ARI Research Note 96-73

Building and Retaining the Career Force: New Procedures for Accessing and Assigning Army Enlisted Personnel--Final Report

John P. Campbell and Lola M. Zook, Editors
Human Resources Research Organization

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

August 1996

19970210 137

United States Army
Research Institute for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

DTIC QUALITY INSPECTED 1

DISCLAIMER NOTICE

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

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
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Henry Busciglio
Elizabeth J. Brady
Alan F. Drisko
Frances C. Grafton
Peter Greenston
Fred Mael
Abraham Nelson
Dale R. Palmer
Jay Silva
Clinton Walker
Mark C. Young

NOTICES

DISTRIBUTION: This report has been cleared for release to the Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or the National Technical Information Service (NTIS).

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 views, opinions, and findings in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED (from. . . to)
1996, August	Final	July 1989-December 1994

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Building and Retaining the Career Force: New Procedures for Accessing and Assigning Army Enlisted Personnel--Final Report	MDA903-89-C-0202
	5b. PROGRAM ELEMENT NUMBER 0603007A
6. AUTHOR(S)	5c. PROJECT NUMBER A792
John P. Campbell and Lola M. Zook (HumRRO)	5d. TASK NUMBER 1222
	5e. WORK UNIT NUMBER C01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Human Resources Research Organization (HumRRO) 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: PERI-RS 5001 Eisenhower Avenue Alexandria, VA 22333-5600	ARI
	11. MONITOR REPORT NUMBER Research Note 96-73

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Prepared under Project Building the Career Force (Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, U.S. Army Research Institute) COR: Michael Rumsey

14. ABSTRACT *(Maximum 200 words)*:

The Career Force research project is the second phase of a two-phase Army program to develop a selection and classification system for enlisted personnel, based on expected future performance. In the first phase, Project A, a large and versatile database was collected from a representative sample of Military Occupational Specialties (MOS) and used to (a) validate the Armed Services Vocational Aptitude Battery (ASVAB) and (b) develop and validate new predictor and criterion measures representing the entire domain of potential measures. Building on this foundation, Career Force research has finished developing the selection/classification system and evaluating its effectiveness, with emphasis on assessing second-tour performance. The final year of the project was devoted to developing optimal test batteries for predicting first- and second-tour performance, attrition, and reenlistment prospects, and estimating gains that might be expected from their use. The present report summarizes the entire Project A/Career Force research program.

15. SUBJECT TERMS

Career Force Criterion measures Longitudinal validation Personnel classification
Personnel selection Predictor measures Project A Second-Tour Performance

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	455	

EDITORS' PREFACE

This is the final annual report for work completed as part of the Building the Career Force Project. It also constitutes the primary technical report of the work completed on several of the project's principal tasks. Consequently, it is a "stand alone" document for Fiscal Year 1994 and does not refer the reader to more detailed descriptions in supplementary reports for that period. The Career Force Project extends the major work on selection and classification of Army enlisted personnel that was completed as part of Project A, and this report summarizes the entire research program.

The Career Force Project included (1) a replication and extension of the Project A Experimental Predictor Battery validities for the selection and classification of first-tour enlisted personnel; (2) validation of the Experimental Battery against end-of-training performance; (3) validation of training performance as a predictor of first-tour job performance; (4) measurement of second-tour performance; (5) validation of the Armed Services Vocational Aptitude Battery (ASVAB), the Experimental Battery, Advanced Individual Training performance, and first-tour performance as predictors of second-tour performance; and (6) identification of the optimal predictor battery for selection and classification, given certain specific sets of goals and constraints.

The annual report for year one (FY 90) described the results of a series of analyses directed at basic score development for (1) the Experimental Predictor Battery, (2) the End-of-Training performance measures, and (3) the second-tour job performance measures that were administered to the second-tour Concurrent Validation sample (CVII). The performance data from this initial sample of second-tour junior noncommissioned officers (NCO) were also used to develop a latent structure model of second-tour performance. The model hypothesizes six basic components for NCO performance.

The annual report for year two (FY 91) dealt with the analysis of performance data from the Longitudinal Validation I (LVI) sample, which is a sample of approximately 10,000 first-tour incumbents who entered the Army during 1986/87. It is the second of the two major cohorts of enlisted personnel that make up the total Project A/Career Force Project database. The criterion score development, data editing, and performance modeling analyses were each described in turn. The remainder of the report presented the results of the basic Longitudinal Validation of the ASVAB and the Project A Experimental Predictor Battery against (1) training performance, (2) first-tour job performance, and (3) second-tour job performance (i.e., the second-tour performance factor scores developed during year one).

The third annual report, for FY 92, focused exclusively on the Longitudinal Validation Second-Tour (LVII) sample. This is a sample of approximately 1,500 individuals in the nine "Batch A" MOS (the primary MOS sample group used from the beginning of Project A) who had reenlisted and were 2-3 years into their second tour of

duty at the time the LVII measures of second-tour job performance were administered. The individuals in the sample had entered the Army in 1986/87 as part of the Project A Longitudinal Validation (LV) sample. The report describes (1) the data collection

procedure, (2) the editing of the data file, (3) the initial analyses of each instrument to develop the basic criterion scores, and (4) the development of a model of second-tour performance based on the LVII sample data. A major feature of the results is the great consistency in the covariance structure of the basic criterion scores across cohorts (CVII vs. LVII) and across organizational levels (LVI vs. LVII).

The fourth annual report, for FY 93, covered (1) a summary of the analyses done with the Experimental Battery to support the work of the Manpower Accession Policy Working Group as it considered future revisions to the ASVAB; (2) the basic validation analyses of ASVAB and ABLE against second-tour performance in the longitudinal sample (LVII), using the LVII performance model factors as criterion scores; (3) the degree to which correlations of performance with performance exhibit convergent and divergent validity across organizational levels (e.g., first-tour performance vs. second-tour performance) relative to the individual components of performance; (4) the results of modeling the predictors of first-tour attrition using event-history analysis; and (5) a description of results obtained with the Army Job Satisfaction Questionnaire.

This final report completes the work of Project A/Career Force. It deals with (1) estimating the validity of an optimal set of predictor batteries, (2) developing and evaluating a new procedure for estimating the gains from classification, (3) estimating the potential gains from classification using the full set of Project A/Career Force predictor information, (4) estimating the gains in the validity of prediction of future performance using information about current performance in combination with the full test battery, (5) estimating the degree of differential prediction across racial and gender subgroups for the optimal predictor batteries, and (6) estimating the potential gains in validity from the development of empirical scoring keys for the AVOICE, an interests inventory that was part of the predictor battery. Consequently, this report contains the final estimates of selection validity and classification efficiency using all the predictor information that the two projects produced. All estimates have been corrected for range restriction and criterion attenuation.

As was the case for previous years, the writing of this report was very much a collaborative effort by a lot of people. The primary authors for each chapter are indicated in the Contents and also on the first page of each chapter. The editors, and the management, are deeply appreciative of their contributions.

(1) The latent structure of performance remains much the same as soldiers move from training through first-tour jobs into second-tour supervisory duties. That is, the specifications of the performance factors that best fit the intercorrelations among the criterion measures were virtually the same for each of the three stages.

(2) Performance in training does predict performance on the first-tour job, and first-tour performance does predict second-tour NCO performance.

(3) The ASVAB is an excellent predictor of future performance for both first tour and second tour. The dimensions of performance that it predicts best are technical task performance and leadership.

(4) The Experimental Predictor Battery that was developed in Project A yielded basic predictor scores that are both reliable and construct valid. While cognitive abilities are the primary predictor of technical task performance, certain aspects of personality and interests contribute to a selection validity that is higher than for ASVAB alone. For example, leadership is predicted with considerable accuracy by the full equation.

(5) In terms of differential prediction and classification efficiency, the predictor battery produces gains from classification over selection. The question of how the potential gain that is possible would be preserved by various kinds of operational job assignment systems is a matter for further investigation.

(6) The components of the methodology used to evaluate selection and classification efficiency together constitute an analytic framework that can be used to evaluate selection validity versus classification efficiency for a variety of personnel selection and job assignment procedures and for a range of organizational constraints. The database itself will provide a comprehensive resource for studying selection/classification systems and personnel decision-making procedures for some time to come.

In addition to the findings and related information obtained in meeting the basic objectives of Project A/Career Force, the projects have yielded much supplementary information and a long list of products and techniques that can be used not only in other Army activities but also across the entire field of human resource management and of personnel research.

Utilization of Findings:

Involved as it was with almost the entire personnel system for entry-level and first-level supervisory jobs in the Army, this long-term research program has provided specific findings and comprehensive information that can be useful in a wide range of policy decisions and operational applications. The knowledge gained will provide guidance in selecting Army recruits in light of the changing circumstances faced by the military, and in continuing efforts to make the best possible assignment decisions both to match individuals with jobs and to meet the needs of the Army.

BUILDING AND RETAINING THE CAREER FORCE: NEW PROCEDURES FOR ACCESSING AND ASSIGNING ARMY ENLISTED PERSONNEL - FINAL REPORT

CONTENTS

Chapter		Page
1	A SUMMARY OF PROJECT A/CAREER FORCE (John P. Campbell and James H. Harris)	1
	CHARACTERISTICS OF THE PRESENT ARMY PERSONNEL SYSTEM	1
	Recruitment	2
	Selection and Classification at the MEPS	4
	Initial Classification	5
	Initial Training	6
	Performance Assessment in Army Units	7
	Reenlistment Screening	8
	Summary	8
	A BRIEF HISTORY OF SELECTION AND CLASSIFICATION	9
	THE GOALS AND DESIGN OF PROJECT A AND CAREER FORCE	10
	Specific Program Objectives	11
	The Research Samples	12
	Procedure and Design	14
	A SUMMARY OF PROJECT A	16
	Predictor Development in Project A	16
	Performance Measurement	18
	The Concurrent Validation	23
	Concurrent Validation Results	29
	Weighting Criterion Components	31
	Scaling the Utility of Individual Performance	32
	Second-Tour Performance Criterion Development	33
	The Longitudinal Validation Data Collection	35
	Summary	37
	The Foundation Provided by Project A	38
	Project A Products and Results	39
	A SUMMARY OF CAREER FORCE	41
	SUMMARY OF PROJECT EFFORTS FOR YEAR ONE	42
	Data Base Design	42
	Basic Scores for the Experimental Battery	42
	Basic Scores for the End-of-Training Measures	45
	Development of Second-Tour Performance Scores (CVII)	45
	Development of the CVII Second-Tour Performance Model	48
	SUMMARY OF PROJECT EFFORTS FOR YEAR TWO	50
	Objectives	50
	Development of Alternative ABLE Factor Composites	52
	Prediction of Performance in Training	55

CONTENTS

Page

Development of Basic Scores for the Longitudinal Validation (LVI) Performance Measures ... 59
The LVI Data File: Final Data Editing and Score Imputation ... 67
Development of the LVI First-Tour Performance Model ... 69
Basic Validation Results for the LVI Sample ... 73
Results of the Concurrent Sample Second-Tour Validation (CVII) ... 84
Prediction of Second-Tour Performance From the Trial Battery and From First-Tour Performance ... 91

SUMMARY OF PROJECT EFFORTS FOR YEAR THREE ... 94
Longitudinal Validation Second-Tour Data Collection ... 95
Development of Basic Scores for LVII Performance Measures ... 98
The LVII Data File: Extent of Missing Data ... 115
Development of the LVII Performance Model ... 115

SUMMARY OF PROJECT EFFORTS FOR YEAR FOUR ... 124
Identification of Optimal Predictor Batteries Using a Subset of the Project A/Career Force Experimental Predictor Battery ... 124
Basic Validation Results for the LVII Sample ... 132
Prediction of Future Performance From Current Performance and From Training Performance ... 139
Prediction of First-Term Military Attrition Using Pre-Enlistment Predictors ... 144
The Role of Job Satisfaction in Performance, Attrition, and Reenlistment ... 153

ORGANIZATION OF THE CURRENT REPORT ... 161

2 **ESTIMATING THE RELIABILITY OF THE SCORES ON PROJECT A/CAREER FORCE PERFORMANCE CRITERION FACTORS** ... 163
(Doug Reynolds, Anthony Bayless, and John P. Campbell)

ESTIMATING CRITERION RELIABILITIES ... 163
Research Samples ... 163
Reliability Computation ... 164
Results ... 167

TRUE SCORE CORRELATIONS OF PAST PERFORMANCE WITH FUTURE PERFORMANCE ... 167

3 **DEVELOPMENT AND EVALUATION OF AVOICE EMPIRICAL KEYS, SCALES, AND COMPOSITES** ... 177
(Cheryl Paullin, Ken Bruskiewicz, Mary Ann Hanson, Kristi Logan, and Mark Fellows)

BACKGROUND ... 177
AVOICE Development and Scoring ... 177
Empirical Scoring Procedures ... 178
Past Research on the Empirical Approach ... 179
Issues in Developing Empirical Scoring Procedures ... 180
Available Project A/Career Force AVOICE Data ... 182
Sex Bias/Fairness Issues ... 183

CONTENTS

Page

THE GENERAL PROCEDURE AND ANALYSIS DESIGN 186
Criterion Variables 186
Samples 187
Data Analysis Procedures 191

RESULTS: COMPARISONS OF SCORING PROCEDURES 195
Comparative Results for Keys vs. Scales vs. Composites 195
Comparative Results Related to Scale Length 197

RESULTS: VALIDITY ESTIMATES 199
Results for Empirical Scales and Composites Developed to Predict Organizationally Relevant Criteria 199
Content of the Empirical Scales and Composites 202
Transportability Results for Core Technical Proficiency 203
Comparison of Empirical CTP Scales/Composites With Rational Scoring Procedures 209
Comparison of Random and Empirical Keys 214
Overall Results for Occupational Keys and Scales 214
Content of the Occupational Scales 218
Male-Female Comparisons on the Occupational Scales 220
Comparison of CTP Empirical Scales and Occupational Scales 224

DISCUSSION 230
Psychometric Issues 230
Empirical Scales and Composites Developed to Predict Organizationally Relevant Criteria 231
Occupational Keys to Predict MOS Membership 234
Comparison of CTP Empirical and Occupational Scales 235
Comparison of Rational and Empirical Scales 235

CONCLUSIONS 236

4 **MAXIMIZING SELECTION VALIDITY FOR PREDICTING FIRST-TOUR PERFORMANCE** 239
(Scott H. Oppler, John P. Campbell, and Norman G. Peterson)

DIFFERENTIAL PREDICTION ACROSS CRITERION CONSTRUCTS AND ACROSS MOS 240
Sample 240
Measures 241
Analysis Procedures 241
Results 243

VALIDITY ESTIMATES FOR FULL AND REDUCED EQUATIONS 244
Analysis Procedures 245
Results 246

SUMMARY 251

CONTENTS

Page

5 **DIFFERENTIAL PREDICTION ACROSS RACIAL AND GENDER GROUPS: OPTIMAL BATTERY ANALYSES** 253
(Teresa L. Russell, Norman G. Peterson, Scott H. Oppler, and John P. Campbell)

PROCEDURES 254
Sample 254
Regression Model Analyses 254

RESULTS 257
Traditional Basic Scores 257
Optimal Selection and Optimal Classification Predictor Composites 259

CONCLUSIONS 259

6 **THE SEQUENTIAL PREDICTION OF INDIVIDUAL JOB PERFORMANCE: THE "ROLLUP"** 263
(Rodney A. McCloy)

METHOD 263
The Samples 263
Analysis Procedure 264

RESULTS 268
LVI Analyses 268
LVII Analyses 271
Summary 274

DISCUSSION AND SUMMARY CONCLUSIONS 274

7 **PERSONNEL CLASSIFICATION AND DIFFERENTIAL JOB ASSIGNMENTS: A REVIEW OF THE ISSUES AND ALTERNATIVE MODELS** 277
(John P. Campbell, Mark Fellows, and Paul Sticha)

THE COMPONENTS OF SELECTION AND CLASSIFICATION 277
The Goal of Selection/Classification 278
Selection Versus Classification 279

MULTIPLE REGRESSION VERSUS MULTIPLE DISCRIMINANT ANALYSIS 280
Differences Between the Two Approaches 281
Regression Models 283
Discriminant Models 285
Conclusions 286
Summary 290

CLASSIFICATION OBJECTIVES AND CONSTRAINTS 290

PREDICTOR SELECTION FOR CLASSIFICATION 292
The Horst Method 293
The Abrahams et al. Method 293

CONTENTS

Page

Evaluating All Possible Combinations ... 294

METHODS FOR ESTIMATING CLASSIFICATION GAINS ... 295
Brogden's Analytic Solution ... 295
Estimates of Gains From Simulations ... 297

ALTERNATIVE DIFFERENTIAL JOB ASSIGNMENT PROCEDURES ... 299
Current Methods ... 300
Summation ... 303

NEW METHODS FOR MAKING JOB ASSIGNMENTS ... 304
Enlisted Personnel Assignment System (EPAS) ... 304
The New PACE ... 306
Cost-Performance Tradeoff Model ... 307
Comparison of Models ... 308

