioral rating scales
- MOS-specific task scales

Experimental Treatments

Training Methods. The following three training methods were used as independent variables:

- Rater Orientation and Error Training Only. Peer and supervisor raters assigned to this experimental condition received training that can be characterized as a combination psychometric error and frame-of-reference program (Bernardin & Pence, 1981; Pulakos, 1984). Briefly, one component of training involved carefully explaining the logic of the behavior-based and task rating scales, as well as urging raters to study and properly use the instruments to arrive at their evaluations. The second major component involved descriptions of halo, stereotyping, one incident of performance, and same-level-of-effectiveness errors in lay terms and provided guidance on how to avoid these errors.

- Rater Orientation and Error Training Plus Practice: Peer Raters. This experimental condition consisted of the training outlined above plus practice using the rating scales in the form of self-appraisals. Specifically, prior to rating their co-workers on each of the four sets of scales, peer raters were asked to evaluate themselves using these instruments.

- Rater Orientation and Error Training Plus Practice: Supervisor Raters. Supervisors assigned to this condition also received the rater orientation and error reduction training discussed above. However, their practice entailed evaluating one hypothetical ratee on the Army-wide behavioral performance dimensions. A vignette describing performance of a first-term soldier was developed for this purpose, using behavioral examples from the pool of items retranslated during Army-wide behavior scale development.

Dependent Variables. We were able to create "ratees" with known performance scores by developing vignettes about first-term soldiers performing tneir jobs, using in the vignettes previously scaled behavioral examples (just as we did for the supervisor practice rating condition). The true or target performance level for a dimension was simply the mean retranslation

effectiveness level for the example included in the vignette for that dimension. Four vignettes were written describing performance in the following Army-wide areas: (a) Effort, (b) Maintain Assigned Equipment, (c) Maintain Living and Work Areas, (d) Physical Fitness, (e) Self-Development, and (f) Self-Control.

By using expert judges' estimates of the true intercorrelation between the six dimensions, along with dimension means of 4.0 and standard deviations of 1.5, a true score matrix (see Table III.60) containing scores for hypothetical ratees on each dimension was generated; this matrix possessed the "correct" covariance structure. Using behavioral examples obtained from the retranslation phase of the Army-wide behavior scaling process, vignettes were then written describing four ratees performing at the effectiveness levels shown in Table III.60. Each incident had been allocated reliably into a single dimension and assigned a narrow range of effectiveness levels in the retranslation process.

Table III.60

True Score[a] Matrix for Vignette Ratees on Six Army-Wide Dimensions

	Ratees			
Dimensions	1	2	3	4
1. Effort	5	2	6	4
2. Maintain Assigned Equipment	5	3	5	2
3. Maintain Living & Work Areas	3	1	5	3
4. Physical Fitness	4	3	5	6
5. Self-Development	7	2	6	4
6. Self-Control	6	1	2	5

[a]Because the rating task required evaluators to select a whole number from 1 to 7 describing each soldier's effectiveness on a dimension, the generated true scores were rounded to the nearest whole number.

After evaluating their co-workers or subordinates on the Army-wide behavioral rating scales, both peers and supervisors read and then rated each of the four vignettes. The materials used to collect these data, including instructions, the actual vignettes, and special rating scales containing only the six dimensions relevant to the vignettes, are presented in ARI Research Note 87-22.

Using the peer and supervisor ratings of the soldiers evaluated (not the vignettes), the following four rating indexes were computed: interrater agreement, halo, leniency/severity, and restriction of range. The vignette ratings were used to assess training effects on accuracy. Each dependent measure was examined separately for peer and supervisory raters.

Procedure

In the field tests first-term soldiers reported to their rating sessions in groups of approximately 15. At each location, one supervisor rating session was conducted for each MOS. Thus, at each post it was necessary to assign supervisor raters within an MOS to one of the two training treatments and then counterbalance the treatment for each MOS across the posts. So, for example, at Fort Stewart, MOS 19Es and 91As received error training only, while MOS 11Bs, 31Cs, and 63Bs received error training plus practice. Conversely, at Fort Lewis, MOS 11Bs, 31Cs, and 63Bs received error training only, while MOS 19Es and 91As received error training plus practice. A similar counterbalancing scheme was used for the remaining three locations. This assignment process resulted in approximately equal numbers of soldiers from each MOS receiving each type of training across the five data collection sites.

Results

Interrater Agreement

Within each training treatment, an intraclass correlation was computed for each dimension of the four instruments on which actual soldier performance was rated. Within each instrument, these correlations were then averaged across the dimensions, resulting in four indexes of rater agreement for each training condition.

For the peer and supervisor rater groups, Table III.61 contains the average intraclass correlations for each of the four rating instruments within training condition. To determine whether there were significant differences between the treatments, chi-square tests were performed. However, because the degree to which these measures are actually correlated was unknown, testing for differences between the correlations proceeded as follows. First, a minimum chi-square, assuming perfect dependency among the measures, was computed. A significant minimum chi-square indicates a difference between the two intraclass correlations. If the minimum chi-square was nonsignificant, a maximum chi-square, assuming perfect independence among the measures, was then computed. A nonsignificant maximum chi-square indicates no difference between the two correlations. The final possibility was that we would obtain a nonsignificant minimum chi-square but a significant maximum chi-square. Such a result would have indicated the possibility of a difference between the two correlations, but no definitive conclusions could be drawn.

Table III.61

Interrater Reliabilities by Training Condition[a] Across All MOS

	Army-Wide Scales		Army-Wide Common Tasks Scales		MOS Scales		MOS Tasks	
Rater Group	EO	E+P	EO	E+P	EO	E+P	EO	E+P
Peers	.26	.31	.18	.17	.16	.18	.11	.11
Supervisors	.32	.37	.21	.26	.30	.38[b]	.21	.34[b]

[a] EO = error training only; E+P = error training plus practice. These are one-rater reliabilities calculated on the unadjusted ratings.

[b] Minimum χ^2 was nonsignificant, but maximum χ^2 was significant.

As shown in Table III.61, for the peers there were no differences between the two training treatments on any of the rating scale types. For the supervisors, interrater agreement was consistent across the training treatments for the Army-wide scales, but practice may have increased rater agreement on the MOS-specific scales.

Given that the supervisors' practice was restricted to only the Army-wide rating dimensions, the finding that practice seemed to facilitate agreement on the MOS scales but not the Army-wide scales seemed counter-intuitive. Hence, the data were inspected further to evaluate the consistency of this effect across MOS. These analyses revealed a significant difference between the two training treatments on the MOS scales only for MOS 91A; there were no differences in interrater agreement as a result of training for any of the other MOS.

Training Effects on Rater Errors

For both the peer and supervisor raters, there were no differences between the two training treatments in terms of halo, leniency/severity, or restriction-of-range errors on either the Army-wide or the MOS-specific rating scales.

Training Effects on Accuracy

A 2 X 2 (Training X Rater Group) fixed-factor analysis of variance was conducted to evaluate training effects on accuracy. (Accuracy was operationalized as the average squared difference between the true scores and each rater's observed ratings, with lower values indicating greater accuracy.) The ANOVA results revealed no significant differences as a function of training or rater group.

Summary and Conclusions

In this experiment to assess whether a practice component of training improved performance rating quality beyond what was obtained by error training alone, results were identical for the peer and supervisor raters. The practice component yielded no significant improvement in ratings in terms of interrater agreement or any of the rating errors assessed here (i.e., halo, leniency/severity, and restriction of range). Further, practice did not facilitate accuracy on a vignette rating task.

It was therefore concluded that the rater orientation and training program to be used for Concurrent Validation would not include the additional peer and supervisor practice components that had been tried out in the experimental study.

Section 14

FIELD TEST RESULTS: COMBAT PERFORMANCE PREDICTION SCALE[1]

Forms A and B of the Combat Performance Prediction Scale were administered at only one post during the Batch B field testing. The scale was administered to peer and supervisor raters during the rating sessions, along with the Army-wide and MOS-specific rating scales. Thus, the rater training described in Section 13 preceded administration of the combat prediction ratings as well.

Statistics From Field Test

Table III.62 presents the means and standard deviations of the combat effectiveness ratings by rating source, scale dimension, and combat vs. non-combat MOS. As can be seen, no meaningful differences were found between supervisor and peer raters, or combat and non-combat MOS, or among the six scale dimensions. All of the means are slightly above the scale midpoint of 7.5.

Table III.63 presents the one-rater intraclass correlations for the total of the 76 items and for each of the category scores. A reliability of .21 was obtained for the total when ratings were pooled across raters and MOS. This reliability is based on all 76 items, some of which may have poor psychometric properties that could attenuate the reliability coefficient. In sum, however, the interrater agreement is disappointing, suggesting strongly that more item analysis is warranted.

Coefficient alphas for the total 76-item scale as well as for each category are presented in Table III.64. The value ranged from .76 to .88 for the dimensions and was .94 for the total.

Revision of Scale for Concurrent Validation

The item statistics used in selecting 40 items for the final scale to be used in Concurrent Validation are presented in Table III.65. Items were selected on the basis of content domain (dimension) coverage and psychometric properties. The 40-item dimension coverage was approximately proportional to the 76-item dimension coverage for the field test. Psychometric properties considered included rescaled t-value, reliability, item-dimension correlation, item-total correlation, and across MOS and rater group means and standard deviations.

Responses to the questions concerning rating confidence and item applicability were also considered. Total scale confidence ratings were slightly above midpoint on a 7-point scale (mean value of 4.25 and 4.20 for

---

[1]Development of this scale was described in Section 6, Part III. Section 14 is based primarily on an unpublished manuscript, "Development of Combat Performance Prediction Scale," by Barry J. Riegelhaupt and Robert Sadacca.

Table III.62

Means and Standard Deviations for Rater-Ratee Pairs on the Combat Performance Prediction Scale[a]

Dimension	Items	Combat (11B, 19E)		Noncombat (31C, 63B, 91A)	
		Peers (N=51)	Supervisors (N=36)	Peers (N=85)	Supervisors (N=77)
Cohesion/Commitment	15	7.73 (2.19)	8.54 (2.08)	8.70 (2.43)	8.52 (2.37)
Self-Discipline/ Responsibility	15	8.97 (2.42)	9.45 (2.40)	9.50 (2.40)	9.55 (2.34)
Mission Orientation	14	9.22 (2.12)	9.87 (1.96)	9.61 (2.15)	9.51 (2.39)
Technical/Tactical Knowledge	12	9.12 (2.04)	8.08 (.216)	9.41 (2.21)	8.78 (2.16)
Initiative	9	8.36 (2.54)	8.37 (2.43)	8.77 (2.70)	8.14 (2.56)
Other	11	9.16 (2.20)	8.74 (2.42)	9.39 (2.44)	8.94 (2.39)
Total	76	8.76 (1.87)	8.96 (1.88)	9.23 (2.04)	8.91 (2.02)

[a] Scale ranged from 1 = Very Unlikely to 15 = Very Likely. Standard deviations are shown in parentheses.

Table III.63

Intraclass Correlations for Estimating Reliabilities for the Combat Performance Prediction Scale

Dimension	Pooled Across MOS[a]		Pooled Across Raters[b]		Pooled Across Raters & MOS
	Peers	Supervisors	Peers	Supervisors	
Cohesion/Commitment	.25	.26	.08	.23	.21
Self-Discipline/ Responsibility	.22	.28	.20	.17	.19
Mission Orientation	.15	.12	.03	.16	.11
Technical/Tactical Knowledge	.19	.19	.11	.17	.15
Initiative	.28	.05	.14	.19	.17
Other	.19	.18	.11	.19	.16
Total	.27	.20	.15	.23	.21

a MOS: 11B, 19E, 31C, 63B, 91A

b Rater/ratee pairs:

	Peers	Supervisors
11B	107	58
19E	93	45
31C	66	47
63B	75	50
91A	124	58
	465	258

Table III.64

Coefficient Alpha for the Combat Performance Prediction Scale

Dimension	Alpha
Cohesion/Commitment	.78
Self-Discipline/Responsibility	.81
Mission Orientation	.79
Technical/Tactical Knowledge	.76
Initiative	.88
Other	.77
Total	.94

Table III.65

Item Statistics Used in Selecting Combat Prediction Scale Items

Item[a]	Category	Reverse Scored Item[c]	Rescaled t-Value	Across MOS & Rater Reliability[d]	Item-Category Correlation[d]	Item-Total Correlation[d]	Across MOS & Rater Groups Mean	Across MOS & Rater Groups SD	Nonapplicability Frequency[e]
1	H		8.32	24	68	64	8.77	3.13	80
2	F	x	-3.63	02	48	35	9.40	3.25	214
3	C	x	-4.52	18	21	22	9.43	3.27	80
4	F	x	-4.25	17	26	26	10.31	3.12	104
5*	A		8.09	30	44	47	8.53	3.33	54
6	E		5.30	14	31	36	10.60	2.93	164
7	C	x	-7.72	06	36	27	12.03	2.93	348
8*	H		4.92	20	71	62	7.85	3.44	104
9*	C	x	-7.08	11	35	37	10.91	3.35	42
10	F	x	-5.73	11	61	52	9.91	3.35	388
11*	F		8.79	17	62	57	8.13	2.97	86
12	O	x	-6.33	07	33	38	9.70	3.29	56
13*	O	x	-7.67	20	59	53	9.27	3.60	174
14	E	x	-4.97	23	44	39	10.13	3.30	126
15*	C	x	-4.47	18	79	64	9.69	3.40	88
16*	E	x	-4.09	14	58	50	10.33	3.02	184
17	F	x	-4.78	11	33	41	11.93	2.82	128
18*	C		7.70	14	61	57	7.73	3.13	130
19	A		4.79	04	54	44	8.38	3.10	84
20	E	x	-3.30	06	32	19	10.36	3.48	260
21	A	x	-6.29	28	32	37	9.71	3.56	112
22	A		3.51	12	53	42	7.53	3.64	142
23*	H		4.53	11	65	60	8.85	2.88	184
24*	E		7.87	15	59	59	8.13	3.26	68
25*	F		5.38	10	36	43	9.22	2.83	92
26	A		6.18	16	59	36	8.20	3.35	184
27*	E		6.41	20	54	52	8.40	3.42	30
28	H		6.71	14	67	62	7.76	3.02	414
29*	C		7.67	20	40	48	8.55	2.91	54
30*	E		5.21	16	62	59	8.12	2.89	110
31	[illegible]	x	-4.79	11	28	22	10.98	3.04	78
32*	F		6.18	22	66	64	7.59	2.97	56
33*	C		6.70	30	58	62	9.36	3.09	32
34*	C	x	-4.53	12	61	51	10.95	3.39	352
35	H		6.11	11	69	59	8.33	2.99	216
36*	O		6.03	09	60	48	9.64	3.02	216
37*	A		6.02	33	44	41	8.91	2.92	70
38*	A		7.27	19	68	63	9.19	3.34	82
39	A		6.19	18	62	30	7.64	3.86	142
40	F	x	-5.16	17	30	38	10.67	2.81	202
41*	F		4.60	14	62	63	8.65	2.88	212
42*	A		6.10	20	51	46	7.47	3.27	60
43	E	x	-4.01	10	48	42	10.03	3.12	140

(Continued)

Table III.65 (Continued)

Item Statistics Used in Selecting Combat Prediction Scale Items

Item[a]	Category	Reverse Scored Item[c]	Rescaled t-Value	Across MOS & Rater Reliability[d]	Item-Category Correlation[d]	Item-Total Correlation[d]	Across MOS & Rater Groups Mean	Across MOS & Rater Groups SD	Nonapplicability Frequency[e]
44*	O		4.61	21	70	68	9.67	2.86	86
45*	H		5.59	26	75	68	9.15	2.82	76
46*	A		5.30	20	64	43	9.05	2.91	122
47*	E		5.74	21	67	66	8.80	2.75	88
48	F	x	-5.20	09	40	34	10.89	2.93	284
49	O		8.00	20	67	60	8.92	2.86	488
50	O		7.09	22	50	45	8.94	2.71	266
51*	C	x	-4.61	20	70	60	8.61	3.53	112
52*	O		6.33	11	70	62	8.13	2.90	246
53	E	x	-4.64	06	40	31	11.93	2.87	250
54	A	x	-4.87	10	53	50	8.71	3.45	70
55*	C		6.72	14	64	56	7.29	3.36	88
56	E		7.73	20	62	56	8.04	3.58	382
57	F	x	-6.08	02	33	35	11.53	3.26	100
58	C		5.44	11	25	28	9.81	3.07	34
59	H		6.42	31	56	55	8.18	3.35	330
60	A		6.04	13	27	11	7.84	3.64	100
61*	O	x	-8.08	18	49	52	10.05	3.25	56
62*	O		6.48	22	54	56	8.33	2.90	32
63	C	x	-6.93	05	19	19	12.53	2.99	180
64*	E		3.50	12	39	40	5.27	3.40	104
65*	O		4.83	27	50	54	10.74	2.73	20
66*	O	x	9.90	16	46	52	10.04	2.76	14
67*	E		6.32	27	52	53	7.67	2.89	36
68	C	x	-4.80	06	30	24	10.17	3.38	72
69	C		10.85	12	28	29	10.94	3.15	72
70	C	x	-4.82	10	42	41	11.14	3.38	108
71*	H		5.63	35	52	51	9.92	2.92	30
72*	F	x	-4.23	15	50	46	9.04	3.16	112
73*	H		6.10	32	63	63	8.34	2.96	70
74*	A		10.50	18	38	50	8.42	3.04	50
75	A	x	-5.99	18	03	04	10.60	3.42	158
76*	A		5.99	17	38	52	9.19	2.87	16

[a]Items with an asterisk were selected for final Combat Performance Prediction Scale.