SUMMARY ... 308

8 ESTIMATING CLASSIFICATION GAINS: DEVELOPMENT OF A NEW ANALYTIC METHOD ... 311
(Rodney L. Rosse, Norman G. Peterson, and Deborah Whetzel)

STATISTICAL BASIS OF eMPP AND reMAP ... 312
Situation ... 312
Relevant Parameters of the Applicant Population ... 313
Mean Values of Predicted Performance ... 314
Classification Efficiency Defined in Terms of Actual Performance:
Definition of the Re-estimate (reMAP) ... 315
Sample Re-estimation of Actual Performance ... 319
Theoretical Expected Job and Overall Mean Predicted Performance ... 320
Numerical Evaluation of the Estimator of Mean Predicted Performance ... 322

MONTE CARLO DEMONSTRATION OF THE ACCURACY OF eMPP AND reMAP ... 324
Defining the Simulated Population ... 324
Simulated Validation Studies ... 326
Numerical Evaluations for Estimated Mean Predicted Performance (eMPP) ... 327
Simulated Actual Assignment of 1,000 Applicants ... 328
Results of Three Monte Carlo Investigations ... 329
Advantages and Disadvantages of Practical Application of eMPP and reMAP Estimators ... 333

APPLICATIONS OF CLASSIFICATION EFFICIENCY ESTIMATES IN LVI ANALYSES ... 333
Weighting Systems ... 335
Predictor Variables ... 336
Criteria ... 337
Sample ... 337
Quota Conditions ... 339
Results ... 340

CONTENTS

Page

Classification Efficiency of Full vs. Reduced Prediction Equations 352
Concluding Comments 354

9 **THE FINAL CHAPTER** 357
(John P. Campbell and James H. Harris)

THE SPECIFIC GOALS AND DESIGN OF PROJECT A AND CAREER FORCE 357
Specific Program Objectives 358
The Research Samples 359
Procedure and Design 360

PREDICTOR DEVELOPMENT IN PROJECT A 362

PERFORMANCE CRITERION MEASUREMENT 371
Criterion Development: First Tour 371
Criterion Development: Second Tour 376
Training Performance Measurement 384
The Correlation of Performance With Performance 384

VALIDATION RESULTS 388
Basic Validation and Incremental Validity Comparisons for the Prediction of Training Performance (EOT), First-Tour Job Performance (LVI), and Second-Tour Job Performance (LVII) 390
Validity Estimates for Optimal Equations 394
The Reenlistment Decision: Optimal Prediction of Second-Tour Performance 396
Selection Validity and Classification Efficiency 400
Performance Component Criticality and the Utility of Performance 400

A FINAL SUMMARY 401
The Basic Objectives 402
Beyond the Objectives 403

A NOTE OF THANKS 405

References 407

APPENDICES
A Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency A-1
B Content of Empirical Scales and Composites Developed to Predict Leadership B-1
C Content of Empirical Scales and Composites Developed to Predict Attrition C-1
D Cross-Sex Transportability for Core Technical Proficiency E-1
E Content of Occupational Scales Developed to Predict MOS Membership D-1
F Least-Squares Regression Weights for Predictor Composites in Chapter 8 ..... E-1

CONTENTS

Page

List of Tables

1.1 The Army Selection, Classification, and Evaluation Process 2
1.2 Hierarchical Map of Predictor Space 19
1.3 Summary of Predictor Measures Used in Concurrent Validation (The Trial Battery) 20
1.4 Summary of Criterion Measures Used in Concurrent Validation Samples 24
1.5 Concurrent Validation Sample Soldiers by MOS by Location 25
1.6 Predictor Construct Scores 27
1.7 Latent Structure Scores 28
1.8 Mean Validity for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs 29
1.9 Mean Incremental Validity for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs 30
1.10 Description of Tests in Experimental Battery 36
1.11 ABLE Rational Composites and Corresponding Content Scales 52
1.12 Distribution of ABLE Scale Items on ABLE-168 and ABLE-114 Factor Composites 54
1.13 Mean of Multiple Correlations Computed Within Job for End-of-Training Sample for ASVAB Factors, Spatial, Computer, JOB, ABLE Rational Composites, and AVOICE 57
1.14 Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for End-of-Training Sample for Spatial, Computer, JOB, ABLE Rational Composites, and AVOICE 58
1.15 Measures Administered to Soldiers in LVI Sample 59
1.16 Comparison of LVI and CVI Army-Wide Factor Analysis Results: Pooled Peer/Supervisor Ratings 62
1.17 Composition and Definition of LVI Army-Wide Rating Composites 63
1.18 LVI Sample Sizes for Performance Measures for Batch A MOS 67
1.19 LVI Sample Sizes for Performance Measures for Batch Z MOS 67
1.20 LVI Combined Criteria Data: Percentage of Missing Data for Basic Scores by MOS 68
1.21 LVI Predictor Data: Amount of Missing Data for Paper-and-Pencil Scale Scores 70
1.22 LVI Predictor Data: Amount of Missing Data for Computer-Administered Scale Scores 71
1.23 Mapping of LVII Performance Measures Onto Latent Performance Factors 74
1.24 Mean Intercorrelations Among 13 Summary Criterion Scores for the Batch A MOS in the LVI Sample 75
1.25 Soldiers in CVI and LVI Data Sets With Complete Predictor and First-Tour Criterion Data by MOS 76
1.26 Mean of Multiple Correlations Computed Within Job for LVI Listwise Deletion Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE 78
1.27 Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for LVI Listwise Deletion Samples for Spatial, Computer, JOB, ABLE Composites, and AVOICE 79
1.28 Mean of Multiple Correlations Computed Within Job for LVI Setwise Deletion Samples for Spatial, Computer, JOB, ABLE Composites, and AVOICE 80
1.29 Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for LVI Setwise Deletion Samples for Spatial, Computer, JOB, ABLE Composites, and AVOICE 81

CONTENTS

		Page
1.30	Comparison of Mean Multiple Correlations Computed Within Job for LVI and CVI Listwise Deletion Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE	82
1.31	CVII Sample Sizes by MOS	85
1.32	Multiple Correlations for ASVAB Factors, ASVAB Subtests, ABLE Composites, and ABLE-114 Scores Against 19 CVII Criterion Variables (All MOS), With Unit Weights	87
1.33	Multiple Correlations for ASVAB Factors Plus ABLE Composites and Plus ABLE-114 Scores, and for ASVAB Subtests Plus ABLE Composites and Plus ABLE-114 Scores Against 19 CVII Criterion Variables, All MOS	88
1.34	Multiple Correlations for 10 Sets of Criterion Composite Weights, All MOS	90
1.35	Numbers of Soldiers With CVI and CVII Data by MOS: Initial Sample	91
1.36	Uncorrected Correlations Between CVI and CVII Raw Criterion Composites Computed Across MOS: Initial Sample	93
1.37	Correlations Between CVI Weighted Predictor Composites, CVI Criterion Composites, and CVII Criterion Composites for Raw Scores, Computed Across MOS: Initial Sample	94
1.38	LVII Data Collection Instruments	96
1.39	LVII Data Collection Schedule	98
1.40	Number of LVII Job Knowledge Tasks and Items by MOS	100
1.41	Number of LVII Hands-On Tasks and Steps by MOS	100
1.42	Comparison of LVII and CVII Army-Wide Factor Analysis Results: All Dimensions	103
1.43	Composition of LVII Army-Wide Rating Composites	104
1.44	Administrative Indices Descriptive Statistics for LVII and CVII	106
1.45	Situational Judgment Test: Definitions of Factor-Based Subscales	109
1.46	LVII Personal Counseling Exercise Scales and Factor Analysis Results	110
1.47	LVII Disciplinary Counseling Exercise Scales and Factor Analysis Results	112
1.48	LVII Training Exercise Scales and Factor Analysis Results	113
1.49	Number of LVII Soldiers With Complete or Partial Data by Criterion Instrument and MOS	116
1.50	Percent of LVII Assigned Values by Type of Instrument and MOS	116
1.51	Number of LVII Soldiers With Complete Array of Basic Criterion Scores (Excluding Combat Performance Prediction Scales) by MOS	117
1.52	Consideration/Initiating Structure Model	119
1.53	Leadership Factor Model	121
1.54	Summary of Results for Predicting Each Criterion at Each Time Interval	128
1.55	"Optimal" Test Batteries for Each Criterion and Time Interval According to the "Maximize Validity/Minimize Average Subgroup Difference" Rule	130
1.56	"Optimal" Test Batteries for Each Criterion and Time Interval According to the "Maximize Discriminant Validity/Minimize Average Subgroup Difference" Rule	131
1.57	Soldiers in LVII Sample Meeting Predictor/Criterion Setwise Deletion Data Requirements for Validation of ASVAB Operational Scores and Spatial, Computer, JOB, ABLE, and AVOICE Experimental Battery Predictor Composites Against Core Technical Proficiency by MOS	133
1.58	Mean of Multiple Correlations Computed Within Job for LVII Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE	135
1.59	Mean of Multiple Correlations Computed Within Job for ASVAB Factors Within Each LVII Predictor/Criterion Setwise Deletion Sample	136
1.60	Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for LVII Samples for Spatial, Computer, JOB, ABLE Composites, and AVOICE	137

CONTENTS

Page

1.61 Comparison of Mean Multiple Correlations Computed Within Job for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE Within LVI and LVII Samples 138
1.62 Zero-Order Correlations of Training Performance (EOT) Variables With First-Tour Job Performance (LVI) Variables: Weighted Average Across MOS 142
1.63 Zero-Order Correlations of First-Tour Job Performance (LVI) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS 143
1.64 Zero-Order Correlations of Training Performance (EOT) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS 144
1.65 AJSQ Score Intercorrelations 155
1.66 Mean Correlations Between Job Satisfaction and Performance Weighted by Sample Size 157
1.67 Intercorrelations Among Major Turnover Analysis Variables 158
1.68 Point-Biserial Correlations Between Reenlistment and Predicted Probability of Reenlistment for Three Models 159
1.69 Point-Biserial Correlations Between Attrition and Predicted Probability of Attrition for Four Models 161

2.1 Reliabilities for CVI Batch A MOS Criterion Composites 167
2.2 Reliabilities for LV-EOT Batch A MOS Criterion Composites 168
2.3 Reliabilities for LVI Batch A MOS Criterion Composites 168
2.4 Reliabilities for LVII Batch A MOS Criterion Composites 169
2.5 Median Reliabilities (Across Batch A MOS) for the LVT (EOT), LVI, and LVII Performance Factor Scores 169
2.6 Zero-Order Correlations of Training Performance (EOT) Variables With First-Tour Job Performance (LVI) Variables: Weighted Average Across MOS 174
2.7 Zero-Order Correlations of First-Tour Job Performance (LVI) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS 175
2.8 Zero-Order Correlations of Training Performance (EOT) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS 176

3.1 Content of the AVOICE 179
3.2 Number of Soldiers in the Concurrent and Longitudinal Validation Cohorts With AVOICE and First-Tour Performance Data by MOS and by Sex 184
3.3 AVOICE Rational Scales and Composites: Male/Female Effect Sizes for the Total LVI Sample 185
3.4 Sample Sizes for Developing and Cross-Validating Empirical Scoring Procedures to Predict Core Technical Proficiency, Leadership, and Attrition 189
3.5 Sample Sizes for Developing and Cross-Validating Occupational Keys, Scales, and Composites to Predict MOS Membership 190
3.6 Comparison of Empirical Keys, Scales, and Composites Developed to Predict Core Technical Proficiencies in the MOS 13B Development and Cross-Validation Samples 195
3.7 Comparison of Occupational Keys, Scales, and Composites Developed to Predict MOS 91A Membership for Males 196
3.8 Cross-Validation Sample Correlations Between Core Technical Proficiency and Empirical Scales of Various Lengths Developed to Predict CTP 198
3.9 Comparison of Occupational Scales of Various Lengths Developed to Predict MOS 91A Membership for Males 199

CONTENTS

		Page
3.10	Cross-Validities for Empirical Scales and Composites Developed to Predict First-Tour Core Technical Proficiency	200
3.11	Development and Cross-Validation Sample Validities for Empirical Scales and Composites Developed to Predict the Second-Tour Leadership Criterion Composite	201
3.12	Development and Cross-Validation Sample Validities for Empirical Scales and Composites Developed to Predict Attrition	202
3.13	Cross-Cohort Transportability of Empirical Scales and Composites Developed to Predict Core Technical Proficiency	204
3.14	Content Overlap Between Empirical Scales/Composites Developed to Predict Core Technical Proficiency in the CV and the LV Cohorts	206
3.15	Transportability Across MOS of Empirical Scales and Composites Developed to Predict Core Technical Proficiency	208
3.16	Male/Female Effect Sizes on Scales Developed to Predict Core Technical Proficiency in the LV Cohort	210
3.17	Comparison of Validity of Rational and Empirical Scales and Composites for Predicting First-Tour Core Technical Proficiency, Second-Tour Leadership, and Attrition	211
3.18	Comparison of Validity of Rational Composites and Empirical Scales for Predicting First-Tour Core Technical Proficiency Before and After Correcting for Range Restriction	212
3.19	Incremental Validity of AVOICE Rational Composites and Empirical Scales Over the ASVAB for Predicting First-Tour Core Technical Proficiency	213
3.20	Mean Occupational Key and Scale Scores in Cross-Validation (LVI) Samples	215
3.21	Mean 12-Item Occupational Scale Scores in the Cross-Validation and Cross-Cohort Samples	216
3.22	Mean All-Significant Occupational Scale Scores in the Cross-Validation and Cross-Cohort Samples	219
3.23	Same-Sex, Cross-Sex, and Combined-Sex Comparisons of 12-Item Occupational Scales Developed in Males Only, Females Only, and Males/Females Combined Samples	221
3.24	Occupational Scales: Effect Sizes for Mean Score by Sex in the LVI Sample	222
3.25	AVOICE Rational Scale Means for the LVI Sample by MOS and by Sex	223
3.26	AVOICE Rational Composite Means for the LVI Sample by MOS and by Sex	225
3.27	Comparison of Mean Scores on 12-Item Occupational Scales and 12-Item CTP Empirical Scales in LVI Sample	226
3.28	Comparison of Mean Scores on All-Significant Occupational Scales and All-Significant CTP Empirical Scales in the LVI Sample	227
3.29	Comparison of 12-Item Occupational and CTP Empirical Scale Correlations With Core Technical Proficiency in the LVI Sample	229
3.30	Comparison of All-Significant Occupational and CTP Empirical Scale Correlations With Core Technical Proficiency in the LVI Sample	229
3.31	Content Overlap Between Occupational Scales and Empirical Scales Developed to Predict Core Technical Proficiency	230
3.32	Summary of Results for Empirical Scales and Composites Developed to Predict First-Tour Core Technical Proficiency, Attrition, and Second-Tour Leadership	232
4.1	Estimates of Differential Prediction of Criterion Construct Scores for the ASVAB, Spatial, and Computerized Tests Predictor Set	243
4.2	Estimates of Differential Prediction for Criterion Construct Scores for the ASVAB, ABLE, and AVOICE Predictor Set	244

CONTENTS

		Page
4.3	SME Reduced (Optimal) Equations for Maximizing Selection Efficiency for Predicting Core Technical Proficiency in LVI	247
4.4	SME Reduced (Optimal) Equations for Maximizing Classification Efficiency for Predicting Core Technical Proficiency in LVI	248
4.5	SME Reduced (Optimal) Equations for Maximizing Selection Efficiency for Predicting Will-Do Criterion Factors	249
4.6	Estimates of Maximizing Selection Efficiency Aggregated Over MOS: Predicting Core Technical Proficiency	251
5.1	Subgroup Sample Sizes for Differential Prediction Analyses	255
5.2	Slope and Intercept Comparisons for Traditional Predictor Composites Against Core Technical Proficiency for White-Black and Male-Female Groups	258
5.3	Slope and Intercept Comparisons for Traditional Predictor Composites Against Will-Do Criteria for White-Black and Male-Female Groups	258
5.4	Slope and Intercept Comparisons for Optimal Selection and Classification Composites Against Core Technical Proficiency for White-Black and Male-Female Groups	260
5.5	Slope and Intercept Comparisons for Optimal Selection and Classification Composites Against Will-Do Criteria for White-Black and Male-Female Groups	260
6.1	Variables Used in the Rollup Analyses	265
6.2	Sample Size for Rollup Analyses by MOS	266
6.3	Rollup Analysis Regression Models	269
6.4	Rollup Validity Analyses: Multiple Correlations for Predicting First-Tour Job Performance (LVI) Criteria From ASVAB and Various Combinations of ASVAB, Selected Experimental Battery Predictors, and End-of-Training Performance Measures	270
6.5	Rollup Validity Analyses: Multiple Correlations for Predicting Second-Tour Job Performance (LVII) Criteria From ASVAB and Various Combinations of ASVAB, Selected Experimental Battery Predictors, and End-of-Training and First-Tour (LVI) Performance Measures	272
8.1	Means of Developmental Sample Statistics for Monte Carlo Investigation Using ASVAB Only	330
8.2	Means of Estimates of Mean Predicted and Mean Actual Performance (eMPP and reMAP) Compared to Simulated Results of Assigning 1,000 "Real" Applicants: First Investigation	330
8.3	Means of Developmental Sample Statistics for Monte Carlo Investigation Using ASVAB, Spatial, and Computer Tests	331
8.4	Means of Estimates of Mean Predicted and Mean Actual Performance (eMPP and reMAP) Compared to Simulated Results of Assigning 1,000 "Real" Applicants: Second Investigation	331
8.5	Means of Developmental Sample Statistics for Monte Carlo Investigation Using ASVAB, ABLE, and AVOICE Tests	332
8.6	Means of Estimates of Mean Predicted and Mean Actual Performance (eMPP and reMAP) Compared to Simulated Results of Assigning 1,000 "Real" Applicants: Third Investigation	332
8.7	Weights Used for Calculating Overall Performance Across Five Criteria	337
8.8	Proportion and Number of Soldiers Selected Into Nine Career Force MOS in Fiscal Year 1993	339

CONTENTS

Page

8.9 Proportion and Number of "Applicants" Assigned to Career Force MOS in Simulations 340
8.10 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB Only and Criterion = Core Technical Proficiency 341
8.11 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB Only and Criterion = Overall Performance 341
8.12 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Spatial and Criterion = Core Technical Proficiency 342
8.13 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Spatial and Criterion = Overall Performance 342
8.14 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Computer-Administered Psychomotor and Criterion = Core Technical Proficiency 343
8.15 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Computer-Administered Psychomotor and Criterion = Overall Performance 343
8.16 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + ABLE and Criterion = Core Technical Proficiency 344
8.17 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + ABLE and Criterion = Overall Performance 344
8.18 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + AVOICE and Criterion = Core Technical Proficiency 345
8.19 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + AVOICE and Criterion = Overall Performance 345
8.20 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + All Experimental Predictors and Criterion = Core Technical Proficiency 346

CONTENTS

Page

8.21 Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB - All Experimental Predictors and Criterion = Overall Performance 346

8.22 Minimum and Maximum Mean Standard Score Values of reMAP Among Six Predictor Sets and Two Criteria of reMAP Estimators by MOS 347

8.23 Values of Two Classification Indices Averaged Across Nine Army Jobs for Two Criteria, Two Assignment Strategies, Two Weighting Systems, and Six Predictor Composites 349

8.24 Values of Two Classification Indices Averaged Across Nine Army Jobs for the Core Technical Proficiency Criterion and Three Types of Prediction Equations 353

9.1 Hierarchical Map of Predictor Space 364

9.2 Sample Sizes From Each Batch A MOS When Performance was Assessed at Three Points in Time for the Project A/Career Force LV Sample 386

9.3 Zero-Order Correlations of Training Performance (EOT) Variables With First-Tour Job Performance (LVI) Variables: Weighted Average Across MOS 388

9.4 Zero-Order Correlations of First-Tour Job Performance (LVI) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS 389

9.5 Average Multiple Correlations Computed Within Job for EOT, LVI, and LVII Validation Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE, and AVOICE 392

9.6 Average Increments in Multiple Correlations over ASVAB Computed Within Job for EOT, LVI, and LVII Validation Samples for Spatial, Computer, JOB, ABLE, and AVOICE 393

9.7 Validity Estimates for Full and SME Reduced (Optimal) Equations for Maximizing Selection Efficiency for Predicting Will-Do Criterion Factors 396

9.8 Estimates of Maximizing Selection Efficiency Aggregated Over MOS: Predicting Core Technical Proficiency 396

9.9 Multiple Correlations for Predicting Second-Tour Job Performance (LVII) Criteria From ASVAB and Various Combinations of ASVAB, Selected Experimental Battery Predictors, and First-Tour (LVI) Performance Measures 399

List of Figures

1.1 Project A/Career Force Military Occupational Specialties (MOS) 13

1.2 Glossary of Terms for Project A/Career Force Research Samples 14

1.3 Project A/Career Force Research Flow and Samples 15

1.4 Linkages Between Literature Review, Expert Judgments, and the Preliminary and Trial Batteries 17

1.5 Experimental Predictor Battery Tests and Relevant Constructs 43

1.6 Longitudinal Validation Experimental Battery: Composite Scores and Constituent Basic Scores 44

1.7 Composite Scores That Reflect End-of-Training (EOT) Performance Factors 46

1.8 Summary List of CVII Basic Criterion Scores 49

1.9 Relationship of Specific Variables to Overall Factors in the CVII Performance Model 51

Page

1.10 Hierarchical Relationships Among Functional Categories, Task Factors, and Task Constructs . . . 65
1.11 Summary List of LVI Basic Criterion Scores . . . 66
1.12 Summary List of LVII Basic Criterion Scores . . . 114
1.13 Final LVII Criterion and Alternate Criterion Constructs Based on More Parsimonious Models . . . 123
1.14 Baseline Survivor Functions for One MOS From Each of the Four Job Groups . . . 147
1.15 Baseline Hazard Functions for One MOS From Each of the Four Job Groups . . . 148
1.16 Survivor Functions for C3 Soldiers Scoring Above Various Truncation Points on the Attrition Composite . . . 151
1.17 Survivor Functions for C4 Soldiers Scoring Above Various Truncation Points on the Attrition Composite . . . 151
1.18 Survivor Functions for NC3 Soldiers Scoring Above Various Truncation Points on the Attrition Composite . . . 152
1.19 Survivor Functions for NC4 Soldiers Scoring Above Various Truncation Points on the Attrition Composite . . . 152

2.1 Six EOT Performance Factor Scores Based on Measures of Training Performance Obtained at the End of Basic and Technical Training . . . 171
2.2 Five LVI First-Tour Performance Factor Scores and the Basic Criterion Scores That Define Them as Obtained From the First-Tour Performance Measures . . . 172
2.3 Six LVII Performance Factor Scores for Second-Tour NCO Performance and the Basic Criterion Scores That Define Them, as Obtained From the Second-Tour Performance Measures . . . 173

3.1 Mean Occupational Scale Scores by MOS for Males in the LV (i.e., Cross-Validation) Sample . . . 217

5.1 White-Black Differential Prediction Analysis for MOS 88M . . . 256

8.1 Bivariate Scatterplot of 10,000 Points in a Population of Applicants . . . 317
8.2 Classification Efficiency, As Measured by reMAP, Comparing Least-Squares and Synthetic Weights Using the Core Technical Proficiency Criterion, Selecting All Applicants . . . 350
8.3 Classification Efficiency, As Measured by reMAP, Comparing Least-Squares and Synthetic Weights Using the Core Technical Proficiency Criterion, Selecting 95% of Applicants . . . 350
8.4 Classification Efficiency, As Measured by reMAP, Comparing Least-Squares and Synthetic Weights Using the Overall Performance Criterion, Selecting All Applicants . . . 351
8.5 Classification Efficiency, As Measured by reMAP, Comparing Least-Squares and Synthetic Weights Using the Overall Performance Criterion, Selecting 95% of Applicants . . . 351

9.1 Project A/Career Force Military Occupational Specialties (MOS) . . . 360
9.2 Glossary of Terms for Project A/Career Force Research Samples . . . 361
9.3 Project A/Career Force Research Flow and Samples . . . 362
9.4 The Project A Experimental Predictor Battery . . . 365
9.5 Longitudinal Validation Experimental Battery: 28 Composite Prediction Scores and Their Constituent Basic Subtest Scores . . . 370
9.6 Criterion Measures Used to Assess First-Tour Performance . . . 374
9.7 Basic Criterion Scores Derived From First-Tour Performance Measures . . . 375

Page

9.8 Five LVI First-Tour Performance Factor Scores and the Basic Criterion Scores That Define Them as Obtained From the First-Tour Performance Measures 377
9.9 Basic Criterion Scores Generated From Second-Tour Performance Measures (LVII) 380
9.10 Leadership Factor Model 383
9.11 Six EOT Performance Factor Scores Based on Measures of Training Performance at the End of Basic and Technical Training 385
9.12 Description of Samples for EOT, LVI, and LVII Comparisons of Basic Validities and Incremental Validities 391

Chapter 1
A SUMMARY OF PROJECT A/CAREER FORCE

John P. Campbell and James H. Harris

This report represents the conclusion of 12 years of research sponsored by the U.S. Army to support the personnel selection and classification system for enlisted personnel. The research was performed in two phases. The first phase was Project A. Its goals were to validate the Armed Services Vocational Aptitude Battery (ASVAB) by collecting data from a representative sample of Military Occupational Specialties (MOS), and to build a large and versatile data base by developing and validating new predictors and criterion measures that represent the entire domain of potential measures.

Phase two of the research program was Building the Career Force, the final results of which are reported here. The goals of Career Force were to determine the longitudinal relationship between the new predictors and first-tour performance, to finalize and administer the measures of second-tour job performance, and to examine how selection and classification tests administered before a soldier's first enlistment, with performance during that enlistment, predict performance in a second enlistment.

In this chapter, to provide the context for the two phases of research, we first describe the present Army personnel system and provide a brief history of military selection and classification research. The bulk of this chapter is then devoted to a summary of Project A and the Career Force project, to provide sufficient background for understanding the results of the research described in the remaining chapters.

The discussion combines the two projects but distinguishes them when appropriate to do so. For example, over the life of the two projects, the military jobs that we examined changed several times; the changes were either to the name or alphanumerical indicator for the job, or in the job itself. In this chapter, we talk about the current sample of jobs only. Complete detailed descriptions of the research leading up to the results summarized here have been presented in a series of annual reports on both projects and a Final Report on Project A, and are available to interested readers (Campbell, 1987a, 1987b, 1988, 1991; Campbell & Zook, 1990, 1991, 1994a, 1994b, 1994c).

CHARACTERISTICS OF THE PRESENT ARMY PERSONNEL SYSTEM

The major stages of the selection, classification, and assignment process for persons entering enlisted service in the Army are presented in Table 1.1, and discussed below. The size, diversity, and widespread geographical distribution of Army activities have long dictated that the initial stages of personnel recruitment, selection, classification, and training be performed across many specialized units or activities and by personnel who have been specifically trained for these functions with guidance from command. Certain other functions are both formalized and carried out at the command level.

Table 1.1
The Army Selection, Classification, and Evaluation Process

Stage/Activity	Process	Outcome
Recruitment (U.S. Army Recruiting Command)	• Recruiting Incentives, Options • Recruiter Interviews • Aptitude Prescreening Test (EST) (CAST) • Records Checks	• To MET Sites or MEPS • Disqualified
Selection/Classification (Military Entrance Processing Stations)	• Aptitude Testing (ASVAB) • Physical Exam (PULHES) • Moral Screening • Special Tests • Counseling • Classification	• To Training Center • Disqualified
Entry Training (Army Training Center and Schools)	• Basic Training • Individual Training • Training Evaluation • Assignment • Disciplinary Reviews • Special Courses	• To Units • Reassigned/Recycled • Discharged
First Term (Operating Units)	• Unit (on-the-job) Training and Mission Activities • Special Courses • Evaluation • Disciplinary Reviews • Promotion Eligibility • Reenlistment Counseling and Screening • Army Continuing Education System	• Promotion/Demotion • Discharge (prior to ETS) • Separation (ETS) • Reenlistment
Second Term (Operating Units)	• Unit Training and Mission Activities • Advanced Technical/ Leadership Training • Evaluation • SDT Scores • Promotion Eligibility	• Promotion/Demotion • Reassigned • Discharge (prior to ETS) • Separation (ETS) • Reenlistment

These include unit or on-the-job training; performance evaluation; and decisions concerning promotion, discipline, reassignment, and retention or separation from service.

Recruitment

It is difficult to discuss recruitment, selection, and classification separately. They are interdependent processes. Their complementary nature should be evident in the ensuing discussion.

The Army has succeeded in meeting or approximating its numerical recruitment quotas in most of the years following the change to an All Volunteer Force. The number of enlisted accessions has averaged in the neighborhood of 100,000, from over twice that many applicants, in the past 15 fiscal years. Many qualified applicants do not enter the Army immediately but enter the delayed entry program (DEP) where they await their date to begin active duty. Though some DEP attrition occurs, in recent years the DEP has been filled to capacity.

The Army seeks to recruit the most capable personnel. Quality is generally defined in terms of high school graduation status and average or above scores on the Armed Forces Qualification Test (AFQT). The AFQT is a composite of four subtests (comprising verbal and math content) from the overall selection and classification instrument, the Armed Services Vocational Aptitude Battery. AFQT scores are reported in percentiles relative to the 1980 national youth population and grouped as follows:

AFQT Category	Percentile Score Range
I	93-99
II	65-92
IIIA	50-64
IIIB	31-49
IVA	21-30
IVB	16-20
IVC	10-15
V	1-9

Categories I and II signify well-above and above average trainability, respectively. Category III denotes average trainability, and Category IV signifies below average. Individuals scoring within Category V are, by law, not eligible for enlistment.

Because of their likelihood of success in training (and now with evidence of the AFQT's relationship to job performance), the Army attempts to maximize the recruitment of those scoring within Categories I through IIIA. In addition, because traditional high school graduates are more likely to complete their contracted enlistment terms, in contrast to nongraduates and alternative credential holders (e.g., GED credential holders), they are actively recruited as well.

Though qualification for initial enlistment into the Army is based upon a number of criteria (including age, and moral and physical standards), education and particularly aptitude are the most important. The Army targets its advertising and aims its recruiting resources to attract quality recruits. Also, to identify recruitment prospects while offering a career guidance tool, the ASVAB is administered to 900,000 high school juniors and seniors annually as part of the Department of Defense Student Testing Program.

At times, the Army has recruited non-high school graduates and applicants scoring in AFQT Category IV in order to meet numerical requirements and budget constraints. And, between 1976 and 1980, the Army erroneously enlisted high proportions of these less-preferred recruits as a result of an ASVAB misnorming. This situation raised

concerns in Congress, and led to the imposition of ceilings on the proportion of non-high school graduates and Category IVs who may be enlisted. One of the outcomes of Project A and Career Force is a much more solid empirical basis for qualification decisions. In fact, this research is particularly timely, given indications that banner recruiting times have tapered off.

To compete with the other Services and with the private sector for the prime target group, the Army has had to offer a variety of special inducements, including "critical skill" bonuses and educational incentives. One of the most popular has been the "training of choice" enlistment to a specific school training program, provided that applicants meet the minimum aptitude and educational standards and other prerequisites, and that training "slots" are available at the time of their scheduled entry into the program. Additional options, offered separately or in combination with "training of choice," include guaranteed initial assignment to particular commands, units, or bases, primarily in the combat arms or in units requiring highly technical skills. In recent years, a large proportion of all Army recruits, especially in the preferred aptitude and educational categories, have been enlisted under one or more of these options. An important research contribution would be to provide counselors with improved data-based aids to help create optimal person-job choices in light of Army manpower needs.

The importance of aptitude in recruiting decisions is exemplified in the prescreening of applicants at the recruiter level. For applicants who have not previously taken the ASVAB and whose educational/aptitude qualifications appear to be marginal in terms of the Army's trainability standards, the recruiter may administer a short Computerized Adaptive Screening Test (CAST) or Enlistment Screening Test (EST) to assess the applicant's prospects of passing the ASVAB. The Army has also experimented with non-cognitive tests to identify individuals who are likely to be poor risks in terms of the probability of completing Army basic training.

Applicants who appear upon initial recruiter screening to have a reasonable chance of qualifying are referred to one of approximately 700 Mobile Examining Team (MET) sites for administration of the ASVAB, or directly to a Military Entrance Processing Station (MEPS) where all aspects of enlistment testing are conducted.

Selection and Classification at the MEPS

For applicants found qualified for enlistment, classification and assignment to a particular training activity are completed on the basis of the information assembled.

The current versions of the ASVAB (Forms 20-22) consist of 10 subtests:

1. Arithmetic Reasoning
2. Numerical Operations
3. Paragraph Comprehension
4. Word Knowledge
5. Coding Speed
6. General Science

7. Mathematics Knowledge
8. Electronics Information
9. Mechanical Comprehension
10. Automotive-Shop Information

In addition to AFQT scores, selected subtest scores are combined to form ten aptitude composite scores, based on subtests that have been found to be most valid as predictors of successful completion of the various Army school training programs. For example, the composite score for administrative specialties is based on the Numerical Operations, Paragraph Comprehension, Word Knowledge, and Coding Speed subtests: the composite score for electronics specialties is based on the scores for Arithmetic Reasoning, General Science, Mathematics Knowledge, and Electronics Information.

As stated above, eligibility for enlistment, in terms of the trainability standard, is based upon a combination of criteria: AFQT score, Aptitude Area composite scores, and whether or not the applicant is a high school diploma graduate. Under the most recent Army regulations, the following standards were in effect[1]:

- High school graduates are eligible if they achieve an AFQT percentile score of 16 or higher and a standard score of 85 in at least one Aptitude Area.

- GED high school equivalency holders are eligible if they achieve an AFQT percentile score of 31 or higher and a standard score of 85 in at least one Aptitude Area.

- Non-high school graduates are eligible only if they achieve an AFQT percentile score of 31 or higher and standard scores of 85 in at least two Aptitude Areas.