[b]Category A - Cohesion/Commitment
C - Self-Discipline/Responsibility
E - Mission Orientation
F - Technical Tactial Knowledge
H - Initiative
O - Other

[c]Items with an x were reverse scored.

[d]Decimal points have been omitted.

[e]Frequencies represent number of rater-ratee pairs across peer and supervisor raters, for which the item was judged nonapplicable. There existed 5194 rater-ratee pairs.

peer and supervisor ratings, respectively). In response to the question about the number of applicable items, on the average peer and supervisor raters felt that about 43 items applied to the soldiers they rated. The frequency with which particular items were judged nonapplicable is presented in Table III.65. In selecting items for Concurrent Validation, preference was given to items that raters judged to be applicable.

The corrected intraclass correlations for the final 40-item scale are shown in Table III.66. Vast improvement (i.e, .21 to .56) resulted when the 40 best items from among the 76 were selected. Total scale coefficient alpha remained at .94.

Table III.66

Corrected Intraclass Correlations for Estimating Reliabilities of Best 40 Items on Combat Performance Prediction Scale

Rater Group	Form A	Form B	Across Forms	Across Forms and Raters
Peers	.55	.66	.56	--
Supervisors	.78	.63	.68	--
				.56

This 40-item scale was judged to have sufficiently good psychometric properties to justify its use for all MOS in the Concurrent Validation data collection. Sample pages from the instrument are shown as Figure III.15. The scale (e.g., development of factors) will be further refined on the basis of the Concurrent Validation data.

1. This soldier volunteered to lead a team to an accident scene where immediate first aid was required before an order was given.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely	
Line up the names	1	O	O	O	O	O	O	O	O	O	O	O	O	O	O
of the soldiers	2	O	O	O	O	O	O	O	O	O	O	O	O	O	O
you are rating	3	O	O	O	O	O	O	O	O	O	O	O	O	O	O
with the rows	4	O	O	O	O	O	O	O	O	O	O	O	O	O	O
to the right.	5	O	O	O	O	O	O	O	O	O	O	O	O	O	O

2. Near the end of a movement when soldiers were ordered to prepare fighting positions, this soldier prepared his position quickly and then assisted other squad members.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely	
Line up the names	1	O	O	O	O	O	O	O	O	O	O	O	O	O	O
of the soldiers	2	O	O	O	O	O	O	O	O	O	O	O	O	O	O
you are rating	3	O	O	O	O	O	O	O	O	O	O	O	O	O	O
with the rows	4	O	O	O	O	O	O	O	O	O	O	O	O	O	O
to the right.	5	O	O	O	O	O	O	O	O	O	O	O	O	O	O

3. This soldier prepared defensive positions without being told to do so.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely	
Line up the names	1	O	O	O	O	O	O	O	O	O	O	O	O	O	O
of the soldiers	2	O	O	O	O	O	O	O	O	O	O	O	O	O	O
you are rating	3	O	O	O	O	O	O	O	O	O	O	O	O	O	O
with the rows	4	O	O	O	O	O	O	O	O	O	O	O	O	O	O
to the right.	5	O	O	O	O	O	O	O	O	O	O	O	O	O	O

4. The battery/company commander instructed everyone to be packed and ready for movement at 0800 hours. This soldier arrived late and missed the movement.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely	
Line up the names	1	O	O	O	O	O	O	O	O	O	O	O	O	O	O
of the soldiers	2	O	O	O	O	O	O	O	O	O	O	O	O	O	O
you are rating	3	O	O	O	O	O	O	O	O	O	O	O	O	O	O
with the rows	4	O	O	O	O	O	O	O	O	O	O	O	O	O	O
to the right.	5	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Figure III.15: Sample items from Combat Performance Prediction Rating Scale (Page 1 of 2)

5. When casualties were to be evacuated from a location identified only by map coordinates, this soldier was able to locate the site by accurate navigation in terrain with few prominent features.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely		
Line up the names	1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
of the soldiers	2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
you are rating	3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
with the rows	4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
to the right.	5	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

6. This soldier talked with other soldiers who were having difficulties coping with the combat conditions.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely		
Line up the names	1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
of the soldiers	2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
you are rating	3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
with the rows	4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
to the right.	5	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

7. Although the unit was in the field for an extended period of time, this soldier constantly cleaned his weapon and carried additional cleaning supplies to be certain they were always available.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely		
Line up the names	1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
of the soldiers	2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
you are rating	3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
with the rows	4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
to the right.	5	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

8. During movement of a convoy from support area to forward base camps, this soldier failed to wear his flak jacket and helmet complaining that they were too heavy and hot.

		Very Unlikely			Fairly Unlikely			About 50-50 Chance			Fairly Likely			Very Likely		
Line up the names	1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
of the soldiers	2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
you are rating	3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
with the rows	4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
to the right.	5	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Figure III.15. Sample items from Combat Performance Prediction Rating Scale (Page 2 of 2)

Section 15

FIELD TEST RESULTS: ARCHIVAL/ADMINISTRATIVE INDICATORS[1]

The Personnel File Information Form (a self-report of 201 File information), which had been developed for use in Batch A field testing, was administered to every soldier at every data collection location. For each soldier tested, project staff also requested his or her 201 File. Using the same form that soldiers completed during the testing sessions, project staff extracted administrative measures information from each soldier's personnel record, thus making possible a comparison of the two approaches to collecting personal information. A revised self-report form was tried out during the Batch B field test.

Results From Batch A

Only soldiers for whom both self-report and 201 File information were available were retained for these analyses. For a small number of cases, self-report data were missing. More often, file data were missing because a soldier did not grant us permission to view his/her 201 File, or for a variety of other reasons. Thus, although data were collected on 548 soldiers during Batch A field tests, only 505 cases were available for administrative measures analyses.

Self-Report vs. File Data

Tables III.67-III.72 show comparisons of information obtained from self-reports and 201 File extraction. Sample sizes below 505 reflect missing data, from one or both sources.

For the Number of Awards variable, as can be seen in Table III.69, there was perfect correspondence between the two sources. For the other measures, which showed varying levels of agreement (i.e., off-diagonal cases), a greater percentage of cases consistently fell below the diagonal. That is, soldiers were reporting more occurrences of administrative measures being received than were found in their 201 Files.

This situation was not surprising in light of the knowledge acquired in our earlier exploration of 201 Files. According to regulations, not all letters, certificates, Articles 15, and so forth, are placed in 201 Files, and some documents are removed after a certain period of time. Also, while 201 Files are the most timely official source of information, they are certainly not updated daily. Thus, discrepancies in the reported direction were not unexpected.

---

[1]Development of these indicators was described in Section 7, Part III. Section 15 is based primarily on an unpublished manuscript, "Army-Wide Administrative Measures," by Barry J. Riegelhaupt.

Table III.67

Comparison of Reenlistment Eligibility Information Obtained From Self-Report and 201 Files: Batch A

	201 File		
Self-Report	Eligible	Ineligible	Total
Eligible	293	45	338
Ineligible	35	21	56
Total	328	66	394

Table III.68

Comparison of Promotion Rate[a] Information Obtained From Self-Report and 201 Files: Batch A

	201 File										
Self-Report	0	.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	Total
0	3	0	1	2	3	1	0	0	0	0	10
.5	0	9	0	4	5	1	0	0	0	0	19
1.0	0	3	34	6	10	0	1	0	0	0	54
1.5	1	2	7	63	26	2	2	1	1	0	105
2.0	0	0	0	7	133	14	4	2	1	0	161
2.5	0	0	1	2	15	48	2	0	0	1	69
3.0	0	0	1	1	2	4	20	0	0	0	28
3.5	0	0	0	0	0	0	3	1	0	0	4
4.0	0	0	0	1	0	1	0	0	0	0	2
7.5	0	0	0	0	1	0	0	0	0	0	1
8.0	0	0	0	1	1	0	0	0	0	0	2
Total	4	14	44	87	196	71	32	4	2	1	455

[a]Grades advanced/year.

Table III.69

Comparison of Awards Information Obtained From
Self-Report and 201 Files: Batch A

	201 File				
Self-Report	0	1	2	3	Total
0	302	0	0	0	302
1	0	158	0	0	158
2	0	0	37	0	37
3	0	0	0	8	8
Total	302	158	37	8	505

Table III.70

Comparison of Letters/Certificates Information Obtained From
Self-Report and 201 Files: Batch A

	201 File							
Self-Report	0	1	2	3	4	5	6	Total
0	178	9	2	0	0	1	0	190
1	80	20	3	1	0	0	0	104
2	60	21	6	0	1	0	0	88
3	38	11	6	3	0	0	0	58
4	24	8	5	4	1	0	1	43
5	7	4	1	0	1	0	0	13
6	5	1	0	1	1	0	0	8
7	0	0	1	0	0	0	0	1
Total	392	74	24	9	4	1	1	505

Table III.71

Comparison of Articles 15/FLAG Information Obtained From Self-Report and 201 Files: Batch A

	201 File				
Self-Report	0	1	2	3	Total
0	320	10	2	0	332
1	73	6	4	0	83
2	38	13	2	1	54
3	18	8	1	0	27
4	2	1	1	1	5
5	1	1	0	0	2
6	1	0	0	0	1
7	1	0	0	0	1
Total	454	39	10	2	505

Table III.72

Comparison of Military Training Information Obtained From Self-Report and 201 Files: Batch A

	201 File		
Self-Report	No	Yes	Total
No	281	15	296
Yes	188	21	209
Total	469	36	505

The intent of this comparison was to be able to address the accuracy of self-report. However, the information contained in 201 Files does not necessarily represent "truth," so determining the accuracy of self-report became an even greater challenge. If soldiers had reported more positive documents, such as letters and certificates, and fewer negative documents, such as Articles 15, when compared with the file data, then the self-report data would surely be suspect. However, soldiers reported receiving more negative as well as more positive documents. An exception was Reenlistment Eligibility, which unlike the other administrative variables are constantly subject to change. For example, if a soldier is not eligible because of being overweight, but subsequently loses weight, he/she once again becomes eligible. In light of the time lag associated with updates to Reenlistment Eligibility status in 201 Files, a number of off-diagonal cases would be expected, and it would be impossible to predict whether these deviations would lie above or below the diagonal. In view of the above results, it seems likely that soldiers were honestly responding to the questions.

Correlations With Rating Variables

Tables III.73 and III.74 present the correlations between the six administrative measures and Army-wide supervisor and peer ratings, respectively. Correlations are shown for both the self-report and the 201 File data. As can be seen, relationships between Army-wide ratings and the administrative measures obtained from the self-report approach were generally higher than those obtained from 201 Files.

Conclusions

While the self-report method in the Batch A field test yielded enhanced variance and stronger relationships with other measures, and was easier and less expensive to use, the selection of self-report over file extraction was still premature. The Batch B field test was used to provide additional information.

A number of revisions were made in the self-report at this point. The Military Training Courses variable was dropped from consideration because it had little variance and showed very low relationships with other measures. Further, recall that in the earlier 201 File-EMF comparison, almost perfect agreement between the two sources was found for the Promotion Rate and Reenlistment Eligibility variables. Since monthly updates of the EMF subsequently became available, there no longer was a need to collect this information from the field. Therefore, the Reenlistment Eligibility question and three questions used to compute Promotion Rate were dropped from the Personnel File Information Form.

Procedural Changes for Batch B

The goal in Batch B field testing--to improve the correspondence between information extracted from 201 Files and that obtained from soldiers' self-reports--was to be accomplished by shortening the form, and by having session administrators "walk" the soldiers through the questions, explaining which things should or should not be counted in responding to certain items.

Table III.73

Correlations Between Army-Wide Supervisor Ratings and Administrative Measures: Batch A

Army-Wide Supervisor Ratings	Reenlistment Eligibility		Promotion Rate		Number of Awards		Number of Lett / Cert		Number of Art 15/Flag		Military Training Courses	
	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File
Technical Knowledge/Skill	05	06	08	06	20	20	14	09	-17	-12	08	04
Initiative Effort	01	05	02	04	14	14	08	07	-12	-10	08	08
Following Regs/Orders	14	10	05	05	09	09	09	05	-25	-18	08	03
Integrity	11	08	03	12	10	10	11	10	-21	-12	06	07
Leading and Supporting	13	18	05	00	12	12	17	14	-16	-13	09	04
Maintaining Equipment	02	01	11	06	11	11	08	08	-16	-13	05	06
Maint. Living/Work Area	10	08	07	02	14	14	07	04	-25	-14	06	10
Military Appearance	20	15	06	01	07	07	10	14	-27	-15	02	05
Physical Fitness	16	05	06	08	08	08	08	06	-22	-14	00	-03
Self-Development	17	08	03	06	13	13	17	08	-21	-14	08	14
Self-Control	23	16	07	15	10	10	14	14	-30	-28	08	14
Overall Effectiveness	15	14	03	07	10	10	21	13	-24	-14	08	09
NCO Potential	16	18	04	06	15	15	16	11	-19	-16	01	07

[a] $N = 505$. Decimal points have been omitted.

Table III.74

Correlations Between Army-Wide Peer Ratings and Administrative Measures: Batch A

Army-Wide Peer Ratings	Reenlistment Eligibility		Promotion Rate		Number of Awards		Number of Lett / Cert		Number of Art 15/Flag		Military Training Courses	
	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File	Self-Report	201 File
Technical Knowledge/Skill	16	08	01	03	06	06	19	17	-19	-07	03	09
Initiative Effort	12	04	05	14	09	09	16	13	-31	-16	03	11
Following Regs/Orders	17	07	01	10	08	08	16	11	-35	-17	03	14
Integrity	08	06	05	07	02	02	14	10	-26	-14	10	07
Leading and Supporting	15	15	03	05	11	11	21	19	-24	-10	04	03
Maintaining Equipment	11	-02	04	12	02	02	07	03	-23	-15	04	06
Maint. Living/Work Area	14	10	-02	00	14	14	07	14	-28	-14	-04	07
Military Appearance	21	05	03	07	06	06	14	12	-27	-15	02	04
Physical Fitness	15	07	00	03	04	04	03	06	-19	-11	03	-03
Self-Development	17	07	00	07	03	03	15	15	-28	-15	10	13
Self-Control	14	10	01	11	05	05	12	11	-30	-14	06	13
Overall Effectiveness	18	09	04	11	09	09	16	12	-31	-10	04	10
NCO Potential	24	14	02	05	10	10	16	17	-34	-13	02	15

[a]N = 505. Decimal points have been omitted.