In addition to these formal minimum requirements, the Army may set higher operational cut scores for one or all of these groups. Physical standards are captured in the PULHES profile, which rates the applicant on General Physical (P), Upper torso (U), Lower torso (L), Hearing (H), Eyes (E), and Psychiatric (S). Scores of 1 or 2 (on a 5-point scale) are required on all six indicators to be accepted for military duty (though waivers may be extended to applicants with a score of 3 on one or two indicators). In addition to the PULHES, the Army also sets general height and weight standards for enlistment.

Initial Classification

The overwhelming majority of enlistees enter the Army under a specific enlistment option that guarantees choice 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 entry into service. This is accomplished at MEPS by referring applicants who have passed the basic screening

[1]Army Regulation 601-201, 1 October 1980, revised, Table 2-2.

criteria (aptitude, physical, moral) to an Army guidance counselor, whose responsibility is to match the applicant's qualifications and preferences to Army current skill training requirements, and to make "reservations" for training assignments, consistent with the applicant's enlistment option.

For the enlistee, this decision will determine the nature of his or her initial training and occupational assignment, future military work environment, and chances of successful advancement in an Army career. For the Army, the relative success of the assignment process will significantly determine the aggregate level of performance and attrition for the entire force.

The classification and training "reservation" procedure is accomplished by the Recruit Quota System (REQUEST) which was implemented in 1973. This computer-based system coordinates the information needed to reserve training slots for volunteers. REQUEST uses minimum qualifications for accessions control. Thus, for an applicant who 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 only limited knowledge about the actual job content and working conditions), (b) availability of training slots, and (c) current priority assigned to filling each military occupational speciality (MOS).

These interactions among recruitment, selection, and classification in the current Army system give rise to several issues. There is an evident need for decision-making algorithms designed to maximize the overall utility of the MOS assignments. This requires that the average differential utilities of alternative assignments be known, as well as the marginal utility of each additional assignment to an MOS. The Army system currently incorporates marginal utilities by specifying desired distributions of AFQT scores, which are termed quality goals.

In general, the parameters of recruit supply and demand (e.g., number of applicants in various categories, selection ratio, percentage of training slots filled, MOS priority) must also be taken into account when developing algorithms for selection and classification. The decision process must also allow for the potentially adverse impacts on recruitment if the enlistee's interests, work values, and preferences are not given sufficient weight. Clear trade-offs must be evaluated between the procedures necessary (a) to attract qualified people, and (b) to put them into the right slots.

Initial Training

After processing at a Reception Battalion, all non-prior service Army recruits are assigned to a basic training program (BT) of 8 weeks which is followed, with few exceptions, by a period of advanced individual training (AIT), designed to provide basic entry-level skills. Entrants into some of the combat arms and the military police receive both their basic training and their AIT at the same Army base (One Station Unit Training, OSUT) in courses of about 3-4 months total duration. Those assigned to other

specialties are sent to separate Army technical schools whose course lengths vary considerably, depending upon the technical complexity of the MOS. The diversity of course offerings is illustrated by the fact that the Army provides initial skills training in approximately 250 separate courses.

In contrast to earlier practice, most enlisted trainees do not currently receive school grades upon completion of their courses, but are evaluated under Pass/Fail criteria. Those initially failing certain portions of a course are recycled. The premise is that slower learners, given sufficient time and effort under self-paced programs, can normally be trained to a satisfactory level of competence, and that this additional training investment is cost-effective. Those who continue to fail the course may be reassigned to other, often less demanding specialties or discharged from service. One consequence of these practices is to limit the usefulness of the selection/ classification practices as predictors of later performance.

Performance Assessment in Army Units

Upon assignment to an Army unit, most personnel actions affecting the career of the first-term enlistee are initiated by his or her immediate supervisor and/or the unit commander. These include the nature of the duty assignment, provision of on-the-job or unit training, and assessments of performance, both on and off the job. These assessments influence such decisions as promotion, future assignment, and eligibility for reenlistment, as well as possible disciplinary action (including early discharges).

To assure that these processes are administered fairly and consistently, in a manner compatible with broader Army objectives, detailed Army regulations govern enlisted personnel management. Army Regulation 600-211, The Enlisted Personnel Management System and related regulations cover such subjects as enlisted personnel evaluation and promotion, while AR 601-280, The Army Reenlistment Program prescribes the qualifications for reenlistment.

During an initial 3-year enlistment term, the typical enlistee can expect to progress to pay grade E-4, although advancement to higher pay grades for specially qualified personnel is not precluded. Authority to promote qualified personnel up to grade E-4 is delegated to unit commanders; promotion to higher grades is numerically restricted and must be approved either by field grade commanders for grades E-5 and E-6 or by HQDA for grades E-7 through E-9.

Promotion to E-2 is almost automatic after 6 months of service. Promotions to grades E-3 and E-4 normally require completion of certain minimum periods of service (12 and 24 months, respectively), but are subject to certain numerical strength limitations and specific commander approval. Unit commanders also have the authority to reduce assigned soldiers in pay grade, based on misconduct or inefficiency.

The Enlisted Evaluation System provides for an evaluation both of the soldier's proficiency in his or her MOS and of overall duty performance. The process includes a subjective evaluation based on supervisory appraisal of performance and on ratings that are conducted at the unit level under prescribed procedures. In addition, objective evaluations of physical fitness (i.e., the Army Physical Fitness Test) and job proficiency generally have been included in the system, particularly in the areas of promotion and retention.

In 1978 the Army replaced the MOS Proficiency Tests with Skill Qualification Tests for skill levels one through four (E-1 through E-7). The latter were criterion-referenced, paper-and-pencil performance-knowledge tests that evaluated soldiers' ability to perform MOS- and skill level-specific critical job tasks satisfactorily. Scores from a soldier's last SQT were used in making promotion decisions for grades E-5 through E-8. The SQT program was canceled in 1991 as a cost-saving measure. Soldiers in grades E-5 and E-6 were administered Self-Development Tests (SDT) on an annual basis for several years but that program has been discontinued.

Reenlistment Screening

The final stage of personnel processing of first-term enlisted personnel is screening for reenlistment eligibility. As described in AR 601-280, this review considers such criteria as disciplinary records; aptitude area scores (based on ASVAB or its predecessors); low SDT scores, when applicable; weight standards; and slow grade progression "resulting from a pattern of marginal conduct and/or performance." Enlisted personnel who wish to reenlist but do not meet certain minimum standards under these criteria must obtain a waiver before being processed for reenlistment.

The cumulative losses due to attrition, reenlistment screening, and non-reenlistment of eligible personnel have resulted in the progressive diminution of initial Army cohorts to only about 10-20 percent of their original numbers, by the time they enter the sixth year of enlisted service. Not all of the latter, moreover, are retained or wish to be retained in their original specialties, since an offer of retraining is often an inducement for reenlistment. The cumulative impact of this skill drain upon the Army is considerable.

Summary

Even this brief description of the current system illustrates the complexity of the Army's personnel decision-making requirements and the large number of parameters that must be taken into account. In addition, these decisions must be made for a very large flow of individuals within a very short time frame. In this regard the Army faces a much more difficult personnel management task than virtually any other organization. More effective selection/classification/promotion strategies would pay large dividends.

A BRIEF HISTORY OF SELECTION AND CLASSIFICATION

Formal personnel selection and classification using standardized measures of individual differences actually began in 1115 B.C. with the system of competitive examinations that led to appointment to the bureaucracy of Imperial China (DuBois, 1964). It soon included the selection/classification of individuals for particular military specialties, as in the selection of spear throwers with standardized measures of long-distance visual acuity (e.g., identification of stars in the night sky).

Systematic attempts to deal with selection/classification issues have been a part of military management ever since. Military organizations are virtually unique in their need to make large numbers of complex personnel decisions in a short period of time. However, the centrality of criterion-related validation to a technology of selection and classification was not fully articulated until World War II, and research and development sponsored by the military has been the mainstay of growth in that technology from that time to the present.

The contributions of military psychologists during World War II are well-known and well-documented. The early work of the Personnel Research Branch of The Adjutant General's Office was summarized in a series of articles in the Psychological Bulletin (Staff, PRB, AGO, 1943 a, b, c, d, e, and f). Later work was published in Technical Bulletins and in such journals as Psychometrika, Personnel Psychology, and Journal of Applied Psychology. The Aviation Psychology Program of the Army Air Forces (AAF) issued 19 volumes, with a summary of the overall program presented in Volume I (Flanagan, 1948). In the Navy, personnel research played a smaller and less centralized role, but here too useful work was done by the Bureau of Naval Personnel (Stuit, 1947).

Much new ground was broken. There were important advances in the development and analysis of criterion measures: Thorndike's textbook based on his Air Force experience presented a state-of-the-art classification and analysis of potential criteria (Thorndike, 1949). Improvements were made in rating scales. Forced-choice methods were developed by the Personnel Research Branch; checklists based on critical incidents were used in the AAF program. The sequential aspect of prediction was articulated and examined: tests "validated" against training measures (usually pass/fail) were checked against measures of success in combat (usually ratings or awards). At least one "pure" validity study was accomplished, when the Air Force sent 1,000 cadets into pilot training without regard to their pilot stanine derived from the classification battery (Flanagan, 1948). This remains one of the few studies that could report validities without correcting for restriction of range. Historically, 1940 to 1946 was a period of concentrated development of selection and classification procedures, and the further accomplishments of the next several decades flowed directly from it.

In part, this continuity is attributable to the well-known fact that many of the psychologists who had worked in the military research establishments during the war became leaders in the civilian research community after the war. In part, it is

attributable to the less widely recognized fact that the bulk of the work continued to be funded by military agencies. The Office of Naval Research, the Army's Personnel Research Branch (and its successors), and the Air Force Human Resources Research installations were the principal sponsors.

The bibliography is very long. Of special relevance to Project A and Career Force is the pioneering work on differential prediction by Brogden (1946a, 1951) and Horst (1954, 1955); on utility conceptions of validity by Brogden (1946b) and Brogden and Taylor (1950); on the "structure of intellect" by Guilford (1957); on the establishment of critical job requirements by Flanagan and associates (Flanagan, 1954); and on the decision-theoretic formulations of selection and classification developed by Cronbach and Gleser (1957) for the Office of Naval Research. The last of these (Psychological Tests and Personnel Decisions) was hailed quite appropriately as a breakthrough--a "new look" in selection and classification. But the authors were the first to acknowledge the relevance of the work cited above of Brogden and Horst. It was the culmination of a lengthy sequence of development.

Project A and Career Force were carried out in the context of an impressive history of selection and classification research, and, taken together, these research projects have become another milestone. It is by far the most comprehensive personnel research and development program ever attempted. It is unique in that a complete personnel system was examined at one time. The jobs (MOS) studied were sampled representatively from the complete population of jobs, new predictor measures were sampled systematically from the full domain of potential information, and job performance was assessed as thoroughly as possible with multiple measures.

Given this data base, and using state-of-the-art analytic techniques, the functioning of the selection/ classification decision process has been modeled and actually evaluated under various goals or constraints. We regard Project A and Career Force as true landmarks in personnel research. The reasons for this judgment are detailed below in the summary descriptions of Project A and Career Force and in the subsequent chapters that report the final stages of our data analyses.

THE GOALS AND DESIGN OF PROJECT A AND CAREER FORCE

Project A and Career Force were designed to provide the greatest possible increase in overall performance and readiness that can be obtained from improved selection, classification, and allocation of enlisted personnel. These two research phases provided an integrated examination of performance measurement, selection/classification, supply and demand parameters, and allocation procedures such that the Army could attempt to optimize the achievement of multiple personnel management goals (e.g., improve performance and decrease attrition).

The responsibilities of the combined projects were to develop (a) a comprehensive set of new predictor measures; (b) multiple measures of job performance; (c) accurate estimates of the predictability of future performance; (d) decision rules for selection/

classification at enlistment and reenlistment to optimize individual and system performance; and (e) a "what-if" gaming capability to illustrate the effects of variations in personnel management policies.

The impetus for the research program came from the practical, professional, and legal need to demonstrate the validity of the ASVAB and other selection variables for predicting job performance. Much of the existing validity data was based on using training measures as criteria. As the Army Research Institute (ARI) began reviewing the design needed to meet that requirement, the concept of a larger program began to emerge. With only a moderate increase in resources, new classification measures in the perceptual, psychomotor, interest, temperament, and biodata domains could be evaluated as well. Also, a longitudinal research database could be developed, linking soldiers' performance on a variety of variables from enlistment, through training, first-tour assignments, reenlistment decisions, and for some, to their second tour.

The research program began in 1982. The research was performed by a consortium composed of the Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), and Personnel Decisions Research Institute, Incorporated (PDRI, Inc.), under contract to and in collaboration with the Army Research Institute.

Specific Program Objectives

The Project A Research Plan spoke to these specific objectives:

(1) Develop new measures of job performance that can be used as criteria against which to validate selection/classification measures.

(2) Develop a general model of performance for entry-level skilled jobs.

(3) Validate existing selection measures against both existing and project-developed criteria.

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

(5) Develop and validate new selection and classification measures.

(6) Develop a utility scale for different performance levels across MOS.

The specific objectives of Career Force were to:

(1) Develop a complete array of valid and reliable measures of second-tour performance as an Army NCO, using the Project A prototypes as a starting point.

(2) Develop a model of second-tour performance that parallels the first-tour performance model from Project A and that identifies the major components of second-tour performance, provides information on their construct validity, and establishes how the major components of performance should be combined for specific prediction or interpretation purposes.

(3) Carry out a complete incremental predictive validation of (a) the ASVAB and the Project A Experimental Battery of predictors, (b) measures of training success, and (c) the full array of first-tour performance criteria developed as part of Project A. The criteria against which these three sets of predictors were validated, both individually and incrementally for each major criterion component, are the second-tour job performance measures.

(4) Estimate the degree of differential prediction across (a) major domains of predictor information (e.g., abilities, personality, interests), (b) major factors of job performance, and (c) different types of 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 the analytic framework needed to evaluate the optimal prediction equations for predicting (a) training performance; (b) first-tour performance; (c) first-tour attrition and the reenlistment decision; and (d) second-tour performance, under the conditions when testing time is limited to a specified amount and when there must be a tradeoff among alternative selection/classification goals (e.g., maximizing aggregate performance vs. minimizing discipline and low-motivation problems vs. minimizing attrition).

(7) Design and develop a fully functional and user-friendly research data base that includes all relevant personnel data on 1981/82, 1983/84, and 1986/87 accessions, including all Project A and Career Force Project data and all relevant Enlisted Master File (EMF), Accession File, and Army Training Requirements and Resources System (ATRRS) data.

The Research Samples

In general, the combined design for Project A/Career Force encompasses two major cohorts of soldiers (new accessions for 1983/84 and for 1986/87), both of which were followed into their second tour of duty and which collectively have produced six major research samples. For each research sample there is a battery of predictor measures and an array of performance measures. For each of the six samples the predictor battery is composed of the ASVAB and either the Trial Battery or the Experimental Battery version of the new tests developed in Project A (see Campbell & Zook, 1991). There were three distinct arrays of performance measures corresponding to the need to assess (a) training performance, (b) first-tour job performance, and (c) second-tour job performance.

The MOS in the two groups were carefully sampled to represent the variation in job content in the Army occupational structure. In addition, they were selected so as to overrepresent both the combat specialties and those MOS with the larger proportions of women and minority groups. The MOS selection procedure has been described in detail in previous Project A reports (e.g., Campbell, 1987).

In each sample the individuals to be assessed were selected from two predetermined sets of MOS -- Batch A and Batch Z. They are listed in Figure 1.1. The two groups differed in that tests administered to Batch A MOS included MOS-specific rating scales and job knowledge and hands-on tests, whereas the only MOS-specific measure administered to the Batch Z MOS was an end-of-training test.

A glossary of terms for the samples and for the different measurement batteries is given in Figure 1.2. The six major samples, their approximate size, and the predictor and/or performance batteries that were to be administered to each are shown in Figure 1.3.

Batch A		Batch Z	
MOS		MOS	
11B	Infantryman	12B	Combat Engineer
13B	Cannon Crewmember	16S	MANPADS Crewman
19E	M60 Armor Crewman	27E	Tow/Dragon Repairer
19K	M1 Armor Crewman[a]	29E	Comm-Electronics Radio Repairer[d]
31C	Single Channel Radio Operator	51B	Carpentry/Masonry Specialist
63B	Light-Wheel Vehicle Mechanic	54B	NBC Specialist[e]
71L	Administrative Specialist	55B	Ammunition Specialist
88M	Motor Transport Operator[b]	67N	Utility Helicopter Repairer
91A/B	Medical Specialist/Medical NCO[c]	76Y	Unit Supply Specialist
95B	Military Police	94B	Food Service Specialist
		96B	Intelligence Analyst[d]

[a] Except for the type of tank used, this MOS is equivalent to the 19E MOS originally selected for Project A testing.
[b] This MOS was formerly designated as 64C.
[c] Although 91A was the MOS originally selected for Project A testing, second-tour medical specialists are usually reclassified as 91B.
[d] This MOS was added after the Concurrent Validation.
[e] This MOS was formerly designated as 54E.

Figure 1.1. Project A/Career Force Military Occupational Specialties (MOS).

Glossary of Terms	
CVI Sample (CVI)	Soldiers who entered the Army between 1 Jul 83 - 30 Jun 84 and were in 1985 Project A Concurrent Validation. They were administered the Trial Predictor Battery and the first-tour job performance measures.
CVII Sample (CVII)	Soldiers who entered the Army between 1 Jul 83 - 30 Jun 84 and were in the 1985 Project A Concurrent Validation (CVI) and the 1988 Second-Tour Concurrent Validation (CVII). They were administered the second-tour job performance measures and were re-administered the ABLE.
LV Sample (LV)	Soldiers in the Longitudinal Validation sample who entered the Army between 20 Aug 86 - 30 Nov 87 and were administered the Experimental Predictor Battery and End-of-Training measures.
LV Training Sample (LVT)	Soldiers in the Longitudinal Validation sample who finished AIT and who were administered the End-of-Training measures.
LVI Sample (LVI)	Soldiers who entered the Army between 20 Aug 86 - 30 Nov 87 and were in the LV Sample and the 1988 First-Tour Longitudinal Validation Sample. They were administered the first-tour job performance measures.
LVII Sample (LVII)	Soldiers who entered the Army between 20 Aug 86 - 30 Nov 87 and were in the LVI Sample and the Longitudinal Validation (LVII) sample. They were administered the second-tour job performance measures in LVII.
Note. Glossary definitions reflect the original research plan. In actuality, some CVII soldiers did not have CVI data, some LVI soldiers did not have LV data, and some LVII soldiers did not have both LV and LVI data.	

Figure 1.2. Glossary of terms for Project A/Career Force research samples.

Procedure and Design

The data collection procedures for each sample have been described in detail in previous reports (e.g., see Campbell & Zook, 1990). Each data collection involved on-site administration by a trained data collection team headed by a team leader from the contractor staff who worked closely with a designated Army point-of-contact (POC) at the site. A brief characterization of each of the six samples in terms of the timing, location, and duration (per soldier) of the data collection is given below.

The Concurrent Validation (CVI) sample. The data were collected at 13 posts in the continental United States and at multiple locations in Germany. Each individual was assessed for 1 1/2 days on the project-developed first-tour job performance measures and for 1/2 day on the new predictor measures (the Trial Battery). Most of the individuals in the sample had been in the Army for 18-24 months.

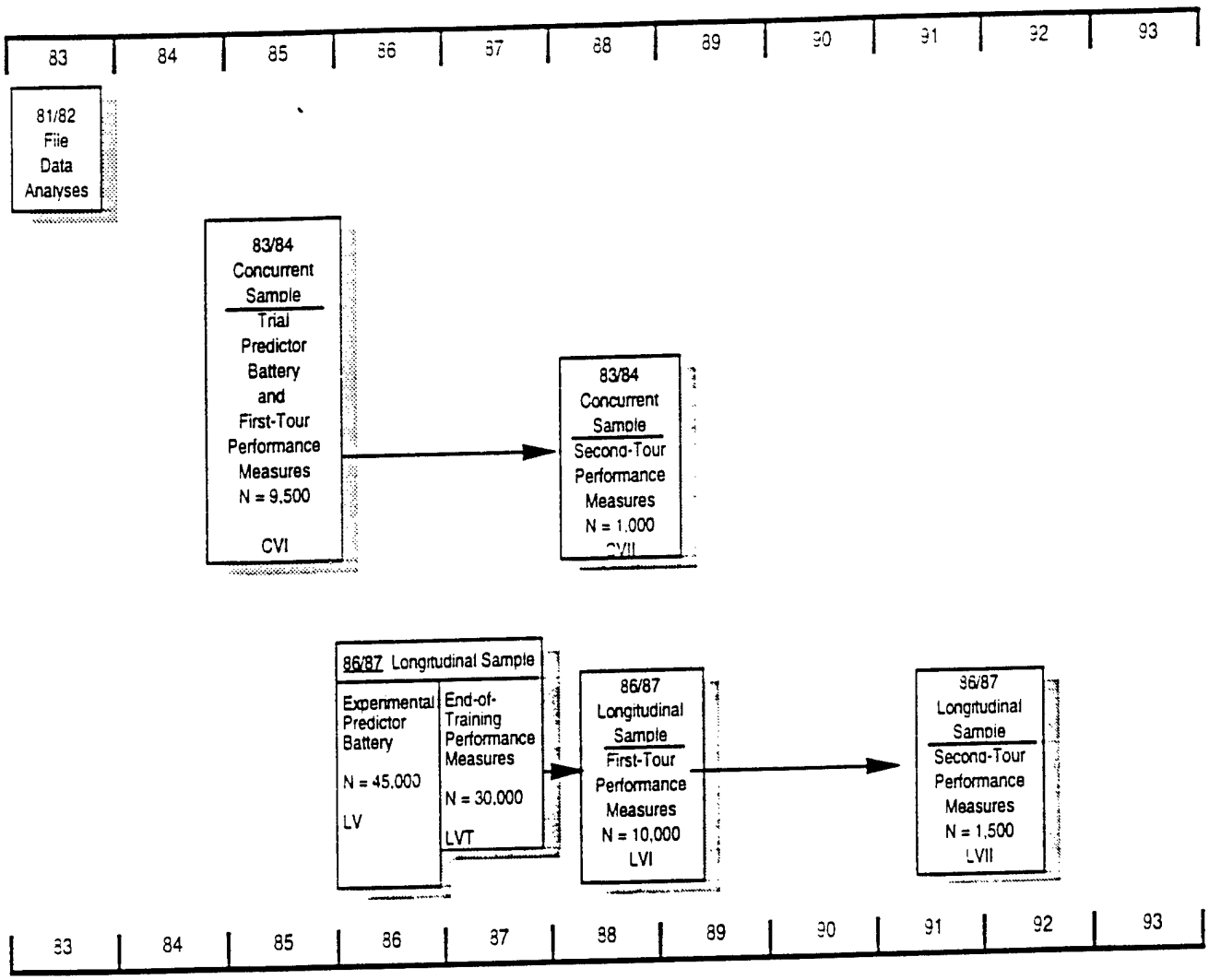


Figure 1.3. Project A/Career Force research flow and samples.

The Longitudinal Validation (LV) Sample. All individuals were assessed on the 4-hour Experimental Predictor Battery within 2 days of first arriving at their assigned Reception Battalion where they would undergo Basic/Advanced Individual training. Data were collected over a 14-month period at eight Reception Battalions by a permanent, on-site data collection team.

The Longitudinal Validation End-of-Training (LVT) Sample. The EOT performance measures were administered to those individuals in the LV sample who completed Advanced Individual Training (AIT), which could take from 2 months to 6 months, depending on the MOS. The training performance measures consisted of an MOS-specific training achievement test and a series of rating scales completed by peers and drill instructors. Data collection took place during the last three days of AIT.

The Longitudinal Performance Measurement (LVI) Sample. The individuals in the 86/87 cohort who were measured with the Experimental Predictor Battery, completed AIT, and remained in the Army were assessed with the full array of first-tour job performance measures when most of them were between 18 and 24 months of service. Data collections were conducted at 13 posts in the United States and multiple locations in Europe (primarily in Germany). The administration of the LVI first-tour criterion measures took one day per soldier.

The Concurrent Validation Second-Tour (CVII) Sample. The same data collection teams that administered the first-tour performance measures to the LVI sample also administered the second-tour performance measures at the same location and during the same time periods to a sample of junior NCOs from the 83/84 cohort in their second tour of duty (4-5 years of service). Every attempt was made to include second-tour personnel from the designated MOS who had been part of CVI. The CVII data collection took one day per soldier.

The Longitudinal Validation Second-Tour (LVII) Sample. This sample included members of the 86/87 cohort from the designated MOS who were part of the LVI (first-tour job performance measures) samples and who reenlisted for a second tour. The revised second-tour performance measures were administered at 15 U.S. posts, multiple locations in Germany, and two locations in Korea. The LVII performance assessment took one day per soldier.

A SUMMARY OF PROJECT A

Substantive summaries of the work in the two projects, in furtherance of the above goals and project design, are given below. The formats differ because of the time context. The Project A summary is less detailed and follows the structure of the project design. For Career Force, which is just now being concluded, the summary is presented in a year by year format for the first four years of work.

Predictor Development in Project A

A major objective was to develop an experimental battery of new selection/classification tests that would be potentially valuable additions to ASVAB and would maximize the Army's capability to make accurate selection/classification decisions. Consequently, the overall Project A strategy was to identify a universe of potential predictor constructs appropriate for the population of enlisted MOS, sample representatively from it, construct tests for each construct sampled, and refine and improve the measures through a series of pilot and field tests. The intent was to develop a predictor battery that was maximally useful for an entire population of jobs.

The long process of predictor development is represented in Figure 1.4. It began with an in-depth search of the entire personnel selection literature. Literature review teams were created for cognitive abilities, perceptual and psychomotor abilities, and

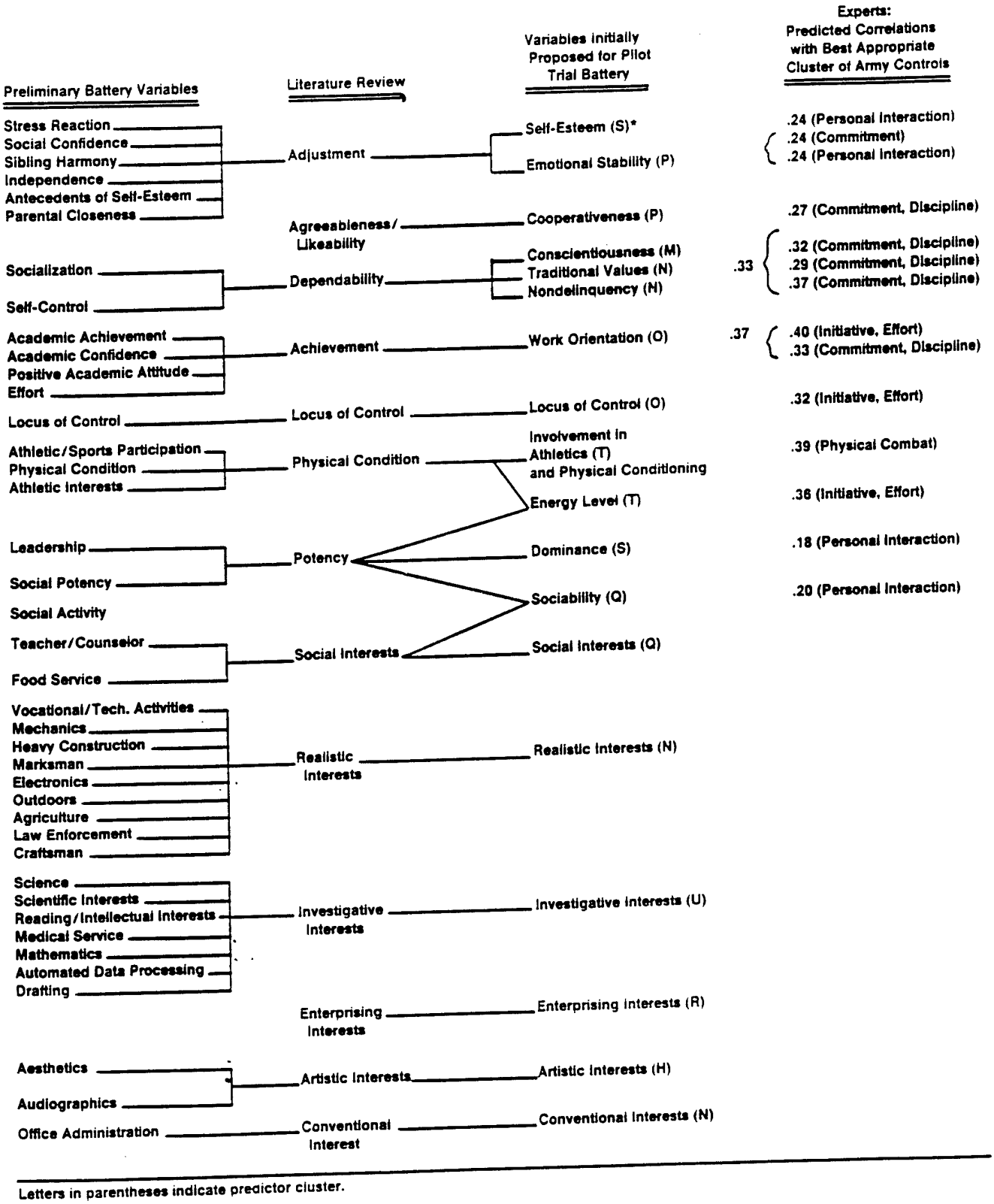


Figure 1.4. Linkages between literature review, expert judgments, and the Preliminary and Trial Batteries.

non-cognitive characteristics such as personality, interest, and biographical history. Every available automated and manual technique was used in the search and an initial list of several hundred variables was compiled. The list went through several waves of expert review and was eventually reduced to a list of 53 potentially useful predictor variables. They are listed in Table 1.2.

A sample of 35 personnel selection experts was then asked to estimate the correlation between each predictor construct and each criterion factor, when that correlation was corrected for restriction of range and criterion unreliability. The resulting judgments could be analyzed for the interjudge agreement, rows and columns could be factor analyzed, and the results could be compared to analogous information from the empirical literature. Most importantly, the exercise provided another substantial set of expert judgments about which predictor constructs should be the most useful. A hierarchical analysis of the predictor validity profiles is also shown in Table 1.2.

All the available information was then used to arrive at a final set of variables for which new measures would be constructed. This represented months of effort by many people to select the variables that would best supplement the ASVAB in predicting job performance across all MOS. What followed were many months more of instrument construction, several waves of pilot tests, and a series of major field tests. Included in these efforts were the development of a computerized battery of perceptual/psychomotor tests, the creation of the software, the design and construction of a special response pedestal permitting a variety of responses (e.g., one-hand tracking, two-hand coordination), and the acquisition of 108 portable computerized testing stations. After each data collection, revisions were made on the basis of item statistics and expert review. Finally on 15 May 1985, the predictor battery was deemed ready for concurrent validation. That battery, known as the Trial Battery, is summarized in Table 1.3.

Performance Measurement

The goals of training performance and job performance measurement in Project A and Career Force were to define, or model, the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor.

Some additional specific goals 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, multi-method approach), (c) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and (d) evaluate existing archival and administrative records as possible indicators of job performance.

Table 1.2
Hierarchical Map of Predictor Space

Constructs	Clusters	Factors
Verbal Comprehension Reading Comprehension Ideational Fluency Analogical Reasoning Omnibus Intelligence/Aptitude Word Fluency	A. Verbal Ability/ General Intelligence	Cognitive Abilities
Word Problems Inductive Reasoning Concept Formation Deductive Logic	B. Reasoning	
Numerical Computation Use of Formula/Number Problems	C. Number Ability	
Perceptual Speed and Accuracy	N. Perceptual Speed and Accuracy	
Investigative Interests	U. Investigative Interests	
Rote Memory Follow Directions	J. Memory	
Figural Reasoning Verbal and Figural Closure	F. Closure	
Two-dimensional Mental Rotation Three-dimensional Mental Rotation Spatial Visualization Field Dependence (Negative) Place Memory (Visual Memory) Spatial Scanning	E. Visualization/Spatial	Visualization/ Spatial
Processing Efficiency Selective Attention Time Sharing	G. Mental Information Processing	Information Processing
Mechanical Comprehension	L. Mechanical Comprehension	Mechanical
Realistic Interests Artistic Interests (Negative	N. Realistic vs. Artistic Interests	
Control Precision Rate Control Arm-hand Steadiness Aiming	I. Steadiness/Precision	Psychomotor
Multilimb Coordination Speed of Arm Movement	D. Coordination	
Manual Dexterity Finger Dexterity Wrist-finger Speed	K. Dexterity	
Sociability Social Interests	Q. Sociability	Social Skills
Enterprising Interests	R. Enterprising Interests	
Involvement in Athletics and Physical Conditioning Energy Level	T. Athletic Abilities/Energy	Vigor
Dominance Self-esteem	S. Dominance/Self-esteem	
Traditional Values Conscientiousness Non-delinquency Conventional Interests	N. Traditional Values/Conventionality/ Non-delinquency	Motivation/ Stability
Locus of Control Work Orientation	O. Work Orientation/Locus of Control	
Cooperativeness Emotional Stability	P Cooperation/Emotional Stability	

Table 1.3
Summary of Predictor Measures Used in Concurrent Validation (The Trial Battery)

Test (Construct) Name	Number of Items
Cognitive Paper-and-Pencil Tests	
Reasoning Test (Induction-figural reasoning)	30
Orientation Test (Spatial orientation)	24
Map Test (Spatial orientation)	20
Object Rotation Test (Spatial visualization - rotation)	90
Assembling Objects Test (Spatial visualization - rotation)	32
Maze Test (Spatial visualization - scanning)	24
Computer-Administered Tests	
Simple Reaction Time (Processing efficiency)	15
Choice Reaction Time (Processing efficiency)	30
Memory Test (Short-term memory)	36
Target Tracking Test No. 1 (Psychomotor precision)	18
Target Shoot Test (Psychomotor precision)	30
Perceptual Speed and Accuracy Test (Perceptual speed and accuracy)	36
Identification Test (Perceptual speed and accuracy)	36
Target Tracking Test No. 2 (Two-hand coordination)	18
Number Memory Test (Number operations)	28
Cannon Shoot Test (Movement judgment)	36
Inventory Name and Subscale Name	**Number of Items**
Non-Cognitive Paper-and-Pencil Inventories	
Assessment of Background and Life Experiences (ABLE)	209
Adjustment	
Dependability	
Achievement	
Physical Condition	
Leadership	
Locus of Control	
Agreeableness/Likability	
Army Vocational Interest Career Examination (AVOICE)	176
Realistic Interests	
Conventional Interests	
Social Interests	
Enterprising Interests	
Artistic Interests	

Given these intentions, the criterion development effort employed 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.