For example, upon completing Head Start a soldier receives a certificate; this is not a certificate of appreciation, commendation, or achievement and thus should not be counted when responding to "How many certificates have you received?" These oral instructions were expected to reduce some of the discrepancies found in Batch A results.

To further investigate why self-report differed from file information, staff personnel conducted an outlier analysis by talking with individual soldiers, trying to determine the extent to which they were counting the items that we intended to be counted. To the extent that the soldier was interpreting the question as we intended, we then asked for possible explanations as to why a self-reported item was not found in the 201 File.

Results From Batch B

Tables III.75-III.78 present Batch B comparisons of information obtained from self-report and file extraction. As before, a greater percentage of non-matches were found below rather than above the diagonal. That is, once again soldiers were reporting that they had received more letters, Articles 15, and so forth, than were found in their 201 Files.

This time information as to the possible causes of the discrepancies was available from the interviews that had been conducted as part of the outlier comparison. The most frequently expressed explanations are presented in Figure III.16. Some of the reasons confirmed earlier suspicions, such as "Counted training certificates," "Counted certificate/letter that accompanied award," and "Recently received, paperwork not completed." Other reasons were unexpected, such as "Counted Levy alert" as a FLAG action; a Levy alert is a notification of an impending transfer.

These interviews provided much-needed information. The lesson learned was a simple one: For the Concurrent Validation data collection the self-report questions needed to be more detailed, and even more clearly specified.

Revisions for Concurrent Validation

After the two field tests of the Personnel File Information Form, three conclusions were drawn. First, self-report yields the most timely data. Second, self-report yields more complete data. Finally, as mentioned above, the questions needed to more detailed.

Acting on this knowledge, we developed a short, simple, but more detailed records report form. The resulting Personnel File Information Form (Form 7) is shown in Figure III.17.

Form 7 is being used as a self-report instrument during Concurrent Validation data collection, and the information obtained will be combined with the Promotion Rate and Reenlistment Eligibility variables obtained from the EMF.

Table III.75

Comparison of Letters/Certificates Information Obtained From Self-Report and 201 Files: Batch B

	201 File				
Self-Report	0	1	2	3	Total
0	38	6	0	0	44
1	14	4	0	0	18
2	13	1	1	1	16
3	7	0	1	0	8
4	4	1	0	1	6
5	3	0	1	2	6
6	5	1	0	1	7
7	0	0	1	0	1
8	4	1	0	0	5
9	1	0	1	0	2
10	1	1	0	0	2
11	1	0	0	0	1
12	1	0	1	0	2
13	0	0	0	0	0
14	0	1	0	0	1
Total	92	16	6	5	119

Table III.76

Comparison of Awards Information Obtained From Self-Report and 201 Files: Batch B

	201 File				
Self-Report	0	1	2	3	Total
0	47	3	0	0	50
1	17	21	4	0	42
2	3	8	9	1	21
3	2	1	0	1	4
Total	69	33	13	2	117

Table III.77

Comparison of Articles 15/FLAG Information Obtained From Self-Report and 201 Files: Batch B

	201 File						
Self-Report	0	1	2	3	4	5	Total
0	93	1	0	0	0	0	94
1	13	1	0	0	0	0	14
2	4	2	1	0	0	0	7
3	3	0	0	0	0	0	3
4	0	0	0	0	0	0	0
5	0	0	0	0	0	1	1
Total	113	4	1	0	0	1	119

Table III.78

Comparison of M16 Qualification Information Obtained From Self-Report and 201 Files: Batch B

	201 File				
Self-Report	Missing	MKM	SPS	EXP	Total
Missing	14	0	0	0	14
MKM	3	8	2	0	13
SPS	6	23	14	3	46
EXP	4	14	15	13	46
Total	27	45	31	16	119

Explanations Given For Discrepancies	AWARDS	LETTERS	CERTIFICATES	ARTICLES 15	FLAG ACTIONS	M16 QUALIFICATIONS
• Self-report is correct	*	*	*	*	*	*
• Recently reviewed, paperwork not complete	*	*	*			
• Received while at previous assignment	*	*	*	*		
• Counted training certificates			*			
• Counted certificate/letter that accompanied award		*	*			
• Counted promotions		*				
• Company level	*		*	*		
• Didn't understand difference (e.g., between Articles 15 and FLAG actions)		*	*	*	*	
• Has been removed		*	*	*	*	
• Counted Levy alert					*	
• Not worth any points		*	*			
• Never forwarded to file		*	*			
• Outdated information						*
• Knows of discrepancy and trying to correct it		*	*			
• Might be in restricted file				*		
• What do you mean it's not there?	*	*	*			

Figure III.16. Results of outlier comparison from Self-Report information.

NAME

LAST FIRST MI

SOCIAL SECURITY NUMBER

DATE

DAY MONTH YEAR

PERSONNEL FILE INFORMATION
(Form 7)

MARKING INSTRUCTIONS

- Use only a No. 2 black lead pencil.
- Read each question carefully. Make a HEAVY BLACK MARK in the circle that corresponds to your answer. Be sure to FILL THE CIRCLE.
- Please do not make any stray marks.

CORRECT MARK ○●○○

INCORRECT MARKS ✓⊗◒⊙

USE NO. 2 PENCIL ONLY

1. Mark the circle(s) corresponding to the awards and decorations listed below that you have received. If you have received any not listed below, use the space(s) to the right of Other to write in the name(s) of the award(s) or decoration(s) and then mark the circle(s).

- ○ Air Assault Badge
- ○ Aircraft Crewman Badge
- ○ Army Achievement Medal
- ○ Army Commendation Medal (Valor or Merit)
- ○ Combat Field Medical Badge
- ○ Combat Infantry Badge
- ○ Diver's Badge
- ○ Driver and Mechanic Badge
- ○ Expert Field Medical Badge
- ○ Expert Infantry Badge
- ○ Explosive Ordnance Disposal Badge
- ○ Good Conduct Medal
- ○ Nuclear Reactor Operator Badge
- ○ Parachutist Badge
- ○ Pathfinder Badge
- ○ Purple Heart
- ○ Ranger Tab
- ○ Other: _____
- ○ Other: _____
- ○ Other: _____

For the next two questions, mark the circle corresponding to the number of Letters and Certificates of Appreciation, Commendation, Achievement that you have received. DO NOT count Letters or Certificates received for:

- Completion of AIT.
- Completion of any training courses taken after AIT.
- Completion of Head Start.
- Announcement of a promotion.
- Announcement of an award or decoration.

2. How many Letters of Appreciation, Commendation, Achievement have you received?

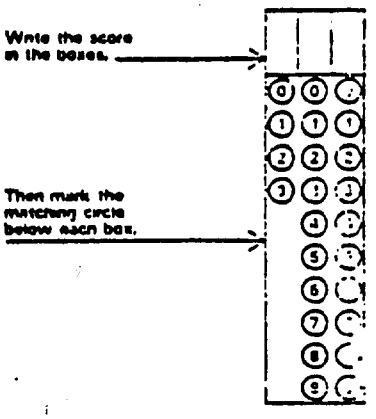
○ 0 ○ 3
○ 1 ○ 4 or more
○ 2

3. How many Certificates of Appreciation, Commendation, Achievement have you received?

○ 0 ○ 3
○ 1 ○ 4 or more
○ 2

Figure III.17. Self-Report Form for use in Concurrent Validation. (Page 1 of 2)

4. What was your last Physical Readiness Test Score? (Scores range from 0-300)

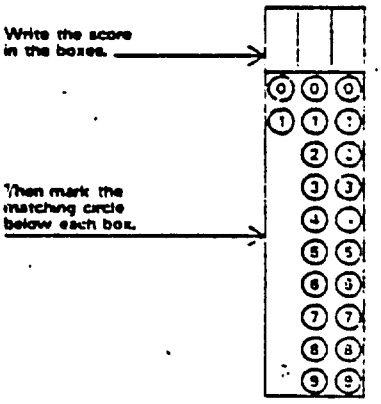


5. What was your last M16 Qualification?

○ Marksman (MKM) ○ Sharpshooter (SPS) ○ Expert (EXP)

6. If you have taken a Skill Qualification Test (SQT), what was your most recent score (SQT scores range from 0-100).

Mark here ○ if you have never taken an SQT.



7. How many Articles 15 have you received?

○ 0
○ 1
○ 2
○ 3
○ 4 or more

8. How many FLAG Actions have you received? DO NOT count a LEVY ALERT as a FLAG Action.

○ 0
○ 1
○ 2
○ 3
○ 4 or more

Figure III.17. Self-Report Form for use in the Concurrent Validation. (Page 2 of 2)

Section 16

FIELD TEST RESULTS: CRITERION INTERRELATIONSHIPS1

Up to this point we have considered the criterion field test results in terms of the item/scale characteristics and reliabilities of each measure. This view is consistent with the principal objective of the field tests which was to provide the information necessary for revising each measure as appropriate for the Concurrent Validation. The covariances among the measures were not a primary concern at this stage, since the field test samples were not large and questions of latent structure and criterion combination could be handled better with the larger samples from the Concurrent Validation.

However, knowledge of the intercorrelations is useful for uncovering potentially aberrant characteristics of the measures and for formulating analytic questions to ask of the concurrent sample data. Consequently, a selected set of intercorrelation matrixes is presented below.

Representative Criterion Intercorrelations

Some might accuse Project A of collecting a bit too much data. Such an accusation becomes credible when the intent is to calculate an intercorrelation matrix among the principal criterion measures. The list is long or short depending on how much aggregation one is willing to tolerate. If the supervisor, peer, and self ratings for all rating scales are counted, along with hands-on and knowledge test scores for each of the 30 tasks, the total number of criterion variables adds up to about 1,600. That is a few too many to interpret at a glance, without further reduction.

One strategy that could be used is cluster or factor analysis. However, rather than use empirical methods with such relatively small samples, we will delay these analyses until the concurrent data are available.

Instead, we reduced the number of variables to a much smaller number, by limiting the list to ratings obtained only from supervisors and peers, and by averaging across the 11 Army-wide scales, the 14 common task scales, the MOS task scales, and the MOS-specific BARS. For the job knowledge tests, scores were totaled for the 15 tasks measured hands-on and separately for the 15 tasks not measured hands-on.

After all this was done, the variable list was reduced to 16. The resulting matrixes are shown in Table III.79 for the nine MOS.

1This section is a revision and expansion of materials in the paper, Criterion Reduction and Combination Via a Participative Decision-Making Panel, by John P. Campbell and James H. Harris, in the ARI Research Note in preparation which supplements this Annual Report.

Table III.79

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

A. Cannon Crewman (138)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	34															
3	Avg. HO Task Rating: Peer	47	46														
	Knowledge Test																
4	All HO Tasks	41	24	18													
5	All Non-HO Tasks	21	24	06	76												
	Training Test																
6	Total Score	20	13	16	59	57											
	AW BARS																
7	Avg. Rating: Supv.	25	59	34	28	26	21										
8	Avg. Rating: Peer	29	48	54	30	24	25	64									
	AW Rating																
9	Overall Perf.: Supv.	19	53	26	28	27	24	77	48								
10	Overall Perf.: Peer	24	41	58	29	21	26	49	73	63							
11	NCO Potential: Supv.	34	47	31	27	20	13	65	44	49	34						
12	NCO Potential: Peer	28	38	50	24	16	16	54	71	43	68	46					
	MOS BARS																
13	Avg. Rating: Supv.	38	63	39	35	30	26	72	48	57	45	65	45				
14	Avg. Rating: Peer	35	53	68	27	17	26	52	74	36	65	44	61	62			
	AW Common Task Rating																
15	Avg. Rating: Supv.	23	51	41	26	28	20	65	44	60	38	46	37	72	53		
16	Avg. Rating: Peer	17	36	67	26	02	30	03	62	27	59	21	44	29	65	36	

(Continued)

*Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

B. Motor Transport Operator (MOS 64C)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	32															
3	Avg. HO Task Rating: Peer	22	70														
	Knowledge Test																
4	All HO Tasks	58	23	10													
5	All Non-HO Tasks	35	33	14	65												
	Training Test																
6	Total Score	31	23	01	58	74											
	AW BARS																
7	Avg. Rating: Supv.	22	67	55	17	21	21										
8	Avg. Rating: Peer	19	60	61	17	22	16	78									
	AW Rating																
9	Overall Perf.: Supv.	19	65	47	13	13	15	23	68								
10	Overall Perf.: Peer	11	59	63	20	18	14	72	79	57							
11	NCO Potential: Supv.	22	56	44	14	20	28	78	60	78	50						
12	NCO Potential: Peer	06	39	50	05	10	05	67	77	57	73	54					
	MOS BARS																
13	Avg. Rating: Supv.	22	25	13	23	24	13	25	21	20	22	23	14				
14	Avg. Rating: Peer	08	56	68	20	28	20	56	72	47	67	42	53	25			
	AW Common Task Rating																
15	Avg. Rating: Supv.	27	69	48	25	19	20	67	59	58	52	56	50	24	48		
16	Avg. Rating: Peer	20	40	51	12	13	06	42	55	39	58	37	55	23	51	57	

(Continued)

* Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

C. Administrative Specialist (MOS 71L)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	23															
3	Avg. HO Task Rating: Peer	16	77														
	Knowledge Test																
4	All HO Tasks	52	12	02													
5	All Non-HO Tasks	43	08	05	68												
	Training Test																
6	Total Score	54	22	23	63	51											
	AW BARS																
7	Avg. Rating: Supv.	24	54	36	19	23	24										
8	Avg. Rating: Peer	10	50	40	04	06	14	80									
	AW Rating																
9	Overall Perf.: Supv.	33	49	23	19	20	25	69	64								
10	Overall Perf.: Peer	17	35	48	13	22	18	67	78	60							
11	NCO Potential: Supv.	24	45	20	07	17	22	64	52	62	37						
12	NCO Potential: Peer	13	35	41	06	15	20	68	76	41	28	40					
	MOS BARS																
13	Avg. Rating: Supv.	23	60	39	15	10	24	68	48	60	26	54	17				
14	Avg. Rating: Peer	23	55	57	09	14	24	62	60	56	37	51	38	78			
	AW Common Task Rating																
15	Avg. Rating: Supv.	08	41	25	02	03	06	49	26	46	23	33	13	60	13		
16	Avg. Rating: Peer	30	20	07	21	24	29	46	37	46	24	49	18	38	10	59	

(Continued)

*Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

D. Military Police (958)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	27															
3	Avg. HO Task Rating: Peer	31	65														
	Knowledge Test																
4	All HO Tasks	11	17	14													
5	All Non-HO Tasks	21	15	08	60												
	Training Test																
6	Total Score	11	05	10	43	56											
	AW BARS																
7	Avg. Rating: Supv.	26	68	53	23	25	22										
8	Avg. Rating: Peer	35	56	70	09	17	21	76									
	AW Rating																
9	Overall Perf.: Supv.	33	65	45	17	24	17	83	66								
10	Overall Perf.: Peer	28	57	70	13	15	15	70	84	64							
11	NCO Potential: Supv.	28	64	49	18	19	21	73	59	77	56						
12	NCO Potential: Peer	28	57	67	13	18	27	67	84	57	82	54					
	MOS BARS																
13	Avg. Rating: Supv.	23	79	58	16	18	12	75	65	75	59	70	57				
14	Avg. Rating: Peer	31	61	79	08	07	19	60	85	56	77	53	75	72			
	AW Common Task Rating																
15	Avg. Rating: Supv.	40	69	59	18	18	20	59	50	62	53	62	47	65	52		
16	Avg. Rating: Peer	40	52	78	03	03	07	45	68	43	67	37	63	46	66	65	

(Continued)

*Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

E. Infantryman (MOS 11B)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	46															
3	Avg. HO Task Rating: Peer	36	61														
	Knowledge Test																
4	All HO Tasks	55	39	30													
5	All Non-HO Tasks	41	29	28	78												
	Training Test																
6	Total Score	40	34	31	70	71											
	AW BARS																
7	Avg. Rating: Supv.	35	67	44	28	22	23										
8	Avg. Rating: Peer	37	56	58	23	26	26	79									
	AW Rating																
9	Overall Perf.: Supv.	29	61	40	24	21	12	81	63								
10	Overall Perf.: Peer	34	52	51	22	25	25	65	85	56							
11	NCO Potential: Supv.	29	55	34	25	20	15	83	60	76	56						
12	NCO Potential: Peer	41	43	47	29	31	27	62	85	47	76	4.					
	MOS BARS																
13	Avg. Rating: Supv.	41	76	53	31	25	21	84	69	77	62	73	55				
14	Avg. Rating: Peer	40	55	69	22	24	24	60	84	55	79	50	71	70			
	AW Common Task Rating																
15	Avg. Rating: Supv.	39	74	48	31	22	17	72	61	67	54	62	48	83	61		
16	Avg. Rating: Peer	43	56	68	32	33	30	59	77	54	71	44	65	63	82	68	

(Continued)

*Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

F. Armor Crewman (MOS 19E)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	09															
3	Avg. HO Task Rating: Peer	10	50														
	Knowledge Test																
4	All HO Tasks	39	19	16													
5	All Non-HO Tasks	32	13	07	74												
	Training Test																
6	Total Score	25	22	13	58	64											
	AW BARS																
7	Avg. Rating: Supv.	15	52	39	05	07	15										
8	Avg. Rating: Peer	15	25	53	07	06	21	66									
	AW Rating																
9	Overall Perf.: Supv.	12	43	33	15	15	20	82	53								
10	Overall Perf.: Peer	13	28	51	17	12	27	50	82	42							
11	NCO Potential: Supv.	18	46	36	08	09	17	77	54	66	43						
12	NCO Potential: Peer	25	32	46	15	10	24	57	79	44	72	50					
	MOS BARS																
13	Avg. Rating: Supv.	14	70	50	27	22	30	69	40	60	37	60	42				
14	Avg. Rating: Peer	12	30	62	24	14	21	41	65	42	64	36	58	59			
	AW Common Task Rating																
15	Avg. Rating: Supv.	20	50	38	28	20	16	52	27	50	31	37	26	54	33		
16	Avg. Rating: Peer	15	34	63	25	14	22	35	53	38	54	26	46	42	64	51	

(Continued)

*Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

G. Radio Teletype Operator (MOS 31C)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	17															
3	Avg. HO Task Rating: Peer	18	71														
	Knowledge Test																
4	All HO Tasks	37	21	20													
5	All Non-HO Tasks	35	18	22	58												
	Training Test																
6	Total Score	26	21	23	50	40											
	AW BARS																
7	Avg. Rating: Supv.	08	59	46	15	13	13										
8	Avg. Rating: Peer	08	42	62	17	23	21	68									
	AW Rating																
9	Overall Perf.: Supv.	11	64	51	11	15	15	91	61								
10	Overall Perf.: Peer	-01	47	61	14	17	17	58	83	54							
11	NCO Potential: Supv.	19	59	44	26	21	19	83	56	80	48						
12	NCO Potential: Peer	20	42	61	19	28	24	59	81	57	71	60					
	MOS BARS																
13	Avg. Rating: Supv.	10	70	59	23	13	26	75	55	75	55	61	50				
14	Avg. Rating: Peer	08	49	74	11	16	21	56	82	52	75	44	64	66			
	AW Common Task Rating																
15	Avg. Rating: Supv.	12	48	39	07	08	14	51	26	52	30	46	34	49	33		
16	Avg. Rating: Peer	03	33	54	02	11	03	32	53	37	50	31	49	25	43	23	

(Continued)

*Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

H. Light-Wheel Vehicle Mechanic (MOS 63B)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	18															
3	Avg. HO Task Rating: Peer	12	59														
	Knowledge Test																
4	All HO Tasks	31	08	23													
5	All Non-HO Tasks	32	15	28	66												
	Training Test																
6	Total Score	37	22	19	52	61											
	AW BARS																
7	Avg. Rating: Supv.	19	71	57	15	18	16										
8	Avg. Rating: Peer	15	49	66	26	30	15	73									
	AW Rating																
9	Overall Perf.: Supv.	24	71	58	16	22	23	85	65								
10	Overall Perf.: Peer	18	49	52	25	22	10	65	81	60							
11	NCO Potential: Supv.	11	60	51	06	10	12	78	54	67	57						
12	NCO Potential: Peer	16	56	54	21	23	12	66	81	63	80	56					
	MOS BARS																
13	Avg. Rating: Supv.	20	80	61	14	16	24	90	68	81	60	72	63				
14	Avg. Rating: Peer	21	64	72	24	26	25	69	70	70	65	54	65	77			
	AW Common Task Rating																
15	Avg. Rating: Supv.	07	59	46	06	23	13	62	57	63	59	54	57	66	63		
16	Avg. Rating: Peer	-03	36	46	18	35	26	43	57	39	49	31	46	43	48	63	

(Continued)

* Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

Table III.79 (Continued)

Intercorrelation Matrixes for 16 Criterion Measures Obtained During Criterion Field Tests, by MOS

I. Medical Specialist (MOS 91A)

	MEASURE*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	Hands-On Test																
1	Total Score																
	Task Performance Rating																
2	Avg. HO Task Rating: Supv.	16															
3	Avg. HO Task Rating: Peer	19	61														
	Knowledge Test																
4	All HO Tasks	21	00	03													
5	All Non-HO Tasks	21	04	05	61												
	Training Test																
6	Total Score	07	-07	-08	39	31											
	AW BARS																
7	Avg. Rating: Supv.	12	59	44	09	14	04										
8	Avg. Rating: Peer	13	42	50	09	09	07	79									
	AW Rating																
9	Overall Perf.: Supv.	14	49	35	13	16	11	89	67								
10	Overall Perf.: Peer	13	39	45	09	05	23	70	89	62							
11	NCO Potential: Supv.	13	51	42	10	17	01	81	62	82	57						
12	NCO Potential: Peer	15	38	48	09	10	17	64	81	58	75	58					
	MOS BARS																
13	Avg. Rating: Supv.	14	61	50	11	20	13	78	67	68	63	59	57				
14	Avg. Rating: Peer	12	39	64	08	05	15	61	80	50	77	43	68	68			
	AW Common Task Rating																
15	Avg. Rating: Supv.	13	55	27	05	19	-10	62	42	54	43	50	32	60	38		
16	Avg. Rating: Peer	20	43	48	02	14	-01	45	54	35	54	30	50	50	64	54	

(Continued)

*Code: AW Army-Wide
BARS Behaviorally Anchored Rating Scale
HO Hands-On

These matrixes illustrate some basic truths. First, the methods correlate more highly within themselves than they do across measures. If one were to examine a multimethod (hands-on, knowledge tests, ratings), multitrait (the 15 tasks) matrix and submit it to a factor analysis, the factors would most likely be defined by methods rather than job tasks. This is not unlike what happens when individual assessment center measures are factored. Factors tend to be defined by the particular exercise or test rather than the trait (Sackett & Dreher, 1982).

However, two points are crucial. Although the variables of task proficiency, job knowledge, and general soldiering performance are certainly not independent, they are also far from being identical in spite of the influence of method variance. Also, one of the great unanswered questions in applied psychology remains: Is what we refer to as method variance (e.g., halo) in ratings, paper-and-pencil knowledge tests, or job sample tests really relevant and valid, or is it simply noise? It is not necessarily error, but may indeed reflect individual differences in performance that are quite relevant.

True Score Relationships

The intercorrelations in the previous table are between uncorrected scores on each variable. To get closer to the "truth" about the criterion space, the intercorrelations were corrected for attenuation, which yielded an estimate of the true score intercorrelation matrix. As illustrations, matrixes for MOS 11B and MOS 71L are shown in Tables III.80 and III.81.

These correlations were computed on the assumption that the most accurate portrayal of the structure of the criterion space is provided by the interrelationships among the true scores. Estimating true score correlations by correcting for attenuation is a dangerous business that must be carefully done. The reliabilities that were used for Tables III.80 and III.81 are conservative in that they do not include all the sources of error that might account for unreliability. For example, variability across testing occasions is not counted here but it might in fact serve to lower the correlations between pairs of variables (e.g., hands-on and knowledge tests). Also, to give more stability to the estimates the adjusted correlations for supervisor and peer ratings were simply averaged.

Looking at the true score intercorrelations, it seems reasonable to conclude that the hands-on measures and the knowledge tests designed to be parallel to them share a significant proportion of their variance. The Army-wide rating measures of general soldier performance are by no means independent of the job sample measures, but they have less in common with job samples than do the knowledge tests.

One large difference between Tables III.80 and III.81 is in the lower correlations for MOS 71L between the ratings and the other variables, particularly the ratings of specific task performance. However, Administrative Specialists tend to work more in isolation than other MOS and are not observed as closely. It all seems to make reasonable sense.

Table III.80

Intercorrelations Among Selected Criterion Measures for Infantryman (MOS 11B)[a]

	Measure	1	2	3	4	5
1	Total Score on all HO Tasks	( )	.67	.86	.60	.57
2	Avg. of 15 HO task ratings (Supv. + Peer)[b]	.41	( )	.44	.39	.73
3	Total score on Job Knowledge Test	.55	.35	( )	.80	.31
4	Total score on Training Knowledge Test	.40	.33	.70	( )	.25
5	AW BARS - Overall Effectiveness (Supv. + Peer)	.32	.51	.23	.19	( )

[a] Correlations corrected for attenuation are above the diagonal.
[b] The corresponding correlations for supervisory ratings and peer ratings were averaged.

Table III.81

Intercorrelations Among Selected Criterion Measures for Administrative Specialist (MOS 71L)[a]

	Measure	1	2	3	4	5
1	Total Score on all HO Tasks	( )	.28	.76	.73	.35
2	Avg. of 15 HO task ratings (Supv. + Peer)[b]	.20	( )	.10	.29	.51
3	Total score on Job Knowledge Test	.52	.07	( )	.82	.22
4	Total score on Training Knowledge Test	.54	.23	.63	( )	.27
5	AW BARS - Overall Effectiveness (Supv. + Peer)	.25	.39	.16	.22	( )

[a] Correlations corrected for attenuation are above the diagonal.
[b] The corresponding correlations for supervisory ratings and peer ratings were averaged.

Summary

In general, the covariance among criteria did not reveal any fatal flaws in the array of measures constructed to cover the domain of job performance. While there is considerable method variance among the ratings, there is also a positive manifold in the matrixes, which suggests that there is indeed a latent structure to be investigated. Performance is certainly not one thing and the pattern of correlations is conceptually sensible. It whets the appetite for a more systematic investigation of the latent structure of job performance. Some speculations about that structure are summarized in the following subsection.

A Plausible Model

Even though the major data collections and analyses are yet to come, a great deal has been learned to date about the domain of first-tour enlisted performance. The total domain was described via two major collections of critical incidents, systematic examination of all available job survey data, review of all job specification documentation, careful analysis of AIT Programs of Instruction, and multiple reviews and elaborations by many panels of expert judges. Subsequent to the job and task descriptions, multiple methods of performance measurement were developed, pilot tested, revised, field tested, revised, reviewed, and revised again.

As a consequence, we have formed some further ideas of how the latent structure of job performance might look when cast against our operational measures. This model is not meant to be definitive or even based on the most relevant data (e.g., the covariances to be obtained in the Concurrent Validation). Rather it is meant to be consistent with what we know so far and to illustrate the kind of job performance framework toward which we are working.

As a first attempt at portraying the latent structure, suppose we suggest that the enlisted performance domain is made up of the following general factors:

(1) Maintaining and upgrading current job knowledge (including common tasks). A legitimate question here might be why the mere possession of job knowledge should be a factor in the performance domain. However, if a major goal of the Army is to be ready to enter a conflict on short notice, then possessing a high degree of current knowledge is performance. Having the proper information and being able to use it (Factor 2) are not the same thing. However, neither are they independent. Consequently, our model must stipulate that these first two factors are significantly correlated and the relationship stems both from sharing common requirements (e.g., general cognitive ability) and from Factor 1 being, in part, a "cause" of performance differences on Factor 2.

(2) Having technical proficiency on the primary job tasks. This factor refers to being able to perform on the technical content, be it complicated or simple. Technical is defined broadly but not so broadly as to include leadership or other interpersonal interaction task requirements. Within this construct the content of the tasks may vary considerably and rely on very different abilities (e.g., playing a musical instrument vs. repairing a truck

generator). For most jobs it might also be possible to think of two such general factors, executing "standard" procedures and troubleshooting special problems.

(3) Exhibiting peer leadership and support. Enlisted personnel often have the opportunity to teach, support, or provide leadership for their peers. This factor refers to the frequency and proficiency with which people do that when the occasion arises. It would also be reasonable to think of this factor as composed of the four subgeneral factors that have been found in leadership research (e.g., Bowers & Seashore, 1966): goal setting, facilitating goal attainment, one-on-one individual support, and facilitating group morale.

(4) Demonstrating commitment to Army regulations and traditions. Performance on this factor refers to maintaining living quarters and equipment, and maintaining a high level of physical fitness and appropriate military appearance. This factor is perhaps a bit more tenuous than the others. Defining it this way assumes that all of the different elements will covary to a high degree.

(5) Continuing to perform under adverse conditions. This factor would share many components in common with the previous three and thus should not be orthogonal. However, the act of carrying out job assignments when wet, tired, or in danger is viewed as a very important and distinct aspect of performance.

(6) Avoiding serious disciplinary problems. Incurring disciplinary actions because of problems with drugs, alcohol, neglect of duty, or serious interpersonal conflict represents a great cost to the Army. Successfully avoiding these costs is viewed as an important factor in overall performance.

Standing in a direct causal relation to the performance factors are knowledges and skills learned dur ng training, abilities and other individual characteristics present at the time of hire, and the choice to perform, which is supposedly under motivational control. For our purposes here, the causal latent variables of most concern are the knowledges, skills, and motivational predispositions acquired during training. Consequently, we might posit that there are three major training performance factors in the latent structure:

- Hands-on task proficiency.
- General job knowledge.
- Exhibition of good soldiering skills and discipline.

A very rough schematic that portrays these latent variables and lists the observable measures of these variables that we have available in Project A is shown in Figure III.18. The arrows between latent variables and operational measures indicate an expected correlation; the expected size of the correlation is not indicated. Arrows between latent variables indicate a hypothesized causal relation.