Performance Modeling

The development efforts were guided by a model that views performance as truly multidimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization.

For the population of entry-level enlisted positions, two major types of job performance components were postulated. The first is composed of components that are specific to a particular job and that would reflect specific technical competence or specific job behaviors that are not required for other jobs. It was anticipated that there would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills and would be reflected by different task content.

The second type includes components that are defined and measured in the same way for every job. Referred to as Army-wide performance factors, these incorporate the basic notion that total performance is much more than task or technical proficiency. It might include such things as contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity.

In sum, the working model of total performance with which Project A began viewed performance as multidimensional within the two broad categories of factors or constructs. The job analysis and criterion construction methods were designed to explicate the content of these factors via an exhaustive description of the total performance domain, several iterations of data collection, and the use of multiple methods for identifying basic performance factors.

Saying that performance is multidimensional does not preclude using just one index to make a specific personnel decision (e.g., select/not select, promote/not promote). 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 particular decision. Determining the specific combinational rules (e.g., simple sum, weighted sum, non-linear combination) that best reflect what the organization is trying to accomplish was to be a matter for research.

Criterion Development

Actual criterion development proceeded from two basic types of information. First, all available task descriptions were used to generate a population of job tasks for each MOS. The principal sources of task description are the Army Occupational Survey Program, which uses questionnaire checklists of several hundred task statements to survey job incumbents about the frequency with which they perform each task, and the

Soldier's Manual for each job, which is a complete specification by management of what the task content of the job is supposed to be. After considerable editing, revising, and a formal review by a panel of subject matter experts (SMEs), a population of 130-180 tasks was enumerated for each MOS in the Project A sample.

An additional series of expert judgments was then used to scale the relative difficulty and importance of each task and to cluster tasks on the basis of content similarity. Sampling tasks for measurement was accomplished via a Delphi procedure. That is, each member of a team of task selectors was asked to select 30 tasks from the population of tasks such that the selected tasks were representative of task content, were important, and represented a range of difficulty. The individual judge's choices were then regressed on the task characteristics and both the choices and the captured "policy" of each person were fed back to the group members, who then revised their choices as they saw fit. Typically, convergence was achieved quickly and the final selection was by consensus. The panels' selections were then thoroughly reviewed by the Army command responsible for that particular job.

Standardized hands-on job samples, paper-and-pencil job knowledge tests, and numerical ratings scales were then constructed to assess knowledge and proficiency on these tasks. Each measure went through multiple rounds of pilot testing and revision.

The second procedure used to describe job content was the critical incident method. Panels of NCOs and officers generated thousands of critical incidents of effective and ineffective performance. There were two basic formats for the critical incident workshops. One asked participants to generate incidents that potentially could occur in any job. The second type focused on incidents that were specific to the content of the particular job under consideration. The behaviorally anchored rating scale procedure was used to construct rating scales for performance factors specific to a particular job (MOS-Specific BARS) and performance factors that were defined in the same way and relevant for all jobs (Army-wide BARS).

The critical incident procedure was also used with workshops of combat veterans to develop rating scales of expected combat effectiveness.

Since a major program objective was to determine the relationships between training performance and job performance and their differential predictability, if any, a comprehensive training achievement test was constructed for each MOS. The content of the program of instruction (POI) was matched with the content of the population of job tasks, and items were written to represent each segment of the match. After pilot testing, revision, field testing, and Army proponent review, the result was a 150-200 item training achievement test for each of the 19 MOS.

The final category of criterion measure was produced by a search of the Army's archival records for potential performance indicators. First, all possibilities were enumerated from three major sources of such records:

Enlisted Master File (EMF) - a central computer record of selected personnel actions.

Enlisted Military Personnel File (EMPF) - the permanent historical record of an individual's military service kept on microfiche at a central location.

Military Personnel Records Jacket (MPRJ) - more commonly known as the 201 File, the personnel folder that follows the individual.

These three sources were systematically compared, using a sample of 750 people and a standardized information recording form. The 201 File looked the most promising in terms of recency and completeness, and six administrative performance indexes were eventually selected.

The complete array of performance measures, after revision on the basis of a large-scale field study (N = 150/MOS for nine MOS), is shown in Table 1.4. These are the measures which were administered to the concurrent sample of 400-600 people in each of the 19 MOS. The distinction between Batch A (9 MOS) and Batch Z (10 MOS) is that not all criterion measures were developed for each job; budget constraints dictated that the job-specific measures could be developed for only a limited number of jobs (i.e., the 9 MOS in Batch A).

The Concurrent Validation

Between 1 July and 1 December 1985 the predictor and criterion batteries were administered to 9,430 job incumbents in the 19 MOS. Four hours were devoted to the predictor tests and 12 hours to the criterion measures. Eight-person teams supported by four or five Army personnel visited each of 14 different Army posts for several weeks at a time. Considerable effort was devoted to training the data collection teams, standardizing testing conditions, keeping logs, and performing data checks each day. The concurrent validation sample sizes by site and by MOS are shown in Table 1.5.

If all the rating scales are considered separately, the MOS-specific measures are aggregated at the task or instructional module level, and the major predictor subscales are used, each individual has approximately 200 criterion scores and 70 predictor scores. There was an obvious need to aggregate variables to reduce collinearity and assist in appropriate interpretation of the results.

For both predictors and criteria, the procedure for getting from the individual task or scale scores to factor or construct scores was similar except for the degree to which the previous literature was of help. Many decades of research on the measurement of abilities, personality, and interests have provided a lot of information about the structure of individual differences. Similar help from the performance side is really not available except for a modest number of descriptive studies of specific occupations such as managers, nurses, police officers, fire fighters, and college professors.

Table 1.4
Summary of Criterion Measures Used in Concurrent Validation Samples[a]

Performance Measures Common to Batch A and Batch Z MOS (Jobs)

1. Ten behaviorally anchored rating scales designed to measure factors of non-job-specific performance (e.g., Giving peer leadership and support, maintaining equipment, self-discipline).

2. Single scale rating of overall job performance.

3. Single scale rating of NCO (noncommissioned officer) potential.

4. Ratings of performance on 13 representative "common" tasks. The Army specifies a series of common tasks (e.g., several first aid tasks) that everyone should be able to perform.

5. Paper-and-pencil Test of Training Achievement developed for each of the 19 MOS (130-210 items each).

6. A 77-item summated rating scale for the assessment of expected combat performance.

7. Five performance indicators from administrative records. The first three were obtained via self-report and the last two from computerized records.
   - Total number of awards and letters of commendation.
   - Physical fitness qualification.
   - Number of disciplinary infractions.
   - Rifle marksmanship qualification score.
   - Promotion rate (in deviation units).

Performance Measures for Batch A Only

8. Job-sample (hands-on) test of MOS-specific task proficiency.
   - Individual was tested on each of the 15 major job tasks.

9. Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.
   - Individual was scored on 150-200 multiple-choice items representing 30 major job tasks. Fifteen of the tasks were also measured hands-on.

10. Rating scale measures of specific task performance on the 15 tasks also measured with the knowledge tests and the hands-on measures.

11. MOS-specific behaviorally anchored rating scales. From 6 to 10 BARS scales were developed for each MOS to represent the major factors that constituted job-specific technical and task proficiency.

Auxiliary Measures Included in Criterion Battery

- A Job History Questionnaire asking for information about frequency and recency of performance of the MOS-specific tasks.
- Work Environment Description Questionnaire -- a 141-item questionnaire assessing situational/environment characteristics, leadership, climate, and reward preferences.

[a]All rating measures were obtained from approximately two supervisors and three peers for each ratee.

Table 1.5
Concurrent Validation Sample Soldiers by MOS by Location

Location	Batch A MOS									Batch Z MOS										Total	Percent Total
	11B	13B	19E	31C	63B	64C	71L	91A	95B	12B	16S	27E	51B	54E	55B	67W	76W	76Y	94B		
Fort Benning	45	23	41	7	13	39	16	9	13	13	15	3	0	12	18	9	13	15	12	316	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	4.61
Fort Stewart	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	1,963	20.80
Total	702	667	503	336	637	686	514	501	692	704	470	147	108	434	291	276	490	630	612	9,430	
Percent 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		

Both expert judgment and factor analytic results from the field tests were used to formulate hypothesized factors. These targets were then subjected to a series of quasi-confirmatory analyses using the Concurrent Validation sample. The resulting predictor construct scores and their associated component scales are shown in Table 1.6.

For the within-MOS criterion intercorrelation matrix, confirmatory analyses were used to test alternative models. The latent structure of performance that both fits the data in each job and seems to make sense is portrayed in Table 1.7.

The model best confirmed by LISREL (Jöreskog & Sörbom, 1981) specified five substantive and two methods factors, which were labeled the "ratings" factor and the "test" factor. The ratings factor was specified to be the first orthogonal component taken from all the ratings scales. The test factor is the first orthogonal component taken from the paper-and-pencil knowledge tests. Given this constraint, five substantive factors were extracted. The first two are based on the knowledge tests and the job sample measures. They are referred to as the Core Technical performance factor and the General Soldiering performance factor. The technical factor reflects content that is central and largely specific to the MOS. The general factor encompasses content that tends to be common across several jobs and is less central to the core performance objectives of each MOS.

The remaining factors are based on the ratings, primarily those developed by the critical incident method and the administrative/personnel records. Factor 3 encompassed the most scales and was the clearest in terms of its loading but appeared to be the most heterogeneous in terms of content. It seems to be a general effort and performance, performance under adverse conditions, peer leadership factor. Factor 4 is much more homogeneous and reflects the rating scales having to do with personal discipline and avoidance of trouble, and the number of negative personnel outcomes people reported. Factor 5 is fairly narrow in content and shows very clear loadings for ratings of military bearing and the physical fitness score that is part of everyone's personnel record. In general, this solution fit the data from all MOS and seemed reasonable and appropriate to Army management.

Table 1.6
Predictor Construct Scores

From Paper-and-Pencil Tests

Overall Spatial Factor
- Assembling Objects test
- Map test
- Maze test
- Object Rotation test
- Orientation test
- Figural Reasoning test

From Computerized Measures

Psychomotor Factor
- Cannon Shoot test (Time score)
- Target Shoot test (Time to fire)
- Target Shoot test (Log distance)
- Target Tracking 1 (Log distance)
- Target Tracking 2 (Log distance)
- Short-Term Memory test (Decision time)

Perceptual Speed/Accuracy Factor
- Short-Term Memory Test (Percent correct)
- Perceptual Speed & Accuracy test (Decision time)
- Perceptual Speed & Accuracy test (Percent correct)
- Target Identification test (Decision time)
- Target Identification test (Percent correct)

Number Speed/Accuracy Factor
- Number Memory test (Percent correct)
- Number Memory test (Initial decision time)
- Number Memory test (Mean operations decision time)
- Number Memory test (Final decision time)

General Reaction Speed Factor
- Choice Reaction Decision Time
- Simple Reaction Decision Time

General Reaction Accuracy Factor
- Choice Reaction Percent Correct
- Simple Reaction Percent Correct

From Non-Cognitive Inventories

Achievement Factor
- Self-Esteem scale
- Work Orientation scale
- Energy Level scale

Dependability Factor
- Conscientiousness scale
- Nondelinquency scale

Adjustment Factor
- Emotional Stability scale

Physical Condition Factor
- Physical Condition scale

Skilled Technician Interest Factor
- Clerical/Administrative
- Medical Services
- Leadership/Guidance
- Science/Chemical
- Data Processing
- Mathematics
- Electronics Communications

Structural/Machines Interest Factor
- Mechanics
- Heavy Construction
- Electronics
- Vehicle/Equipment operator

Combat-Related Interest Factor
- Combat
- Rugged Individualism
- Firearms Enthusiast

Audiovisual Arts Interest Factor
- Drafting
- Audiographics
- Aesthetics

Food Service Interest Factor
- Food Service - Professional
- Food Service - Employee

Protective Services Interest Factor
- Law Enforcement
- Fire Protection

Table 1.7
Latent Structure Scores

1. Core Technical Proficiency--MOS (Job) specific core technical skills: Represents the proficiency with which the individual performs the tasks that are "central" to his or her job (MOS). The tasks represent the core of the job and they are its primary definers from job to job.

   - The subscales representing core content in both the knowledge tests and the job sample tests that loaded on this factor were summed within method, standardized, and then added together for a total factor score. The factor score does not include any rating measures.

2. General Soldiering Proficiency--General or common skills: Reflects that, in addition to the core technical content specific to an MOS, individuals in every MOS are responsible for being able to perform a variety of general or common tasks -- e.g., use of basic weapons, first aid. This factor represents proficiency on these general tasks.

   - The same procedure (as for factor one) was used to sum the general task scales, standardize within methods, and add the two standardized scores.

3. Peer Support and Leadership, Effort, and Self Development: Reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers.

   - Five scales from the Army-wide BARS rating form (general technical performance, peer leadership, demonstrated effort, self-development, general maintenance), the expected combat performance scales, the job-specific BARS scales, and the total number of commendations and awards received by the individual were summed for this factor.

4. Maintaining Personal Discipline: Reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self control, demonstrates responsibility in day-to-day behavior, and does not create disciplinary problems.

   - Scores on this factor are composed of three Army-wide BARS scales (adherence to traditions and regulations, exercising self-control, demonstrating integrity), a subscale from the combat rating pertaining to avoidance of trouble, and two indices from the administrative records (disciplinary actions and promotions rate).

5. Physical Fitness and Military Bearing: Represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

   - Factor scores are the sum of the physical fitness qualification score from the individual's personnel record and the "military bearing and appearance" rating scale.

Concurrent Validation Results

The different criterion components were not predicted by the same things. Table 1.8 shows the multiple correlation of the components in these domains (corrected for shrinkage and for restriction of range, but not for unreliability) with the five criterion factors.

The entries in the table represent the average across all MOS. The level of validity of ASVAB for the first two factors is about the same as, or higher than, that usually observed when ASVAB is correlated with training criteria. ASVAB does predict job performance. For the third factor the validity of the cognitive tests drops, but is still substantial, and the validity of the non-cognitive temperament inventory increases. This reversal becomes even more distinct for factors 4 and 5. Notice that the interest scales are also a reasonably good predictor of task performance and do not predict factors 3, 4, and 5 as well as the temperament scales do.

Table 1.8
Mean Validity[a] for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs

	Predictor Domain					
Job Performance Construct	General Cognitive Ability (K=4)[b]	Spatial Ability (K=1)	Perceptual-Psychomotor Ability (K=6)	Temperament (K=4)	Vocational Interests (K=6)	Job Reward Preferences (K=3)
Core Technical Proficiency	.63	.56	.53	.25	.35	.29
General Soldiering Proficiency	.65	.63	.57	.25	.34	.30
Effort and Leadership	.31	.25	.26	.33	.24	.19
Personal Discipline	.16	.12	.32	.32	.13	.11
Physical Fitness and Military Bearing	.20	.10	.11	.37	.12	.11

[a]Validity coefficients were corrected for range restriction and adjusted for shrinkage.

[b]K is the number of predictor scores.

Incremental Validity

An important question for the Army is how to improve on the validity of decisions made using the Army's current selection and classification instrument, the ASVAB. To help answer that question, the validity of the General Cognitive Ability scores (computed from the ASVAB) was compared to the validity obtained when the scores from a predictor domain were used to supplement the General Cognitive Ability composite. This was done for each performance construct within each of the nine jobs. Validities were then averaged across the nine jobs. The resulting mean validities are reported in Table 1.9.

Table 1.9
Mean Incremental Validity[a] for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs

	Predictor Domain					
		General Cognitive Ability Plus:				
Job Performance Construct	General Cognitive Ability (K=4)[b]	Spatial Ability (K=5)	Perceptual Psychomoter Ability (K=10)	Temperament (K=8)	Vocational Interests (K=10)	Job Reward Preferences (K=7)
Core Technical Proficiency	.63	.65	.64	.63	.64	.63
General Soldiering Proficiency	.65	.68	.67	.66	.66	.66
Effort and Leadership	.31	.32	.32	.42	.35	.33
Personal Discipline	.16	.17	.17	.35	.19	.19
Physical Fitness and Military Bearing	.20	.22	.22	.41	.24	.22

[a] Validity coefficients were corrected for range restriction and adjusted for shrinkage. Incremental validity refers to the increase in R afforded by the new predictions above and beyond the R for the Army's current predictor battery, the ASVAB.

[b] K is the number of predictor scores.

Relative Contribution of Individual Predictors

Because there were virtually no predictor by MOS interactions, a stepwise multiple regression solution within each of the six categories of predictor constructs was computed on the combined samples from the nine MOS in Batch A for each of the last four Army-wide performance factors (i.e., General Soldiering, Effort/Leadership, Personal Discipline, and Physical Fitness/Military Bearing).

Some comparisons of interest are the following:

- Among ASVAB scores the quantitative and technical scores contribute the most to the prediction of General Soldiering Proficiency. The verbal score plays a more prominent role in predicting the Core Technical performance factor.

- While ASVAB does not contribute much to the prediction of performance factors 4 and 5, the ASVAB technical score does make a relatively large contribution to the prediction of factor 3, the Effort/Leadership factor.

- The differential contributions of the temperament (ABLE) scores to prediction of performance factors 3, 4, and 5 are clear, significant, and pronounced. The profiles look like they should.

- The combat interests score was the most predictive interest score among the scores generated from the AVOICE.

The profile of regression coefficients for predicting the Core Technical Proficiency factor was significantly different across MOS. The greatest differential is within the ASVAB and the AVOICE, and to a lesser extent within the spatial and computerized tests.

To look at the coefficients in another way, stepwise regressions were carried out with all 24 predictor scores used to predict each performance factor. The analyses for the four Army-wide criterion factors were carried out on a combined sample while the analyses against the Core Technical factor were done MOS by MOS. Again the differential patterns appear across the four Army-wide performance factors and across MOS for the Core Technical factor. However, a surprise was the strong role played by the spatial factor and the combat interest factor in predicting the technical performance factor in the combat specialties.

Weighting Criterion Components

The Concurrent Validation results indicated that each of the five criterion components can be predicted with considerable validity and that the validity of the different predictor domains varies systematically across criterion components. A subsequent focus was on the best method for obtaining their relative importance weights when the five components are combined into an overall composite index of performance.

Consequently, weighting judgments were systematically gathered from carefully chosen samples of both NCOs and officers familiar with each MOS.

The five Project A performance constructs received significantly different patterns of weights in different MOS and the different groups of experts agreed, in general, on the relative ranking of the weights. For example, the exercising leadership construct tends to be rated highest among the combat MOS.

Multiple judges per MOS, about 30 on the average, produced average rater reliabilities that are quite respectable (above .95 for most MOS). High intermethod correlations (about .95 on the average) between the construct weights obtained by a direct estimation method and a conjoint scaling method for the separate MOS further document the reliability of the mean weights.

Scaling the Utility of Individual Performance

The utility problem for Project A was one of assigning utility values to MOS by performance level combinations. That is, if it is true that personnel assignments will differ in value to the Army, depending on the specific MOS to which an assignment is made and on the level at which an individual will perform in that MOS, then the value of a classification strategy that has a validity significantly greater than zero will increase to the extent that the differential values (utilities) can be estimated and made a part of the assignment system.

The general procedure used to obtain utility scale values for different levels of predicted performance in each MOS used field grade officers as expert judges and was divided into three phases. Phase one was exploratory and used a series of workshop meetings with various officer groups to uncover the major issues. The goal of Phase two was to evaluate alternative expert judgment scaling methods and develop the procedure to be used. In Phase three the selected methods were used to obtain the final scale values.

Perhaps the most significant finding was that Army officers would be willing and able to assign differential utility values across MOS and performance levels. Perhaps the next most significant finding was that stable scale values could be obtained from averaging across a relatively small number of officer judges.

The analyses supported the conclusions that (a) for both methods the reliability of the average value produced by 11 judges or more is very high; (b) reliabilities are high even when performance level is controlled and differences are due only to MOS differences within performance level; (c) judges from different posts or MOS backgrounds do not produce different patterns of scale values; (d) within the limits of the methods used, the 1,365 MOS by performance level combinations have been placed on the same ratio scale of judged utility.

However, a number of problems need to be addressed before utilities similar to the ones obtained in Project A can be used operationally. One problem concerns the

optimal distribution within MOS, considering both within- and between-MOS utilities as well as the available recruit pool and the quality of existing personnel. This is the issue of average vs. marginal utility (Nord & White, 1988). Another issue concerns the duration of time that the recruits actually remain in the Army and how to aggregate values over time.

Second-Tour Performance Criterion Development

Over the course of its life cycle, Project A was able to complete the necessary job analyses and begin the criterion development work for the assessment of second-tour NCO performance for the Batch A MOS.

The specific goals of the job-analytic work were to:

- Describe the major differences between entry-level and second-tour performance content, within MOS.
- Describe the major differences across MOS, within the second-tour jobs.
- Describe the specific nature of the supervisory/leadership component of these higher level jobs.

Once these objectives were achieved, the information was used to address four questions:

(1) What should be the content of the new criterion measures?

(2) What kinds of measurement methods are needed?

(3) Are separate measures needed for each job? Or are the jobs so similar that the same measures can be applied to all?

(4) To what extent can measures developed for entry-level soldiers be used among higher level soldiers?

By Army policy, all soldiers are responsible for being able to perform all tasks at lower skill levels, as well as the tasks at their skill level. Because of these policies, the first-tour job analyses were used as a starting point and additional job analysis information was collected to describe the second-tour changes. In addition, the issue of leadership/supervision performance was of special concern.

To capture both the technical and the supervisory aspects of an MOS, four methods of job analysis were used: task analysis, a standardized questionnaire measure of supervisory/leadership responsibilities, critical incident analyses, and interviews with small groups of senior NCOs.

The information gathered via the different methods was summarized in a job analysis summary booklet for each MOS. Each booklet contains sections corresponding to (a) a table that shows estimates of the percentage of time per week that a second-tour soldier would spend on various kinds of activities under different conditions (field vs. garrison); (b) a description of the population of second-tour tasks, task clusters, and the major differences between first- and second-tour task content; (c) the second-tour critical incident analysis, including a discussion of the differences between the first-tour and second-tour dimensions; and (d) a comparison of task-based analysis, supervisory analysis, and behavior-based analysis results.

A separate job analysis book shows the results of the analysis of Army-wide critical incidents and describes the major commonalities and differences across MOS in terms of second-tour performance dimensions.

Given available resources, constraints on testing time, guidance from the literature, previous Project A work, and the second-tour job analysis results, a potential set of measurement methods was identified and reviewed by the project staff and the Scientific Advisory Committee. Some of the measurement methods had been used for the first tour and some were newly developed.

Briefly, the array of second-tour measures included the following:

(1) The original first-tour Army-wide behavioral and combat performance rating scales as modified on the basis of the second-tour job analysis.

(2) The MOS-specific rating scales as modified by the second-tour job analysis.

(3) Hands-on tests for 8-15 tasks for each MOS.

(4) Job knowledge tests for approximately 30 tasks for each MOS.

(5) A self-report measure of administration indexes, as modified by a reexamination of the records available for second-tour incumbents.

(6) A paper-and-pencil situational judgment test designed to measure knowledge of what action to take in a series of critical supervisory leadership situations.

(7) A role-play simulation involving the counseling of a soldier with a personal problem.

(8) A role-play simulation involving the counseling of a soldier with a performance problem.

(9) A role-play simulation of one-on-one remedial training.

(10) A set of rating scales designed to reflect the major dimensions of supervisory/leadership performance.

The above measures were administered to the 83/84 cohort second-tour followup sample in the fall and winter of 1988/89.

The Longitudinal Validation Data Collections

The Longitudinal Validation (LV) began with the administration of the Experimental Battery at the reception battalions to more than 49,000 accessions from the 86/87 cohort. These soldiers were then followed through their Advanced Individual Training or One Station Unit Training, where they were administered several criterion measures of performance during training. They were then followed into their first tour, where the job performance measures were administered.

Experimental Battery of Predictors

A summary of the predictor testing sites and the data collection period for each site is as follows:

Site	Predictor Testing Period
Fort Sill	20 Aug 86 - 20 Aug 87
Fort Benning	27 Aug 86 - 27 Aug 87
Fort Bliss	4 Sep 86 - 4 Sep 87
Fort Knox	10 Sep 86 - 10 Sep 87
Fort McClellan	17 Sep 86 - 17 Set 87
Fort Dix	24 Sep 86 - 24 Sep 87
Fort Leonard Wood	1 Oct 86 - 1 Oct 87
Fort Jackson	19 Nov 86 - 19 Nov 87

The complete array of tests and inventories in the Experimental Battery, the number of items in each, and the time limit (for the timed tests) or approximate time to finish (for the untimed inventories) are shown in Table 1.10.

The information obtained from CV data analysis was used to make the final revisions to the predictor battery for the Longitudinal Validation. Since the battery had already been through several iterations of data collection, analysis, and revision, the revisions were not substantial.

Training Performance Measures

Measures of training performance were collected on each individual at the end of AIT. The measures consisted of a number of the Army-wide BARS scales collected from the individual's drill instructor and the training achievement test previously developed for

Table 1.10
Description of Tests in Experimental Battery

	Number of Items	Time Limit (minutes)
Cognitive Paper-and-Pencil Tests		
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	36	18
	Number of Items	**Approximate Time**
Computer-Administered Tests		
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
	Number of Items	**Approximate Time**
Non-Cognitive Paper-and-Pencil Inventories		
Assessment of Background and Life Experiences (ABLE)	199	35
Army Vocational Interest Career Examination (AVOICE)	182	20
Job Orientation Blank (JOB)	31	5

each MOS. In order to cover the ASVAB aptitude area composites more comprehensively, two MOS were added to the Batch Z domain. They were 29E, Communications Electronics Repairer and 96B, Intelligence Specialist. In addition, 19E/19K was split into two distinct MOS for measures development, and 76W was dropped because it was redundant with 76Y. These changes resulted in 21 MOS in Project A.

Job Performance Measurement

The longitudinal data collection concluded with the administration of the performance measures to both first-tour and second-tour incumbents between July 1988 and February 1989. Data were collected from an estimated 11,300 first-tour soldiers and 1,050 second-tour soldiers at 13 CONUS installations and throughout USAREUR. The first-tour criterion measures were essentially the same as those used for the Concurrent Validation. The second-tour measures were the prototypes described in the previous section.

Summary

While analyses of the final longitudinal validation data for first-tour and second-tour soldiers were to be completed as part of the second-stage Career Force project, at this point it can be said that Project A had reached its basic goals:

- Multiple criterion measures had been developed and used to formulate five components of job performance.
- ASVAB was shown to be a highly valid predictor of job performance as reflected in the Core Technical performance and General Soldiering performance components.
- There was considerable differential prediction for the total test battery across the five performance components within each MOS.
- The non-cognitive predictors added significantly to the prediction of the "will do" components of performance and should prove to be valuable additions to the total system.
- As was expected, differential prediction across MOS was limited largely to the Core Technical performance factor. Both the ASVAB and the new experimental cognitive tests should contribute to differential prediction equations across major MOS clusters. However, the full analyses necessary to determine the prediction equations remained to be done.
- The importance weights for the five performance components had been scaled within each of the 19 MOS.

- Reliable scale values had been obtained for comparing the average utility of 1,365 MOS (273) by performance level (5) combinations.

- Comprehensive job analyses and prototypic criterion measures had been completed for second-tour NCO performance in nine MOS.

The results are impressive; however, for their full benefit to be realized a number of things must happen. Both the covariance structures and the estimates of predictive validity must be cross-validated with a genuine predictive design (i.e., the Longitudinal Validation); the rules for forming criterion composites must be developed; the marginal utility of accurate predictions must be estimated; valid measures of NCO performance must be constructed and an NCO performance model developed; the specifics of the full selection/classification/promotion decision system must be modeled; and the effects of using the new predictors in various combinations under a variety of goals and constraints must be evaluated.

The Foundation Provided by Project A

Project A was an innovative and ambitious undertaking. Contrary perhaps to even the most optimistic expectations, it met virtually all its objectives. The project received high praise from its Scientific Advisory Group, the National Academy of Sciences (NAS) Committee on the Performance of Military Personnel, and many others.

The first three years of the project were devoted largely to completing a comprehensive series of development steps. Much time, effort, and resources were devoted to predictor development, performance criterion development, and model development. It was indeed a very large initial investment in project R&D and the actual production of the major operational products was deferred. Another year and one half was devoted to planning and conducting the first major data collection (the Concurrent Validation) and completing the basic data analysis. During the remainder of the project, data analysis continued, utility values for performance outcomes were estimated, criterion component weights were obtained, and the Longitudinal Validation data collection was begun.

Although the time spent in the initial research and development phase created some delays in gratification, in retrospect it seems to have been a wise decision. This long development process can be thought of as a large initial investment that in the end produced a much larger return than if the objectives had been pursued in more piecemeal fashion and the Army had tried for more short-run gains. As shown by the preceding summary of Project A, some of the payoff was already evident. However, the major portion of the profit, particularly as it pertains to optimizing the entire selection/classification/promotion decision-making procedure, was realized under Career Force.

For example, during the development phase of Project A, a 4-hour battery of new selection/classification tests was constructed so as to sample systematically the most relevant applicant characteristics not presently covered by ASVAB. Also during the

development phase, a 12-hour training achievement and job performance measurement procedure was constructed to provide multiple measures of every major component of performance for each job in a representative sample of entry-level MOS.

These MOS were sampled representatively from the population of entry-level MOS. Consequently, for (a) jobs, (b) performance components, and (c) selection/classification measures, a population had been defined and then sampled systematically. This makes the results of the Concurrent and Longitudinal Validations generalizable and extremely useful for guiding future selection/classification practices. A wide variety of comparative, "what if" questions can be asked about the differential prediction (by different kinds of test information) of each major performance component under varying sets of constraints, and the answers generalized to the entire system. No other organization in the world, public or private, has such an extensive, carefully developed, and generalizable body of information with which to build and evaluate future selection, assignment, and promotion strategies. It can be used for many years to come.

In addition to developing a comprehensive battery of selection/ classification tests and a full multiple-method array of first-tour performance measures and using them to generate the most extensive data base in the history of personnel research, Project A resulted in a number of both scientific and applied products. These are summarized below.

Project A Products and Results

The products are of two general kinds: products for the "science" (personnel research) and products for the organization (the Army). The list is intended to move from the scientific to the applied. However, the distinction is not always easy to make because many products are useful for both.

(1) There exist, in technical report form, comprehensive reviews of all validity evidence pertaining to selection and classification for skilled jobs. These are the most comprehensive such reviews ever done.

(2) As a by-product of the analyses involving ASVAB, there exists a much clearer idea of its factor structure, of what the factors are measuring, and of the nature of its strengths and limitations.

(3) A set of new experimental tests has been developed to measure non-cognitive, psychomotor, perceptual, and cognitive characteristics that are not now measured by the ASVAB. The scope of the project made it possible to examine virtually the entire domain of selection information, sample from it, and investigate the basic incremental validity produced by each major piece of information.

(4) The results of an expert judgment study of expected correlations between predictor constructs and performance factors are available. In brief, a large sample of personnel experts considered the population of predictor and

criterion variables appropriate for entry-level jobs and forecasted what the validity coefficients would be. The consistency in the judgments and their correspondence with known data points make these a potentially valuable tool for future test selection and synthetic validation work.

(5) Much has been learned about the nature of performance in entry- level skilled jobs (e.g., first-tour MOS). We now have a much clearer idea of what major factors constitute performance and how they can be measured. The "criterion problem" is better understood. This knowledge should better inform future enlistment and promotion policy, as well as future personnel research.

(6) The Concurrent Validation data support the assertion that supervisor ratings of subordinate performance have considerable construct validity if a careful measurement procedure is followed. The data also support the conclusion that supervisors seem to assess both the technical performance of individuals and their general dependability/motivation at the same time. Apparently, when raters rate the aggregate performance of subordinates, the "can do" and "will do" aspects cannot be separated and one cannot be rated while the other is "controlled" in some fashion. If the effectiveness of an individual's volitional behaviors are consistent across tasks and this consistency is accurately assessed, then perhaps halo error has been explained.

(7) Using much more comprehensive samples than ever before, new ASVAB Aptitude Area composites have been developed which are firmly data based and empirically defensible.

(8) The question of whether ASVAB does or does not predict job performance (in addition to training performance) has been answered definitively, in the affirmative. The Army and DoD are now in a better position to support their quality goals. In addition, it is now known what aspects of performance ASVAB predicts best and which aspects of performance could be predicted better with other types of selection instruments.

(9) Within the limits of the Concurrent Validation design, the incremental validity of appropriate ABLE scales for predicting the "will do" components of performance has been demonstrated.

(10) The potential of the AVOICE for differentially predicting "can do" performance in combat vs. technical vs. administrative support MOS has been established. What is needed to make this finding operational is empirical scoring keys.

(11) The Project A job/task analysis procedures worked well and can be used by the Army in the future to develop training curricula, test content, performance measures, and field exercises. The job analysis summaries for each MOS serve as a model for future job analysis work in the Army as well as in the public and private sector.