Several points can be made about this picture. First, the principal data upon which the list of latent constructs is based are the results of the critical incident workshops conducted during the development of the behaviorally anchored ratings scales. We have not yet had the opportunity to

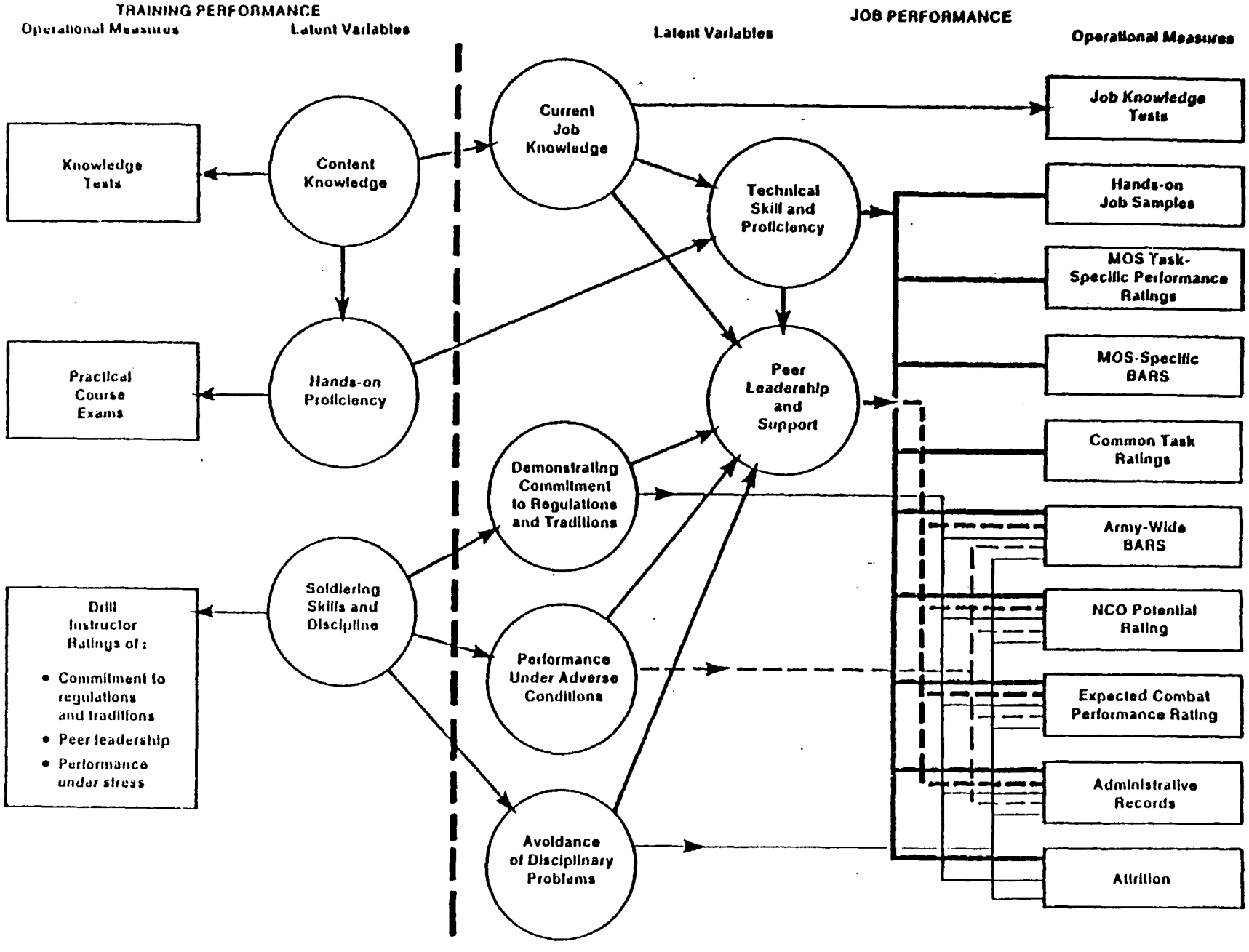


Figure III.18. Job Performance -- A Proposed Structural Model.

examine the factor structure of the hands-on measures or knowledge tests, or even to look comprehensively at the factor structure of all the rating scales. These analyses really must wait until larger sample sizes are available with the revised measures.

Second, the manifest job performance variables are by no means "pure" measures of the latent constructs. For example, Factor 2 would seem to underlie virtually all of the observable measures. By contrast, the "avoidance of disciplinary problems" should influence only some of the Army-wide BARS scales, NCO potential, attrition, and perhaps expected combat performance. However, in general, most of the observable variables are probably multiply determined.

Third, the above reasoning suggests that if the operational criteria share so many common determinants, they probably should not receive grossly differential weights when combined into composites for the purpose of test validation.

Fourth, differential prediction of job performance across jobs must come from different requirements for success on Factors 1 and 2 (e.g., psychomotor abilities vs. verbal ability). To a certain extent it could also result from a greater weight being given in some MOS to peer leadership and performance under adverse conditions.

Fifth, limiting measures of training success to paper-and-pencil tests of knowledges mastered is probably not sufficient. To more completely determine the relationship of training performance, additional measures would be required.

Sixth, the causal relations among the individual differences present at the time of entry, the latent variables making up training performance, the latent variables that constitute job performance, and the operational criterion measures can be described with brilliant understatement by saying that they are complex. As part of that complexity it seems reasonable to assert that:

- Among the latent variables describing training performance, hands-on proficiency and content knowledge are most likely more highly related to each other than either is to soldiering skills and discipline. Further, content knowledge stands in at least a partial causal relation to hands-on proficiency.

- Among the latent variables describing job performance, job knowledge would seem to come first in the causal chain since it at least partially determines technical proficiency. However, both these factors most likely cause at least some of the individual differences in peer leadership performance. A causal relation between technical proficiency and either commitment to regulations/traditions or avoiding disciplinary problems does not seem so likely. However, some may wonder whether commitment to regulations/traditions and avoidance of disciplinary problems are bipolar.

- If the first two factors were measured with high construct validity, Factor 1 (current job knowledge) should have a direct effect only on job knowledge tests. Job knowledge should create differences on other operational measures only through its influence on technical proficiency. Consequently, if technical proficiency could be held constant, the observed correlations between job knowledge tests and all other variables should be reduced to zero.

- Since peer leadership and support was given a broad definition (in terms of leadership theory), greater knowledge, higher technical skill, higher commitment, demonstrated performance under stress, and an exemplary record would all "cause" an individual to exhibit more effective peer leadership.

- As somewhat of a contrast, performance under adverse conditions is conceptualized as a dispositional variable. Consequently it would be under motivational control and not a function of knowledge or ability.

The reader should keep firmly in mind that the above comments are still speculative. Such a model of performance will go through many iterations before Project A is finished. However, the first major empirical specifications will be based on the Concurrent Validation, which began in FY85 and which will be analyzed in FY86. The concluding sections of this report (Part IV) briefly outline the data collection procedure and the analysis plan for this data set.

PART IV

CONCURRENT VALIDATION

Included in Part IV are a listing of the predictor and criterion arrays used in the Concurrent Validation, a description of the samples, a brief summary of the procedures used for data collection, and an outline of the analyses that will be carried out. The data collection itself and the basic data analyses will be completed during FY86.

Section 1

CONCURRENT VALIDATION: PREDICTOR AND CRITERION VARIABLES[1]

Parts II and III of this report have described the development and field testing of the predictor tests and performance measures to be used in the Concurrent Validation phase of Project A. The predictors were common across all jobs, but not all of the performance measures were used with every MOS in the total Concurrent Validation sample.

The nomenclature for MOS groupings also changed slightly for the Concurrent Validation phase. Batch A and Batch B MOS are now known collectively as Batch A; they are the nine MOS that were used in the criterion field tests. The remaining 10 MOS are still designated as Batch Z. Batch A and Batch Z MOS are listed in Table IV.1.

Table IV.1

Project A MOS Used in the Concurrent Validation Phase

Batch A		Batch Z	
11B	Infantryman	12B	Combat Engineer
13B	Cannon Crewman	16S	MANPADS Crewman
19E	Armor Crewman	27E	TOW/Dragon Repairer
31C	Radio Teletype Operator	51B	Carpentry/Masonry Specialist
63B	Light-Wheel Vehicle Mechanic	54E	Chemical Operations Specialist
64C	Motor transport Operator	55B	Ammunition Specialist
71L	Administrative Specialist	67N	Utility Helicopter Repairer
91A	Medical Specialist	76W	Petroleum Supply Specialist
95B	Military Police	76Y	Unit Supply Specialist
		94B	Food service Specialist

While the same predictor battery was used for each MOS, the criterion measures used for Batch A MOS were different than those used for MOS in Batch Z. The major distinction is that the MOS-specific job performance and job knowledge measures developed for the Batch A MOS were not prepared for the 10 MOS in Batch Z. For these jobs only the Army-wide measures and the training achievement tests were available.

The complete array of predictors in the Trial Battery is listed in Table IV.2. The list of criteria for Batch A and Batch Z are shown in Table IV.3.

---

[1]Part IV is based in part on a paper, Criterion Reduction and Combination Via a Participative Decision-Making Panel, by John P. Campbell and James H. Harris, in the ARI Research Note in preparation, which supplements this Annual Report.

Table IV.2

Summary of Predictor Measures Used in Concurrent Validation: The Trial Battery

Name	Number of Items
COGNITIVE PAPER-AND-PENCIL TESTS	
Reasoning Test (Induction - Figural Reasoning)	30
Object Rotation Test (Spatial Visualization - Rotation)	90
Orientation Test (Spatial Orientation)	24
Maze Test (Spatial Visualization - Scanning)	24
Map Test (Spatial Orientation)	20
Assembling Objects Test (Spatial Visualization - Rotation)	32
COMPUTER-ADMINISTERED TESTS	
Simple Reaction Time (Processing efficiency)	15
Choice Reaction Time (Processing efficiency)	30
Memory Test (Short-term memory)	36
Target Tracking Test 1 (Psychomotor precision)	18
Perceptual Speed and Accuracy Test (Perceptual speed and accuracy)	36
Target Tracking Test 2 (Two-hand coordination)	18
Number Memory Test (Number operations)	28
Cannon Shoot Test (Movement judgment)	36
Identification Test (Perceptual speed and accuracy)	36
Target Shoot Test (Psychomotor precision)	30
NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES	
Assessment of Background and Life Experiences (ABLE)	209
Adjustment Dependability Achievement Physical Condition Leadership Locus of Control Agreeableness/Likeability	
Army Vocational Interest Career Examination (AVOICE)	176
Realistic Interests Conventional Interests Social Interests Enterprising Interests Artistic Interests	

Table IV.3

Summary of Criterion Measures Used in Batch A and Batch Z Concurrent Validation Samples

Performance Measures Common to Batch A and Batch Z

- Army-Wide Rating Scales (all obtained from both supervisors and peers).
  - Ten behaviorally anchored rating scales (BARS) designed to measure factors of non-job-specific performance.
  - Single scale rating of overall effectiveness.
  - Single scale rating of NCO potential.
- Combat prediction scale containing 41 items.
- Paper-and-Pencil Test of Training Achievement developed for each of the 19 MOS (130-210 items each).
- Personnel File Information form developed to gather objective archival records data (awards and letters, rifle marksmanship scores, physical training scores, etc.).

Performance Measures for Batch A Only

- Job Sample (Hands-On) tests of MOS-specific task proficiency.
  - Individual is tested on each of 15 major job tasks in an MOS.
- Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.
  - Individual is scored on 150 to 200 multiple-choice items representing 30 major job tasks. Ten to 15 of the tasks were also measured hands-on.
- Rating scale measures of specific task performance on the 15 tasks also measured with the knowledge tests. Most of the rated tasks were also included in the hands-on measures.
- MOS-specific behaviorally anchored rating scales (BARS). From 6 to 10 BARS were developed for each MOS to represent the major factors that constitute job-specific technical and task proficiency.

Performance Measures for Batch Z Only

- Army-Wide Rating Scales (all obtained from both supervisors and peers).
  - Ratings of performance on 11 common tasks (e.g., basic first aid).
  - Single scale rating on performance of specific job duties.

Auxiliary Measures Included in Criterion Battery

- A Job History Questionnaire which asks for information about frequency and recency of performance of the MOS-specific tasks.
- Work Questionnaire - a 44-item questionnaire scored on 14 dimensions descriptive of the job environment.
- Measurement Method Rating obtained from all participants at the end of the final testing session.

Section 2

CONCURRENT VALIDATION: SAMPLES AND PROCEDURES

The original Project A Research Plan specified a Concurrent Validation target sample size of 600-700 job incumbents for each of the 19 MOS, and a tentative starting date of 1 May 1985, using procedures that had been tried out and refined during the predictor and criterion field tests. The Research Plan specified 13 data collection sites in the United States (CONUS) and two in Europe (USAREUR). The number of sites was the maximum that could be visited within the project's budget constraints, which dictated that sites should be chosen to maximize the probability of obtaining the required sample sizes.

The data collection actually began 10 June 1985 and was not yet concluded by the end of FY85 (30 September 1985). The projected schedule, by site, is shown in Figure IV.1. Although the starting date shifted slightly, it was still within the permissible "window" that would maintain the project's original schedule.

The data were collected by on-site teams made up of project staff. Each square in Figure IV.1 represents 1 week of one team's time. For example, during the week of 7 July seven teams were operating, one at each of seven posts.

Fort Lewis	10 Jun - 18 Jul	Fort Knox	15 Aug - 13 Sep
Fort Benning	17 Jun - 25 Jul	Fort Sill	3 Sep - 4 Oct
Fort Riley	8 Jul - 9 Aug	Fort Campbell	3 Sep - 14 Oct
Fort Carson	8 Jul - 23 Aug	Fort Polk	30 Sep - 13 Nov
Fort Hood	8 Jul - 27 Aug	Fort Bliss	1 Oct - 31 Oct
Fort Stewart	22 Jul - 30 Aug	Fort Ord	7 Oct - 15 Nov
Fort Bragg	1 Aug - 13 Sep	USAREUR	12 Jul - 8 Oct
			12 Jul - 9 Aug
			20 Sep - 18 Oct

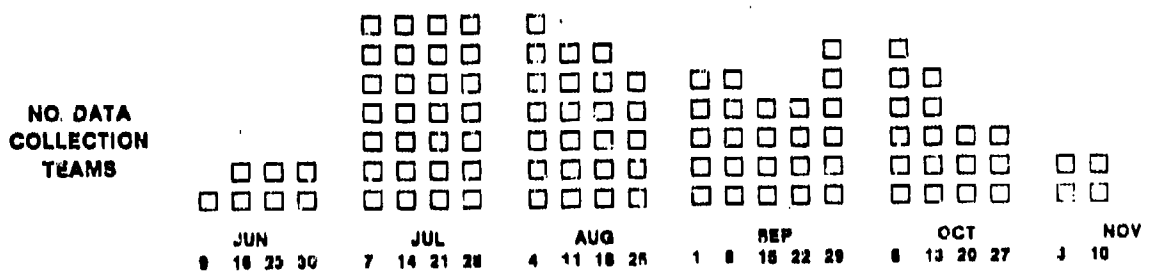


Figure IV.1. Concurrent Validation schedule.

Since data collection was not concluded until FY86, a detailed description of the Concurrent Validation procedure and results must wait for a future report. However, the basic sampling plan, team training, and data collection procedures are summarized below. An outline of the planned data analysis steps is presented in the following section (i.e., Section 3 of Part IV).

Cross-Validation Sample

The general sampling plan was to use the Army's World-Wide Locator System to identify, at the specified sites, all the first-term enlisted personnel in Batch A and Batch Z MOS who entered the Army between 1 July 1983 and 31 July 1984. If possible, the individual's unit identification was also to be obtained. The steps described below were then followed. The intent was to be as representative as possible while preserving enough cases within units to provide a "within rater" variance estimate for the supervisor and peer ratings.

Sampling Plan: Concurrent Validation

A. Preliminary activities:

1. Identify the subset of MOS (within the sample of 19) for which it would be possible to actually sample people within units at specific posts. That is, given the entry date "window" and given that only 50-75 percent of the people on any list of potential subjects could actually be found and tested, what MOS are large enough to permit sampling to actually occur? List them.

2. For each MOS in the subset of MOS for which sampling is possible, identify the smallest "unit" from which 6-10 people can be drawn. Ideally, we would like to sample 4-6 units from each post and 6-12 people from each unit. For the total concurrent sample this would provide enough units to average out or account for differential training effects and leadership climates, while still providing sufficient degrees of freedom for investigating within-group effects such as rater differences in performance appraisal.

3. For the four MOS in the Preliminary Battery sample, identify the members of the PB sample who are on each post.

B. The ideal implementation would be to obtain the Alpha Roster list of the total population of people at each post who are in the 19 MOS and who fit our "window". The lists would be sent to HumRRO where the following steps would be carried out:

1. For each MOS, randomize units and randomize names within units.

2. Select a sample of units at random. The number would be large enough to allow for some units being truly unobtainable at the time of testing.

3. Instruct the Point-of-Contact (POC) at the post to obtain the required number of people by starting at the top of the list and working down (as in the Batch A field test) within each of the designated units. If an entire unit is unavailable, go on to the next one on the list.

4. In those MOS for which unit sampling is not possible, create a randomized list of everyone on the post who fits the window. Instruct the POC to obtain the required number by going down the list from top to bottom (as in the Batch A field tests).

C. If it is not possible to bring the Alpha Roster to HumRRO, provide project staff at the post to assist the POC in carrying out the above steps.

1. If it is not possible to randomize names at the post, first use the World-Wide Locator to obtain a randomized list, carry the list to the post and use it to sample names from units drawn from a randomized list of units. If there are only six to eight units on the post, then no sampling of units is possible. Use them all.

D. If it is not possible for project personnel to visit the post, then provide the randomized World-Wide Locator list to the POC. Ask him or her to follow the sampling plan described above; supply written and telephone assistance. That is, the POC would identify a sample of units (for those MOS for which this is possible), match the unit roster with the randomized World-Wide Locator list, and proceed down each unit until the required number of people was obtained. If the POCs can generate their own randomized list from the Alpha Roster, so much the better. The World-Wide Locator serves only to specify an a priori randomized list for the POC.

E. If none of the above options is possible, then present the POC with the sampling plan and instruct him or her to obtain the required number of people in the most representative way possible (the Batch B procedure).

Actual Samples Obtained

The final sample sizes are shown by post and by MOS in Figure IV.2. Note that it was not always possible in all MOS to find as many as 600 incumbents with the appropriate accession dates at the 15 sites. Some MOS simply are not that big.

	BATCH A									BATCH Z											
MOS / Location	11B	13B	19E	31C	63B	64C	71L	91A	95B	12B	16S	27E	51B	54E	55B	67N	76W	76Y	94B	Total	% Total
Fort Benning	45	23	41	7	13	39	16	9	13	13	15	3	0	12	18	9	13	15	12	318	3.35
Fort Bliss	0	20	30	15	61	45	17	0	44	15	5	2	0	14	0	12	6	31	30	347	3.68
Fort Bragg	68	46	0	0	37	25	41	10	72	82	75	13	19	72	20	7	42	39	62	730	7.74
Fort Campbell	90	28	0	20	60	45	54	44	43	90	23	10	0	32	18	42	51	61	46	757	8.03
Fort Carson	60	50	77	30	49	53	30	33	46	49	57	13	0	25	7	0	23	40	47	689	7.31
Fort Hood	26	56	0	30	40	28	38	50	60	51	60	4	12	62	36	44	72	41	57	767	8.13
Fort Knox	29	32	111	16	38	48	22	45	31	43	10	6	0	8	12	0	10	29	34	524	5.56
Fort Lewis	75	46	13	11	43	46	23	27	56	27	25	1	11	51	31	20	48	41	36	631	6.69
Fort Ord	30	0	0	14	30	42	31	43	51	51	7	8	1	4	7	15	23	40	28	425	4.51
Fort Polk	73	47	19	29	47	47	18	46	44	60	45	9	8	16	7	23	26	51	35	648	6.87
Fort Riley	30	43	55	27	26	45	35	30	40	31	20	8	8	25	52	0	20	39	45	579	6.14
Fort Sill	0	108	0	20	43	51	44	0	29	42	11	0	0	0	0	15	7	35	32	437	4.63
Fort Steward	44	46	39	17	28	51	31	45	45	30	39	9	8	17	29	26	44	34	35	617	6.54
USAREUR	132	122	120	130	122	121	114	119	118	120	78	61	41	96	54	63	105	134	113	1963	20.8
Total	702	667	503	366	637	686	514	501	692	704	470	147	108	434	291	276	490	630	612	9430	
% Total	7.44	7.07	5.33	3.88	6.76	7.27	5.45	5.31	7.34	7.47	4.90	1.56	1.15	4.60	3.09	2.93	5.20	6.68	6.49		

Figure IV.2. Concurrent Validation sample soldiers by MOS by location.

Data Collection Team Composition and Training

Team Composition

Each data collection team was composed of a Test Site Manager (TSM) and six or seven project staff members who were responsible for test and rating scale administration. The teams were made up of a combination of regular project staff and individuals specifically recruited for the data collection effort (e.g., graduate students). The test site manager was an "old hand" who had participated extensively in the field tests. This team was assisted by eight NCO scorers (for the hands-on tests), one company-grade officer POC, and up to five NCO support personnel, all recruited from the post.

Team Training

The project data collection teams were given 3 days of training at a central location (Alexandria, VA). During this period, Project A was explained in detail, including its operational and scientific objectives. After the logistics of how the team would operate (transportation, meals, etc.) were discussed, the procedures for data entry from the field to the computer file were explained in some detail. Emphasis was placed on reducing data entry errors at the outset by correct recording of responses and correct identification of answer sheets and diskettes.

Next, each predictor and criterion measure was examined and explained. The trainees took each predictor test, worked through samples of the knowledge tests, and role played the part of a rater. Considerable time was spent on the nature of the rating scales, rating errors, rater training, and the procedures to be used for administering the ratings. All administrative manuals, which had been prepared in advance, were studied and pilot tested, role playing exercises were conducted, and hands-on instruction for maintenance of the computerized test equipment was given.

The intent was that by the end of the 3-day session each team member would (a) be thoroughly familiar with all predictor tests and performance measures, (b) understand the goals of the data collection, (c) have had an opportunity to practice administering the instruments and to receive feedback, and (d) be committed to making the data collection as error-free as possible.

Hands-On Scorer Training

As noted above, eight NCO scorers were required for Hands-On test scoring. They were recruited and trained using procedures very similar to those used at each post in the criterion field tests. Training took place over 1 full day and consisted of (a) a thorough briefing on Project A, (b) an opportunity to make the tests themselves, (c) a check of the specified equipment, and (d) multiple practice trials in scoring each task, with feedback from the project staff. The intent was to develop high agreement for the precise responses that would be scored as GO or NO-GO on each step.

Data Collection Procedure

The Concurrent Validation administration schedule for a typical site (Fort Stewart, Georgia) is shown in Figure IV.3. The first day (22 Jul 85) was devoted to equipment and classroom set-up, general orientation to the data collection environment, and a training and orientation session for the post POC and the NCO support personnel.

BATCH A MOS 4 Blocks, 4 Hours Each		BATCH Z MOS 2 Blocks, 4 Hours Each	
Block 1	Predictor Tests	Block 1	Predictor Tests
Block 2	School and Job Knowledge Tests Army-Wide Ratings	Block 2	School and Job Knowledge Tests Army-Wide Ratings
Block 3	MOS-Specific Hands-On Tests		
Block 4	MOS Ratings MOS-Specific Written Tests		

Figure IV.3 Concurrent Validation test outline.

On the first day of actual data collection (23 Jul 85), 30 MOS 12B soldiers arrived at the test site at 0745. The 30 soldiers were divided randomly into two groups of 15 soldiers each, identified as Group 1 or 2. Each group was directed to the appropriate area to begin the administration for that group. The measures administered in each block of testing are shown in Figure IV.3. The groups rotated under the direction of the test site manager through the appropriate blocks according to the schedule shown in Figure IV.4.

For soldiers in a Batch Z MOS, like 12B, the procedure took 1 day. For soldiers in a Batch A MOS, like MOS 91A, the procedure was similar but took 2 days to rotate the soldiers through the appropriate blocks, as shown in the 6-7 August schedule at Fort Stewart.

Fort Stewart, GA
Concurrent Validation
22 July - 30 August 1985

	Soldiers for 2 Days* (Batch A)				(Batch Z) Soldiers for 1 Day*	
Groups of 15	1	2	3	4	1	2
22 July AM	Training/Orientation for Data Collection					
PM	"				"	
					30 12B 10-20 Soldiers	
23 July AM					P	K/R
PM					R/K	P
					32 27E 10-20 Soldiers	
24 July AM					K/R	P
PM					P	R/K
					30 55B 10 - 20 Soldiers	
25 July AM					P	K/R
PM					R/K	P
					30 55B 10-20 Soldiers	
26 July AM					K/R	P
PM					P	R/K
					30 76W 10-20 Soldiers	
29 July AM					P	K/R
PM					R/K	P
					30 76W 10-20 Soldiers	
30 July AM					K/R	P
PM					P	R/K
					30 94B 10-20 Soldiers	
31 July AM					P	K/R
PM					R/K	P

Figure IV.4. Sample schedule for Concurrent Validation administration.
(Page 1 of 4)

		Soldiers for 2 Days* (Batch A)				(Batch Z) Soldiers for 1 Day*	
Groups of 15		1	2	3	4	1	2
						30 16S 10-20 Soldiers	
1 Aug	AM					K/R	
	PM					P	R/K
						15 16S & 14 51B 10-20 Soldiers	
2 Aug	AM					P	K/R
	PM					R/K	P
						30 51B 10-20 Soldiers	
5 Aug	AM	Train 8 91A Scorers				P	K/R
	PM	- -				R/K	P
		45 91A 10-20 Soldiers				15 54E 10-20 Soldiers	
6 Aug	AM	P	K_3R_1	HO		K/R	
	PM	HO	R_2K_5	R_1K_5		P	
7 Aug	AM	R_1K_3	HO	P		Train 8 11B Scorers	
	PM	K_5R_2	P	R_2K_3		- -	
		45 11B 10-20 Soldiers				15 54E 10-20 Soldiers	
8 Aug	AM	P	K_3R_1	HO		K/R	
	PM	HO	R_2K_5	R_1K_5		P	
9 Aug	AM	R_1K_3	HO	P		Train 8 13B Scorers	PM
K_5R_2	P	R_2K_3			- -		
		45 13B 10-20 Soldiers				15 54E 10-20 Soldiers	
12 Aug	AM	K_3R_1	HO	K_3R_1		P	
	PM	HO	P	R_2K_5		R/K	

Figure IV.4. Sample schedule for Concurrent Validation administration.
(Page 2 of 4)

		Soldiers for 2 Days* (Batch A)				(Batch Z) Soldiers for 1 Day*	
Groups of 15		1	2	3	4	1	2
		45 13B 10-20 Soldiers				15 54E 10-20 Soldiers	
13 Aug	AM	P	R_1K_5	HO		Train 8 63B Scorers	
	PM	R_2K_5	R_2K_3	P		- -	
		30 63B 10-20 Soldiers					
14 Aug	AM	P	HO				
	PM	HO	P				
						30 67N 10-20 Soldiers	
15 Aug	AM	K_3R_1	R_1K_5	Train 8 95B Scorers		P	K/R
	PM	R_2K_5	R_2K_3	- -		R/K	P
		45 95B 10-20 Soldiers					
16 Aug	AM	K_3R_1	HO	K_3R_1			
	PM	HO	P	R_2K_5			
19 Aug	AM	P	R_1K_5	HO		Train 8 71L Scorers	
	PM	R_2K_5	R_2K_3	P		- -	
		45 71L 10-20 Soldiers				15 67N 10-20 Soldiers	
20 Aug	AM	K_3R_1	HO	K_3R_1		P	
	PM	HO	P	R_2K_5		R/K	
21 Aug	AM	P	R_1K_5	HO		Train 8 31C Scorers	
	PM	R_2K_5	R_2K_3	P		- -	

Figure IV.4. Sample schedule for Concurrent Validation administration.
(Page 3 of 4)

		Soldiers for 2 Days* (Batch A)				(Batch Z) Soldiers for 1 Day*	
Groups of 15		1	2	3	4	1	2
		30 31C 10-20 Soldiers				30 76Y 10-20 Soldiers	
22 Aug	AM	HO	K_3R_1			K/R	P
	PM	R_1K_5	HO			P	R/K
23 Aug	AM	R_2K_3	P	Train 8 64C Scorers			
	PM	P	R_2K_5	- -			
		45 64C 10-20 Soldiers				15 76Y 10-20 Soldiers	
26 Aug	AM	K_5R_1	K_5R_1	HO		P	
	PM	P	HO	R_1K_5		R/K	
27 Aug	AM	HO	K_3R_2	P	Train 8 19E Scorers		
	PM	K_3R_2	P	R_2K_3	- -		
		60 19E 10-20 Soldiers					
28 Aug	AM	K_5R_1	P	K_3R_1	HO		
	PM	HO	R_1K_3	P	K_3R_1		
29 Aug	AM	K_3R_2	HO	R_2K_5	P		
	PM	P	K_5R_2	HO	R_2K_5		
30 Aug	AM	Make-up Day					
	PM	- -					

*Legend:
R - Rating Scales
R_1 Batch A (Army-Wide, MOS BARS, Job History)
R_2 Batch A (Combat Prediction, Work Questionnaire, Personnel File Information)
R Batch Z (Army-Wide, Overall Performance, Common Tasks, Combat Prediction, Work Questionnaire, Personnel File Information)
K - Knowledge Tests
K_3 Training Achievement Tests
K_5 MOS Task-Based Tests
P - Predictor Tests
R - Rating Scales
HO - Hands-On Tests

In addition, at the end of their final session, all soldiers filled out the Measurement Method Rating (MMR)

Figure IV.4. Sample schedule for Concurrent Validation administration.
(Page 4 of 4)

Section 3

CONCURRENT VALIDATION: ANALYSIS PLAN

The analysis plan for the Concurrent Validation data is outlined in this section. The overall goal is to move systematically from the raw data, which consist of thousands of elements of information on each individual, to estimates of selection validity, differential validity, and selection/classification utility.

General Steps

The overall analysis plan consists of the following steps:

1. Prepare and edit individual data files.
2. Determine basic scores for the predictor variables.
3. Determine basic scores for the criterion variables.
4. Describe the latent structure of the predictor and criterion covariance matrices.
5. Determine how well each predictor predicts each criterion variable (for each MOS).
6. Determine how well predictive relationships generalize across criterion constructs and across MOS.
7. Examine subgroup differences in the predictive relationships and their generalizations.
8. Evaluate alternative sets of predictors in terms of maximizing classification efficiency while minimizing any predictive bias.

Data Preparation

For initial processing, the data from the field are being divided into the following groups:

Predictor Measures -

- Computer Tests - diskettes are sent to project staff for processing.
- Paper-and-Pencil Tests - booklets sent to vendor for scanning.

Criterion Measures -

- Hands-on Measures - score sheets sent to project staff for keypunching.

- Ratings, Knowledge Tests, Background, Job History, Work Questionnaire, Method Measurement, Personnel File Information - sent together to vendor for scanning.

The Roster Control File will be merged with the most recent Enlisted Master Files extracts for the FY83/84 cohort and with Applicant/Accession files. Any unmatched cases will be [illegible]cked for incorrect identifiers.

Score Generation

While the data are still separated into the different types, initial score variables will be generated. For the paper-and-pencil tests, we will generate number correct and number omitted scores, as was done in the field tests. Subsequent revisions to these scores will be implemented to reflect any scoring changes suggested during item analyses and as part of subsequent outlier analyses.

For the non-cognitive predictor tests, scale scores will be generated in the same manner as was done for the field test. A missing data screen will be implemented identifying any score where more than 10% of the component items were omitted.

For the computer-administered tests, response time, error, and other derivative scores will be generated as per the guidance from the field test results.

For the hands-on measures, the percentage of steps passed will be computed for each task. We will also compute the number of steps not scored for each task. If more than 10% of the steps were not scored for a given soldier, the task score will be identified as missing; otherwise, scores for the missing steps will be imputed as described under Missing Values below.

For the rating data, adjusted ratee means will be computed for each rating scale as was done in the field test. A separate file of the individual ratings will be maintained for reliability and other analyses.

Initial summary measures will be defined for each data collection method (hands-on, rating, and knowledge test) on the criterion side. These measures will be means of task scores or rating scales and will serve as performance factor scores pending more precise definitions of performance composites.

Outlier Analysis

Outlier analysis will be conducted for each of the predictor and criterion score variables. For test data, distribution of the number correct and number omitted will be examined. Residual errors in predicting these scores from other variables will also be examined. Any residuals larger than three standard errors will be questioned.

For the non-cognitive predictor measures, additional screening will be performed. The ABLE includes built-in validity scales. The cutoffs used with the field test data will be reviewed with project staff. The default will be to use the same cutoffs and identify all of the ABLE scale values as missing for any soldier exceeding the cutoff.

For the rating data, we will conduct outlier analysis and individual rater adjustments in the same way that we did for the field test data. This procedure involves looking at the correlation of each rater's ratings with the mean of all other raters' ratings of the same soldiers and also looking at the mean signed error. Outliers are identified in terms of these statistics and also in terms of measures of halo (lack of variability across dimensions).

Missing Values

Because extensive multivariate analyses requiring complete data will be performed, the treatment of missing values is an important concern, much more so than was the case with the field test data. Prior efforts have amounted to substituting either examinee means or variable means for missing values. We plan to use PROC IMPUTE to derive proxy values for missing scale scores (and for missing step scores in the hands-on analyses).

This procedure essentially substitutes a value observed for a respondent who was very similar to the examinee with the missing variable. This procedure has been shown to be significantly better than straight regression procedures (e.g., BMDPAM) in reproducing correlation and variance estimates, as the regression approaches tend to underestimate variances and to spuriously inflate correlations.

Data completion flags will be generated for each battery. For each set, the flag will indicate whether the data are complete, partially complete and partially missing or imputed, or entirely missing or imputed.

Predictor Score Analyses

After data preparation, basic item analyses, and the initial score generation, the principal objectives for the predictor analyses are to generate the basic summary scores that will enter the initial prediction equation in each MOS, examine the latent structure of those scores, and determine MOS and subgroup differences. The basic steps are as follows:

1. Using the initial scores, items/scale score analyses will be conducted as the final opportunity to (a) identify faulty items, (b) revise the scale by eliminating some items, and (c) arrive at the final array of predictor scores that will be entered in the predictor intercorrelation matrices.

2. Scale reliabilities and descriptive statistics will then be computed.

3. The latent structure of the predictor space will be examined via confirmatory factor analysis. Factor structure matrixes based on the field test data were estimated for each of the predictor domains. Once the scale scores have been created and their reliabilities estimated, we will proceed to confirmatory factor analysis. We anticipate a hierarchy of predictor factors. At the most detailed level the factors will include:

   Cognitive Factors - Verbal (from ASVAB), Quantitative, Technical Knowledge, Speed, Visualization

   Psychomotor Factors - Reaction Time, One-Hand Tracking, Two-Hand Tracking, etc.

   Non-cognitive Factors - Surgency, etc.

   LISREL runs will be examined to determine the goodness of fit of hypothesized factor structures. In addition to testing an a priori model, we will look for potential simplifications (loadings not significantly different from zero) and also consider potential improvements in fit by adding additional dimensions or additional loadings on the existing dimensions. In considering modifications to the original model, interpretability will be given somewhat more weight than empirical statistics.

4. Predictor factor (construct) scores will be estimated with a least squares procedure. The chief alternatives would be weighted sums (means) on the one hand and a more complex maximum likelihood or multistage least squares approach to factor score estimation.

5. MOS differences in the predictor constructs will be examined. The purpose of these analyses will be to see what kind of applicants go into (and remain in) the different MOS. Particular attention will be focused on the interest and temperament measures. This work will be necessarily exploratory since only concurrent data will be available for most measures. Special analyses with the Preliminary Battery MOS will attempt to determine whether the same construct differences existed prior to training.

6. Subgroup analyses will be run separately for each of the MOS with at least 100 in each of the different race or gender groups. At this stage, we are just looking for mean differences in the predictor scores, not for differences in regression slopes.

Criterion Score Analyses

After data preparation has been completed, the objectives for the criterion analyses are to identify an array of basic criterion variables (i.e., scores), investigate the latent structure of those variables, and determine the criterion construct scores. The following steps will be taken:

1. A final set of items by a priori scale analyses will be used to identify faulty or misplaced items. At this point the number of criterion variables will still be too large to enter into an

intercorrelation matrix. For example, the job knowledge test will still contain 30 task scores, and the number of individual rating scales will still be quite large.

2. A more manageable set of basic criterion scores will be obtained by factoring/clustering rating scales, hands-on test steps, and knowledge test items. In general, exploratory factor analysis will be used to reduce the individual rating scales to clusters of scales that will be averaged. For example, analyses of the field test data suggested that it might be reasonable to group the 11 Army-wide BARS scales into two or three clusters. For the hands-on and knowledge tests, items will be clustered via expert judgment sorts.

3. After step 2 yields a basic array of criterion scores, an intercorrelation matrix will be calculated for each MOS. Exploratory factor analyses will be used to generate hypotheses about the latent structure of the criterion space.

4. The "theories" about the criterion space generated in step 3 will be subjected to confirmatory analyses in an attempt to make a reasonable choice about the best-fitting model for the total domain of job performance for each MOS.

5. After the variables that comprise each criterion factor (construct) are identified, factor scores will be generated in a similar fashion as for the predictors.

Predictor/Criterion Interrelationships

After the above analyses have been carried out, the basic variables and the best-fitting model for both the predictors and the performance measures will have been identified. They can be compared to the "best guess" that was offered at the conclusion of the field tests. They also provide the variables to be used for establishing the selection/classification validity of the new predictor battery and for determining differential validity across criterion constructs, across jobs, and across subgroups. The basic picture and the analysis steps are summarized in Figures IV.5 and IV.6.

The validity analyses will begin by regressing the predictor battery against each criterion factor on each MOS. Within MOS we will then proceed to determine how much information is lost by reducing the number of prediction equations from K, the number of criterion factors, to just one equation. The predictive accuracy of the composite equation can then be compared to an equation developed for the composite criterion measure for each MOS. The method for obtaining the criterion composite score is described below.

The generalizability across MOS of the prediction equation for each criterion factor will be determined by using the predictor weights developed for one MOS to compute predicted scores for the same criterion factor (e.g., peer leadership and support) in each of the other MOS. The loss in predictive information as the number of equations is reduced from 19 to 1 will be determined.

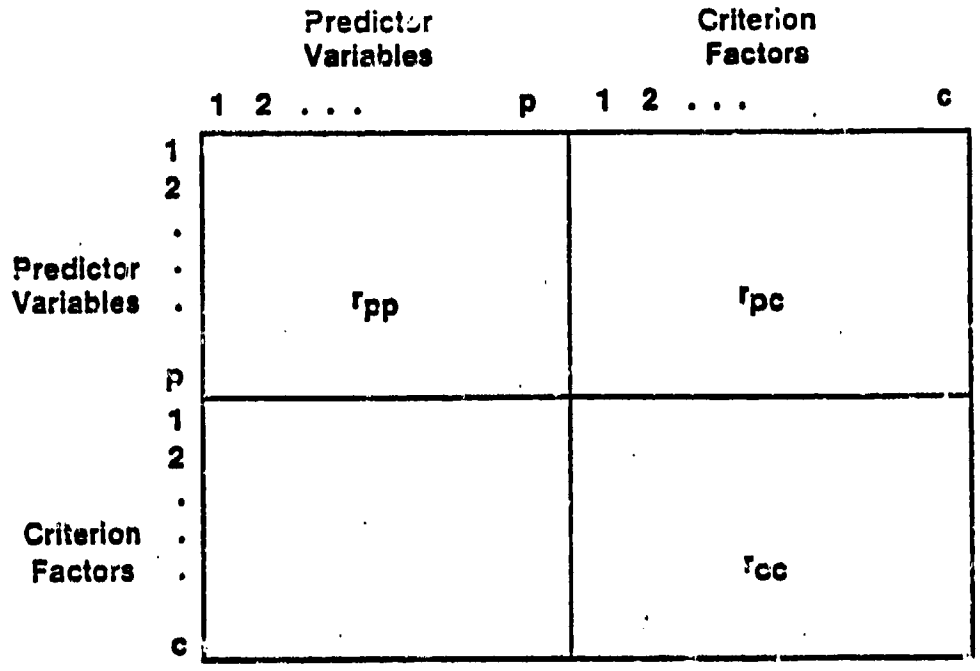


Figure IV.5. Predictor variable/criterion factor matrix.

A. Within MOS.
   1. Compute "best" prediction equation for each criterion factor.
   2. Compute best prediction equation for overall composite score.
   3. Determine loss of predictability as number of equations is reduced from (1) to (2).
   4. Determine incremental validity of Project A measures (over ASVAB).

B. Between MOS.
   1. Determine generalizability of each performance factor's prediction equation across MOS.
   2. Determine generalizability of composite prediction equations across MOS.
   3. Determine generalizability of incremental validity across MOS.

C. Differential Prediction Across Gender/Minority Groups (within selected MOS).
   1. Differential prediction for criterion specific equations.
   2. Differential prediction for criterion composite equation.

D. Estimation of Classification Efficiency.
   1. Brogden/Horst approximation.
   2. Project A approximation.

Figure IV.6. Analysis plan for predictor variable x criterion factor matrix.

For each major prediction equation the degree of differential prediction across racial and gender groups will be determined for MOS in which sample sizes are sufficient. For equations that produce significant differential prediction, the predictors that seem to be the source will be isolated and the validity of the equation recomputed with those variables omitted.

Also, for each major prediction equation the incremental validity of the Trial Battery compared to ASVAB will be computed. For the first analyses the degree of incremental validity will be expressed in terms of variance accounted for (ΔR^2). Additional indexes will be presented in the form of expectancy charts where specific base rates on the criterion are used to compare the percentage of correct predictions using ASVAB and ASVAB plus the Trial Battery.

Criterion Composite Scores

At the operational level, a selection and classification system makes two fundamental decisions about individuals: to select or not select into the organization, and, if selected, to choose the job (MOS) to which the individual will be assigned. For the Army these have the form of two sequential single-choice decisions. Developing the appropriate decision rules and estimating the overall validity, or accuracy, with which each decision is made requires a single algorithm. To compute a single validity estimate, the individual pieces of validity information must be aggregated. One straightforward way of doing this is to compute the decision rules and validity estimates using a single aggregate or composite measure of job performance.

For each MOS an overall criterion composite score will be computed for each individual in the following way. The latent structure, or criterion model, which receives the strongest support in the confirmatory analysis will be used to designate the specific measures that will be aggregrated to obtain factor or construct scores. Since the observed scores that enter the confirmatory analysis will already have been through a considerable amount of item and scale aggregation, the most straightforward scoring procedure would be to use an unweighted sum of standard scores to obtain a construct score.

Once the latent structures, or criterion constructs, for each MOS have been identified and defined, an overall composite will be obtained by expert weighting. To accomplish this, a series of workshops will be held with NCOs and company officers from each MOS, and the NCOs and officers will act as SMEs to scale the relative importance of each criterion construct for overall performance. Once the construct weights are obtained, they can be used to generate a weighted composite score.

During FY86, several different scaling methods for obtaining importance weights will be tried out in three exploratory workshops. The method judged to be the most feasible and most informative will be used in the final criterion weighting workshops which will begin in the summer of 1986. The judgment methods being explored are ratio estimation, paired comparisons, and conjoint measurement. Methods will be compared in terms of their ease of use, perceived validity, acceptability to the Army, and interjudge agreement.

After the importance weights for the performance factors are obtained, the composite scores for each MOS can be computed and used to derive a single prediction equation for each MOS. An additional issue is the effect of scenario (e.g., European conflict vs. peacetime) on the construct weights. Scenario effects will also be explored in the preliminary workshops.

Performance Utility

A full determination of classification validity or efficiency requires a common performance or criterion metric across all jobs (MOS). If the analysis goal is to rank order the validity, or efficiency, of alternative classification systems, then the precise meaning of the common metric is not crucial. However, if an organization wants to evaluate the cost vs. benefit of a new system, then some sort of utility metric is necessary.

In the private sector, great deference is paid to the dollar metric as the appropriate utility scale. If only the costs and benefits of personnel programs could be portrayed in dollar terms, then decisions among alternatives would be straightforward, or so it is hoped. Attempts to deal with the dollar metric in personnel research is well documented (e.g., Boudreau, 1983; Cascio, 1982; Hakel, 1986; Hunter & Schmidt, 1982) and need not be repeated here.

The general procedure that Project A has used to approach this problem is summarized in Sadacca and Campbell (1985). We began by exploring the problem in eight workshops with 8-12 field grade Army officers in each workshop. Each workshop incorporated both a general discussion of the "utility of performance" issue and a tryout of one or more potential scaling methods with which to scale the utility of MOS x performance level combinations (e.g., an MOS 11B Infantryman who performs at the 70th percentile). Five performance levels were defined as simply the 10th percentile, 30th percentile, 50th percentile, 70th percentile, and 90th percentile performer. The findings and conclusions from this first series of workshops are reported in Sadacca and Campbell. The more important conclusions were as follows:

1. For an organization such as the Army, expressing the utility of performance in dollars makes little conceptual sense. The Army is not in business to sell a product or service. Its job is to be in a high state of readiness so as to be able to respond to external threats of an unpredictable nature.

2. When using any one of several techniques (e.g., ratio estimation, paired comparisons), Army officers can provide reliable and consistent utility judgments. There does seem to be a commonly held value system about the relative utility of specific MOS x performance level combinations.

3. When participants were asked to describe the performance behaviors of individuals at the different performance levels, the model descriptions were very similar to the behavioral anchors of the BARS scales described in Part III. Again, at some general level, there does seem to be more common understanding of what performers at different levels are like.

During FY85 and on into FY86, an additional series of five utility scaling workshops were conducted. This second series examined specific scaling issues and evaluated alternative scaling methods. Workshop participants were asked to judge the difficulty of using each method and to state their perception of its validity. Methods are being compared in terms of their time demands on the judges, the amount of information they provide, and the degree of interjudge agreement.

At the conclusion of FY85, the following utility scaling methods were still in the process of being compared and evaluated.

- Card sort - A sample of MOS x performance level combination was sorted into 7-10 piles such that there were equal-appearing intervals between piles on a scale defined as the priority for filling force strength requirements. A variant of this procedure was to designate one of the piles as having zero utility for the Army. Cards sorted into piles below that point represented combinations with negative utility.

- Scaling against a standard - This was a ratio estimation technique in which a particular MOS x performance level combination was assigned a utility of 100 points (e.g., a 90th percentile 11B) and the remaining combinations were assigned points proportionately.

- Paired comparisons (100 point distribution) - The relative utility of MOS x performance level combinations presented in pairs was judged by dividing 100 points between each pair such that the difference represented the difference in relative potential utility to the Army.

- Paired comparisons (Equivalent Manning levels) - For each MOS x performance level combination, the judges were asked to estimate the number of individuals of one combination that it would take to equal some given number of the second combination (e.g., X number of 11Bs at the 50th percentile would be equivalent to Y number of 71Ls at the 30th percentile). The MOS x performance level combination pairs were judged independently.

The exploratory workshops will be concluded during FY86. After that the evaluation information will be analyzed and the two most promising methods will be used to do the actual scaling of MOS performance-level utility during FY86 and FY87. The next steps will be a full-scale proponent review of the utility scaling results and a series of Monte Carlo studies to determine the effects of different ranges of utility scale values on personnel assignments.

Three crucial issues must be resolved during FY86 and FY87:

- First, the precise nature of the desired metric must be incorporated in the scaling directions. For example, if ratio estimation against a standard is used, then the metric could be in terms of gains or losses of standard equivalents (e.g., "The gain from using the system is equivalent to having 3,000 more than 50th percentile 11Bs in the force.").

- Second, the issue of average vs. marginal utility must be addressed. That is, does the utility of a specific MOS/performance level combination change as more and more personnel are added to the enlisted force?

- Since the answer seems certain to be yes, an appropriate scenario must be developed that will allow scale values to be determined in a relatively straightforward manner.

EPILOGUE

This volume has presented the first 3 years of work on the Army Selection and Classification Project. At this point a full array of selection/classification tests and new measures of training and job performance have been developed. Also, all the newly developed measures have been administered to a sample of nearly 10,000 job incumbents in a Concurrent Validation design. This may very well be the richest single data set ever generated in personnel psychology, and more is yet to come. What is perhaps even more startling is that the project reached this point on schedule and with the original research plan intact. It has done what it set out to do with no compromises in the original objectives, so far.

It is also true that the work to date has been largely of a developmental nature. A great deal of time and energy was poured into the painstaking development of multiple instruments for multiple jobs and into the planning and execution of the data collection procedures. All this has been done under virtually continual evaluation by several review bodies. However, it is now time for the fun part. It is the time to analyze the data, to determine what more we can learn about performance and its antecedents, and to plan for the most significant data collection of all, the longitudinal validation. To falter now would be a disaster, both for the operational needs of the Army and for the benefit of the discipline.

REFERENCES

Alley, W. E., & Matthews, M. D. (1982). The vocational interest career examination. Journal of Psychology, 112, 169-193.

Anderson, J. W. (1984a). Heroism - a review of the literature and an operational definition (Working Paper 84-8). Alexandria, VA: U.S. Army Research Institute.

Anderson, J. W. (1984b). The warrior spirit (Working Paper 84-9). Alexandria, VA: U.S. Army Research Institute.

Anderson, J. W. (1984c). Warrior spirit II (Working Paper 84-10). Alexandria, VA: U.S. Army Research Institute.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.

Bentler, P. M. (1980). Multivariate analysis. Annual Review of Psychology, 30.

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. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 412-421.

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1987a). Development of a model of soldier effectiveness (ARI Technical Report 741). In preparation.

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1987b). Development of a model of soldier effectiveness: Retranslation Materials and Results (ARI Research Note 87-29). AD A181 832

Borman, W. C., Rosse, R. L., Abrahams, N. M., & Toquam, J. L. (1979). Investigating personality and vocational interest constructs and their relationships with Navy recruiter performance. Minneapolis, MN: Personnel Decisions Research Institute.

Borman, W. C., White, L. A., & Gast, I. F. (1985, August). Performance ratings as criteria: What is being measured? Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Boudreau, J. W. (1983). Effects of employee flows on utility analysis of human resources productivity improvement programs. Journal of Applied Psychology, 68, 396-406.

Bowers, D. G., & Seashore, S. E. (1966). Predicting organizational effectiveness with a four factor theory of leadership. Administrative Science Quarterly, 11, 238-263.

Bownas, D. A., & Heckman, R. W. (1976). Job analysis of the entry-level firefighter position. Minneapolis, MN: Personnel Decisions Research Institute, Inc.

Brown, E. M. (1968). Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 52, 195-199.

Brown, F. L., & Jacobs, T. O. (1970). Developing the critical combat performance required of the infantry platoon leader (HumRRO TR 70-5). Alexandria, VA: Human Resources Research Organization.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986a). Development and field test of task-based MOS-specific criterion measures (ARI Technical Report 717). In preparation.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1987). Appendixes to ARI Technical Report 717: Development and field test of task-based MOS-specific criterion measures (ARI Research Note in preparation).

Volume 1	Appendixes A-F
Volume 2	Appendix G, Part 1 (Batch A MOS)
Volume 3	Appendix G, Part 2 (Batch B MOS)
Volume 4	Appendix H, Part 1 (MOS 13B, 64C, 71L, 95B)
Volume 5	Appendix H, Part 2 (MOS 11B, 19E, 31C, 63B, 91A)
Volume 6	Appendixes I-U
Volume 7	Appendix V, Part 1 (MOS 13B, 64C)
Volume 8	Appendix V, Part 2 (MOS 71L, 95B)
Volume 9	Appendix V, Part 3 (MOS 11B, 19E, 31C)
Volume 10	Appendix V, Part 4 (MOS 63B, 91A)

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.

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

Cascio, W. F. (1982). Costing human resources: The financial impact of behavior in organizations. Boston: Kent.

Davis, R. H., Davis, G. A., Joyner, J. N., & de Vera, M. V. (1987a). Development and field test of job-relevant knowledge tests for selected MOS (ARI Technical Report 757). In preparation.

Davis, R. H., Davis, G. A., Joyner, J. N., & de Vera, M. V. (1987b). Development and field test of job-relevant knowledge tests for selected MOS: Appendixes to ARI Technical Report 757 (ARI Research Note in preparation).

Volume 1 Appendix A
Appendix B, Part 1 (MOS 11B - 54E)
Volume 2 Appendix B, Part 2 (MOS 55B - 95B)

Eaton, N. K., & Goer, M. H. (Eds.) (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Technical appendix to the Annual Report (ARI Research Note 83-37). AD A137 117

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (Eds.) (1984). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1984 fiscal year (ARI Technical Report 660). AD A178 944

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Fiedler, F. E., & Anderson, J. W. (1983). Hari kari, kamikaze, and other indications of commitment. Unpublished manuscript.

Flanagan, J. C. (1965). Flanagan Industrial Test Manual. Chicago: Science Research Associates.

Fleishman, E. A. (1967). Performance assessment based on an empirically derived task taxonomy. Human Factors, 9, 1017-1032.

Frost, D. E., Fiedler, F. E., & Anderson, J. W. (1983). The role of personal risk taking in effective leadership. Human Relations, 36, 185-202.

Gibson, J. J. (Eds.) (1947). Motion picture testing and research. Army Air Forces Aviation Psychology Research Program Reports, 7. Washington, DC: Government Printing Office.

Gough, H. G. (1975). Manual for the California Psychological Inventory. Palo Alto, CA: Consulting Psychologists Press.

Guilford, J. P., & Lacy, J. I. (Eds.) (1947). Printed classification tests. Army Air Forces Aviation Psychology Research Program Reports, 5. Washington, DC: Government Printing Office.

Guion, R. M. (1977). Content validity--The source of my discontent. Applied Psychology Measurement, 1 (winter), 1-10.

Hakel, M. D. (1986). Personnel selection and placement. Annual Review of Psychology.

Henriksen, K. F., Jones, D. R., Jr., Hannaman, D. L., Wylie, P. B., Shriver, E. L., Hamill, B. W., & Sulzen, R. H. (1980). Identification of combat unit leader skills and leader-group interaction processes (ARI Technical Report 440). AD A084 977

Holland, J. L. (1966). The psychology of vocational choice. Waltham, MA: Blaisdell.

Hollander, E. P. (1954). Buddy ratings: Military research and industrial implications. Personnel Psychology, 7, 385-393.

Hollander, E. P. (1965). Validity of peer nominations in predicting a distinct performance criterion. Journal of Applied Psychology, 49, 434-438.

Hough, L. M. (1984a). Development and evaluation of the "Accomplishment Record" method of selecting and promoting professionals. Journal of Applied Psychology, 69, 135-146.

Hough, L. M. (1984b). Identification and development of temperament and interest constructs and inventories for predicting job performance of Army enlisted personnel. Minneapolis, MN: Personnel Decisions Research Institute.

Hough, L. M. (Ed.) (1985). Literature review: Utility of temperament, biodata, and interest assessment for predicting job performance (ARI Research Note in preparation).

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J. S., & Peterson, N. C. (1984, August). Covariance analyses of cognitive and non-cognitive measures of Army recruits: An initial sample of Preliminary Battery Data. Paper presented at the Annual Convention of the American Psychological Association, Toronto. (In ARI Technical Report 660). AD A178 944

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983a). Improving the Selection, Classification and Utilization of Army Enlisted Personnel. Project A: Research Plan (ARI Research Report 1332). AD A129 728

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1983b). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report (ARI Research Report 1347). AD A141 807

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1984a). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report synopsis, 1984 fiscal year (ARI Research Report 1393). AD A173 624

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1984b). Improving the selection, classification, and utilization of Army enlisted personnel: Appendixes to the Annual Report, 1984 fiscal year (ARI Research Note 85-14).

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute. (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1985 fiscal year - Supplement to ARI Technical Report 746 (ARI Research Note in preparation).

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. Dunnette & E. Fleishman (Eds.), Human performance and productivity (Vol. 1). Hillsdale, NJ: Erlbaum.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process approach. In B. M. Staw & L. L. Cummings (Eds.), Research in organization behavior (Vol. 5), 141-197.

Jackson, D. N. (1967). Personality Research Form Manual. Goshen, NY: Research Psychologists Press.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills, CA: Sage.

Jensen, A. R. (1982). Reaction time and psychometric g. In M. J. Eysenck (Ed.), A model for intelligence, Springer-Verlag.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factor structure of the Army Services Vocational Aptitude Battery (ASVAB), Forms 8, 9, and 10: 1981 Army applicant sample. Educational and Psychological Measurement, 43, 1077-1088.

Kelley, C. R. (1969). The measurement of tracking proficiency. Human Factors, 11, 43-64.

Kern, R. P. (1966). A conceptual model of behavior under stress, with implications for combat training (HumRRO TR 66-12). Alexandria, VA: Human Resources Research Organization.

Keyes, M. A. (1985, in press). A review of the relationship between reaction time and mental ability. Minneapolis, MN: Personnel Decisions Research Institute.

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

Landy, F. J., & Farr, J. L. (1983). The measurement of work performance: Methods, theory, and application. New York: Academic Press.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

McHenry, J. J., & Rose, S. R. (1985). Literature review: Validity and potential usefulness of psychomotor ability tests for personnel selection and classification (ARI Research Note in preparation).

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

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY1981 and FY1982 enlisted accessions (ARI Technical Report 651). AD A156 807

Melton, A. W. (Ed.) (1947). Apparatus tests. Army Air Forces Aviation Psychology Program Research Reports, 4. Washington, DC: U.S. Government Printing Office.

Olson, D. M., Borman, W. C., Robertson, L., & Rose, S. R. (1984, August). Relationships between scales on an Army Work Environment Questionnaire and measures of performance. Paper presented at the Annual Convention of the American Psychological Association, Toronto. (In ARI Technical Report 600). AD A140 220

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. Journal of Applied Psychology Monographs, 64, 569-607.

Peterson, N. (Ed.) (1987a). Development and field test of the Trial Battery for Project A (ARI Technical Report 739). In preparation.

Peterson, N. (Ed.) (1987b). Test appendixes to ARI Technical Report 739: Development and field test of the Trial Battery for Project A (ARI Research Note 87-24). AD B113 802L

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

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. J., & 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: Validity analyses (Technical Report No. 4). Atlanta, GA: Life Office Management Association.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating formats. Organizational Behavior and Human Decisions Processes. In preparation.

Pulakos, E. D., & Borman, W. C. (Eds.) (1986). Development and field test of Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716). AD B112 857

Pulakos, E. D., & Borman, W. C. (Eds.) (1987b) Development and field test of Army-wide rating scales and the rater orientation and training program: Appendixes to ARI Technical Report 716 (ARI Research Note 87-22). In preparation.

Riegelhaupt, B. (1985). Army-wide administrative measures. Unpublished manuscript, Human Resources Research Organization.

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1987). The development of administrative measures as indicators of soldier effectiveness (ARI Technical Report 754). In preparation.

Riegelhaupt, B., & Sadacca, R. (1985). Development of Combat Prediction Scales. Unpublished manuscript, Human Resources Research Organization.

Rosse, R. L., Borman, W. C., Campbell, C. H., & Osborn, W. C. (1983, October). Grouping Army occupational specialties by judged similarity. Paper presented to the Military Testing Association. (In ARI Research Note 83-37). AD A137 117

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. Los Angeles: Psychological Services, Inc.

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985, August). Comparing work sample and job knowledge measures. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles. (In ARI Research Note in preparation).

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubleshooting empirical findings. Journal of Applied Psychology, 67, 401-410.

Sadacca, R., & Campbell, J. P. (1985, March). Assessing the utility of a personnel/classification system. Paper presented at the meeting of the Southeastern Psychological Association, Atlanta. (In ARI Research Note in preparation).

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., & Kaplan, L. B. (1971). Composite vs. multiple criterion: A review and resolution of the controversy. Personnel Psychology, 24, 419-434.

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

Sterling, B. (1984, March-April). Predictors of combat performance. Soldier Support Journal, 7-9.

Sternberg, S. (1966). High speed scanning in human memory. Science, 153, 652-654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. Acta Psychologica, 30, 276-315.

Tellegen, A. (1982). Brief manual for the Differential Personality Questionnaire. Unpublished manuscript, University of Minnesota.

Thorson, G., Hochhaus, L., & Stanners, R. F. (1976). Temporal changes in visual and acoustic codes in a letter-matching test. Perception and Psychophysics, 19, 346-348.

Toquam, J. L., Corpe, V. A., Dunnette, M. D., & Keyes, M. A. (1985). Literature review: Cognitive abilities--theory, history, and validity (ARI Research Note in preparation).

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1985a). Development and field test of behaviorally anchored rating scales for nine MOS (ARI Technical Report in preparation).

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1985b). Development and field test of behaviorally anchored rating scales for nine MOS: Appendixes to ARI Technical Report (ARI Research Note in preparation).

Volume 1 Appendixes A-B (MOS 13B, 64C)
Volume 2 Appendixes C-D (MOS 71L, 95B)
Volume 3 Appendixes E-G (MOS 11B, 19E, 31C)
Volume 4 Appendixes H-I (MOS 63B, 91A)

White, L. A., Gast, I. F., Sperling, H. K., & Rumsey, M. G. (1984, November). Influence of soldiers' experiences with supervisors on performance during the first tour. Paper presented at the meeting of the Military Testing Association, Munich, Germany. (In ARI Research Note in preparation).

Wing, H., Peterson, N. G., & Hoffman, R. E. (1984, August). Expert judgments of predictor-criterion validity relationships. Paper presented at the Annual Convention of the American Psychological Association, Toronto. (In ARI Technical Report 660). AD A178 944

Zedeck, S., & Cascio, W. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67, 752-758.

APPENDIX A

PROJECT A FY85 TECHNICAL PAPERS

A number of technical papers dealing with specialized aspects of Project A were prepared during Fiscal Year 1985. These papers are available in an ARI Research Note (in preparation), Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Annual Report, 1985 Fiscal Year - Supplement to ARI Technical Report 746. The following papers are included in the Research Note:

Borman, W. C. (1985). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. Manuscript submitted for publication.

Borman, W. C., White, L. A., & Gast, I. F. (1985, August). Performance rating as criteria: What is being measured? Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

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

Eaton, N. K. (1985, August). Measurement of entry-level job performance. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Hough, L. M., Barge, B. N., Houston, J. S., McGue, M. K., & Kamp, J. D. (1985, August). Problems, issues, and results in the development of temperament, biographical, and interest measures. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

McHenry, J. J., & McGue, M. K. (1985, August). Problems, issues, and results in the development of computerized psychomotor measures. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

McLaughlin, D. H. (1985, August). Measurement of test battery value for selection and classification. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Peterson, N. G. (1985, August). Overall strategy and methods for expanding the measured predictor space. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Rosse, R. L., & Peterson, N. (1985, August). Advantages and problems with using portable computers for personnel measurement. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Rossmeissl, P. G., & Brandt, D. A. (1985, August). Modeling the selection process to adjust for restriction in range. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985, August). Comparing work sample and job knowledge measures. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Sadacca, R., & Campbell, J. P. (1985, March). Assessing the utility of a personnel/classification system. Paper presented at the meeting of the Southeastern Psychological Association, Atlanta.

Toquam, J. L., Dunnette, M. D., Corpe, V., McHenry, J. J., Keyes, M. A., McGue, M. K., Houston, J. S., Russell, T. L., & Hansen, M. A. (1985, August). Development of cognitive/perceptual measures: Supplementing the ASVAB. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Walker, C. B. (1984, November). Validation of the Army's Military Applicant Profile (MAP) against an expanded criterion space. Paper presented at the meeting of the Military Testing Association, Munich, Germany.

Walker, C. B. (1985, February). The fakability of the Army's Military Applicant Profile (MAP). Paper presented at the Combined National and Western Region Meeting of the Association of Human Resources Management and Organizational Behavior, Denver.

White, L. A., Gast, I. F., Sperling, H. M., & Rumsey, M. G. (1984, November). Influence of soldiers' experiences with supervisors on performance during the first tour. Paper presented at the meeting of the Military Testing Association, Munich, Germany.

Wing, H. (1985, August). Expanding the measurement of predictor space for military enlisted jobs. Symposium presented at the Annual Convention of the American Psychological Association, Los Angeles.

Wise, L. L., & Mitchell, K. J. (1985, August). Development of an index of maximum validity increment for new predictor measures. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.