(13) AIT training achievement measures have been developed for 21 MOS. The training measures will allow a determination of whether training performance predicts job performance, and whether it does so differentially for different groups of trainees (race, gender), and different groups of MOS (combat, combat support, combat service support).

(14) The package of rating scale administration procedures can be used in future personnel research in the Army. A major effort in the Project A research was to develop an effective and very efficient set of procedures for administering performance rating scales to large numbers of people. These procedures and the package of materials can be adapted for use in other Army personnel research where ratings of many persons are required.

(15) The Supervisory Description Questionnaire (which came out of second-tour job analyses work) is a very useful instrument for future work in the design of leadership training or the evaluation of leadership/supervisor performance. The questionnaire is based on a clear rationale and is straightforward to use.

(16) The Project A performance measures, against which new selection/ classification decision procedures will be calibrated, have been demonstrated not to be discriminatory. The Project A samples that examined the interactions of rater and ratee race exceed the magnitude of the combined samples from all previous research on this issue.

(17) Project A developed a common utility scale for making comparisons across MOS and performance levels within MOS. Although it does not speak to marginal utility issues, it can be used to enhance the comparison of alternative selection/classification procedures.

(18) One very real, and very important product, is the Project A data base itself. It is by orders of magnitude the largest and most completely documented personnel research data base in existence.

A SUMMARY OF CAREER FORCE

The Career Force project spanned a period of five years corresponding to FY90, FY91, FY92, FY93, and FY94. A detailed year by year summary of the first four years is given below; this work is more fully reported in the project annual reports (Campbell & Zook, 1990, 1994a, 1994b, 1994c). The work completed during the last year of the project is presented in Chapters 2-9 of the current report.

SUMMARY OF PROJECT EFFORTS FOR YEAR ONE

As described in the first Career Force annual report (Campbell & Zook, 1990), the objectives of the project's first year were focused on developing a full design for the data base and on analyzing basic scores for (a) the final version of the Experimental Predictor Battery (EB), (b) the End-of-Training (EOT) performance measures, and (c) the second-tour criterion measures used to assess NCO performance in the second-tour Concurrent Validation (CVII) sample. The data from the End-of-Training and second-tour Concurrent Validation (CVII) performance assessments were also used to formulate both a model of training performance and a model of second-tour (junior NCO) job performance. That is, the basic scores from the individual performance measures were aggregated into factor scores that represented, as well as possible, the major components, or latent structure, of training performance and second-tour job performance.

By the end of year one, the data collection for the Longitudinal Validation first-tour performance assessments had been completed, but the data cleaning and editing were still in progress and the analysis of the LVI performance measures had not yet begun.

Data Base Design

As described in the first-year annual report, the Career Force data base design allows access at any level of score aggregation. The report describes each variable and the amount of information that is available. The data are accessed via a secure system that requires prior approval by the Army. The data base also includes data from several operational files maintained by the Army.

Basic Scores for the Experimental Battery

During year one, much effort was devoted to analyzing the data that had been obtained by administering the Experimental Predictor Battery to almost 50,000 new accessions in the Longitudinal Validation sample. A number of data editing procedures were compared and evaluated, and great care was taken to maximize data quality for the information that was entered into the final data file. The psychometric properties and subgroup differences for each measure were analyzed, and a series of exploratory and confirmatory analyses were conducted to identify the basic predictor scores within each domain that would be used in the validation analyses.

The final array of tests in the Experimental Battery and the constructs they are intended to measure are shown in Figure 1.5. The 31 basic scores that are obtained from the specific test indicators are shown in Figure 1.6 (Campbell & Zook, 1990).

There was a very high degree of consistency between the Concurrent Validation and the Longitudinal Validation in terms of the factor structures of the various measures. The resulting definitions of the basic predictor scores to be used in the validation analyses were quite similar.

Test/Measure	Construct
Paper-and-Pencil Spatial Tests	
Assembling Objects	Spatial Visualization-Rotation
Object Rotation	Spatial Visualization-Rotation
Maze	Spatial Visualization-Scanning
Orientation	Spatial Orientation
Map	Spatial Orientation
Reasoning	Induction
Computer-Administered Tests	
Simple Reaction Time	Reaction Time (Processing Efficiency)
Choice Reaction Time	Reaction Time (Processing Efficiency)
Short-Term Memory	Short-Term Memory
Perceptual Speed and Accuracy	Perceptual Speed and Accuracy
Target Identification	Perceptual Speed and Accuracy
Target Tracking 1	Psychomotor Precision
Target Shoot	Psychomotor Precision
Target Tracking 2	Multilimb Coordination
Number Memory	Number Operations
Cannon Shoot	Movement Judgment
Temperament, Interest, and Job Preference Measures	
Assessment of Background and Life Experiences (ABLE)	Adjustment
	Dependability
	Achievement
	Physical Condition
	Leadership (Potency)
	Locus of Control
	Agreeableness/Likability
Army Vocational Interest Career Examination (AVOICE)	Realistic Interest
	Conventional Interest
	Social Interest
	Investigative Interest
	Enterprising Interest
	Artistic Interest
Job Orientation Blank (JOB)	Job Security
	Serving Others
	Autonomy
	Routine Work
	Ambition Achievement

Figure 1.5. Experimental Predictor Battery tests and relevant constructs.

ASVAB Factor Composites

- Quantitative
  - Mathematics Knowledge
  - Arithmetic Reasoning
- Technical
  - Auto/Shop Information
  - Mechanical Comprehension
  - Electronics Information
- Speed
  - Coding Speed
  - Number Operations
- Verbal
  - Word Knowledge
  - Paragraph Comprehension
  - General Science

Paper-and-Pencil Test Composite

- Spatial
  - Assembling Objects Test
  - Object Rotation Test
  - Maze Test
  - Orientation Test
  - Map Test
  - Reasoning Test

Computer-Administered Test Composites*

- Psychomotor
  - Target Tracking 1 Distance
  - Target Tracking 2 Distance
  - Cannon Shoot Time Score
  - Target Shoot Distance
- Movement Time
  - Pooled Movement Time
- Perceptual Speed
  - Perceptual Speed & Accuracy (DT)
  - Target Identification (DT)
- Basic Speed
  - Simple Reaction Time (DT)
  - Choice Reaction Time (DT)
- Perceptual Accuracy
  - Perceptual Speed & Accuracy (PC)
  - Target Identification (PC)
- Basic Accuracy
  - Simple Reaction Time (PC)
  - Choice Reaction Time (PC)
- Number Speed and Accuracy
  - Number Memory (Operation DT)
  - Number Memory (PC)
- Short-Term Memory
  - Short-Term Memory (PC)
  - Short-Term Memory (DT)

ABLE Composites

- Achievement Orientation
  - Self-Esteem
  - Work Orientation
  - Energy Level
- Leadership Potential
  - Dominance
- Dependability
  - Traditional Values
  - Conscientiousness
  - Nondelinquency
- Adjustment
  - Emotional Stability
- Cooperativeness
  - Cooperativeness
- Internal Control
  - Internal Control
- Physical Condition
  - Physical Condition

JOB Composites

- High Job Expectations
  - Pride
  - Job Security
  - Serving Others
  - Ambition
- Job Routine
  - Routine
- Job Autonomy
  - Autonomy

AVOICE Composites

- Rugged/Outdoors
  - Combat
  - Rugged Individualism
  - Firearms Enthusiast
- Audiovisual Arts
  - Drafting
  - Audiographics
  - Aesthetics
- Interpersonal
  - Medical Services
  - Leadership/Guidance
- Skilled/Technical
  - Science/Chemical
  - Computers
  - Mathematics
  - Electronic Communication
- Administrative
  - Clerical/Administrative
  - Warehousing/Shipping
- Food Service
  - Food Service - Professional
  - Food Service - Employee
- Protective Services
  - Fire Protection
  - Law Enforcement
- Structural/Machines
  - Mechanics
  - Heavy Construction
  - Electronics
  - Vehicle Operator

*DT - Decision Time and PC - Proportion Correct

Figure 1.6. Longitudinal Validation Experimental Battery: Composite scores and constituent basic scores.

Basic Scores for the End-of-Training Measures

During year one, the data from the school knowledge test and seven training performance rating scales administered at the end of training were analyzed in terms of their psychometric properties and factor structure. Confirmatory techniques were used to identify the "model" of training performance that best represented the covariances among the observed measures. That is, an a priori set of alternative models was proposed and evaluated in terms of the degree to which they fit the data. In the end six basic scores were proposed, two based on the knowledge tests and four based on the rating scales. A brief characterization of the six scores is given in Figure 1.7 (Campbell & Zook, 1990).

These six scores will serve both as criterion measures (for the Experimental Battery) and as predictors (of first-tour and second-tour job performance) in later validation analyses.

Development of Second-Tour Performance Scores (CVII)

The performance measures used in the CVII sample, and their development, have been described in detail in previous Project A and Career Force reports (e.g., Campbell, 1991; Campbell & Zook, 1991). First-tour measures were revised for use with second-tour personnel and new measures reflecting the unique components of second-tour jobs were added. A summary description of the specific measures is given below.

Rating Scales

On the basis of second-tour critical incident analyses, the Army-wide Behaviorally Anchored Ratings Scales (BARS) and MOS-specific BARS were revised and scales having to do with leadership and supervision were added. Further, based on job analysis data, seven new scales pertaining to supervision and leadership responsibilities were also added. A full list of the Army-wide rating scales is shown below. Not shown are the MOS BARS for each MOS, which were revised to reflect second-tour performance demands, and the Combat Performance Prediction Scales, which were the same as those used in LVI, and which were not administered to female NCOs.

Army-Wide Behavior Scales:

1. Demonstrating Technical Knowledge and Skill
2. Demonstrating Effort
3. Supervising Subordinates
4. Following Regulations and Orders
5. Demonstrating Integrity
6. Training and Development of Subordinates
7. Maintaining Equipment
8. Physical Fitness
9. Self-Development
10. Showing Consideration for Subordinates
11. Demonstrating Appropriate Military Bearing
12. Demonstrating Appropriate Self-Control

EOT RATING SCALE BASED SCORES

1) **Effort and Technical Skill (ETS)**

Technical Knowledge/Skill:	How effective is each soldier in acquiring job/soldiering knowledge and skill?
Effort:	How effective is each soldier in displaying extra effort?

2) **Maintaining Personal Discipline (MPD)**

Following Regulations and Orders:	How effective is each soldier in adhering to regulations, orders, and Standard Operating Procedures (SOP) and displaying respect for superiors?
Self Control:	How effective is each soldier in controlling own behavior related to aggressive acts?

3) **Physical Fitness and Military Bearing (PFB)**

Military Appearance:	How effective is each soldier in maintaining proper military appearance?
Physical Fitness:	How effective is each soldier in maintaining military standards of physical fitness?

4) **Leadership Potential (LEAD)**

Leadership Potential:	Evaluate each soldier on his or her potential effectiveness as a leader. Do not necessarily rate on the basis of present performance.

EOT KNOWLEDGE TEST BASED SCORES

5) **Basic Knowledge Score:**	Items measuring knowledge requirements common to all MOS.
6) **Technical Knowledge Score:**	Items measuring technical knowledge requirements specific to each MOS.

Figure 1.7. Composite scores that reflect end-of-training (EOT) performance factors.

Additional Leadership Scales:

13. Serving as a Role Model
14. Communication With Subordinates
15. Personal Counseling
16. Monitoring Subordinate Performance
17. Organizing Missions/Operations
18. Personnel Administration
19. Performance Counseling

General Scales:

20. Overall Effectiveness
21. Senior NCO Potential

Situational Judgment Test (SJT)

A new paper-and-pencil measure of supervisory judgment was developed by describing prototypical judgment situations and asking the respondent to select the most appropriate and the least appropriate courses of action. The situation descriptions and the scoring keys were refined through extensive subject matter expert judgments.

Supervisory Simulation Exercises

These measures were developed to assess NCO performance in job areas that were judged to be best assessed through the use of interactive exercises. The simulations were designed to evaluate performance in counseling and training subordinates. A trained evaluator (role player) played the part of a subordinate to be counseled or trained and the examinee assumed the role of a first-line supervisor who was to conduct the counseling or training. Evaluators also scored the examinee's performance, using a standard set of rating scales.

Here are brief descriptions of the three simulation exercises:

- Personal Counseling Simulation: A private first class (PFC) is exhibiting declining job performance and personal appearance. Recently, the PFC's wall locker was left unsecured. The supervisor has decided to counsel the PFC about these matters.

- Disciplinary Counseling Simulation: There is convincing evidence that the PFC lied to get out of coming to work today. The PFC has arrived late to work on several occasions and has been counseled for lying in the past. The PFC has been instructed to report to the supervisor's office immediately.

- Training Simulation: The commander will be observing the unit practice formation in 30 minutes. The private, although highly motivated, is experiencing problems with the hand salute and about face.

For each exercise, examinee performance was evaluated on 3-point rating scales reflecting specific behaviors tapped by the exercises and a 5-point overall effectiveness rating scale. Factor analyses of the ratings data suggested that each simulation could be scored in terms of the content of the NCO's behavior (i.e., did he or she do or say the right things) and the process, or style, with which the counseling steps were carried out.

Administrative Measures

The self-report Personnel File Form (PFF) used in LVI was modified for use with second tour and six administrative indices of performance were obtained.

Job Knowledge and Hands-On Measures

The content of each of these measures was revised on the basis of the second-tour job analyses and the revised instruments were subjected to extensive SME review. Analyses of alternative aggregations of item and scale scores from both of these measures resulted in the adoption of a general (Army-wide) and an MOS-specific score for each of them.

Final Array of Second-Tour Basic Performance Scores

After extensive analyses of their psychometric properties and factor structures, based on CVII data, the final array of basic second-tour performance scores was as shown in Figure 1.8. There were 22 basic scores. Scores from this array became the basis for the second-tour performance modeling analysis in CVII.

Development of the CVII Second-Tour Performance Model

The basic CVII performance scores served as input to the development of a latent structure model for second-tour performance (Campbell & Zook, 1990). Based on a consensus of the project staff, three major alternatives could be used to explain the observed correlations. Consequently, the competing models that were evaluated for comparative goodness of fit, using the LISREL VI program (Jöreskog & Sörbom, 1986), were the following:

(1) First-Tour Model: Included five substantive and two methods factors, with the SJT and Simulation variables all loading on the Effort and Leadership factor.

(2) Leadership Factor Model: Included a sixth substantive factor with the SJT, Simulation, and Leadership Rating factor variables all loading on this factor. This model was evaluated with and without a separate simulation "methods" factor.

(3) Training and Counseling Factor Model: Included a sixth substantive factor with just the Simulation variables. No separate simulation methods factor could be estimated under this model.

Hands-On Performance Test

1. MOS-specific task performance score
2. General (common) task performance score

Job Knowledge Test

3. MOS-specific task knowledge score
4. General (common) task knowledge score

Army-Wide Rating Scales

5. Leadership/supervision composite
6. Technical skill and effort composite
7. Personal discipline composite
8. Physical fitness and military bearing composite

MOS-Specific Rating Scales

9. Overall MOS composite

Combat Performance Prediction Scales

10. Overall Combat Prediction scale composite (available for males only)

Personnel File Form

11. Awards and Certificates
12. Articles 15/Flag Actions (Disciplinary Actions)
13. Physical Readiness
14. M16/M19 Qualification
15. Military Training Courses
16. Promotion Rate

Situational Judgment Test

17. Total score obtained by subtracting the total "ineffectiveness" score from the total "effectiveness" score

Supervisory Simulation Exercises

18. Personal Counseling: Process
19. Personal Counseling: Content
20. Disciplary Counseling: Process
21. Disciplary Counseling: Content
22. Training: Total composite score

Figure 1.8. Summary list of CVII basic criterion scores.

Of the three models, the Training and Counseling Factor Model provided the closest fit to the observed data. A result of considerable interest was that the SJT (a paper-and-pencil measure) fit best with the Effort and Leadership factor, in spite of the method variance involved.

The basic scores that have been used to represent the latent variables are as shown in Figure 1.9. For validation analysis purposes, the six substantive factor scores are obtained by standardizing and summing the basic scores within each factor.

SUMMARY OF PROJECT EFFORTS FOR YEAR TWO

As described in the annual report for the 1991 fiscal year (Campbell & Zook, 1994a), year two was a period of score development, model building, and basic validation analyses for (a) training performance (EOT), (b) first-tour performance (LVI), and (c) second-tour performance (CVII). During year two, the second-tour longitudinal data collection (LVII) began and was ongoing.

Objectives

The specific objectives for the second-year annual report were as follows:

(1) Describe the development of alternative scores for the Assessment of Background and Life Experiences (ABLE) instrument.

(2) Describe the basic validation analyses for the prediction of performance in training.

(3) Describe the development of basic scores for the longitudinal sample first-tour performance measures.

(4) Describe the replication/confirmation of the first-tour performance model and the basic Longitudinal Validation analyses for the Experimental Predictor Battery against first-tour performance.

(5) Describe the basic validation analyses for the prediction of second-tour performance, using the CVII sample.

(6) Report the results of a preliminary analysis of the prediction of second-tour performance from first-tour predictors and performance.

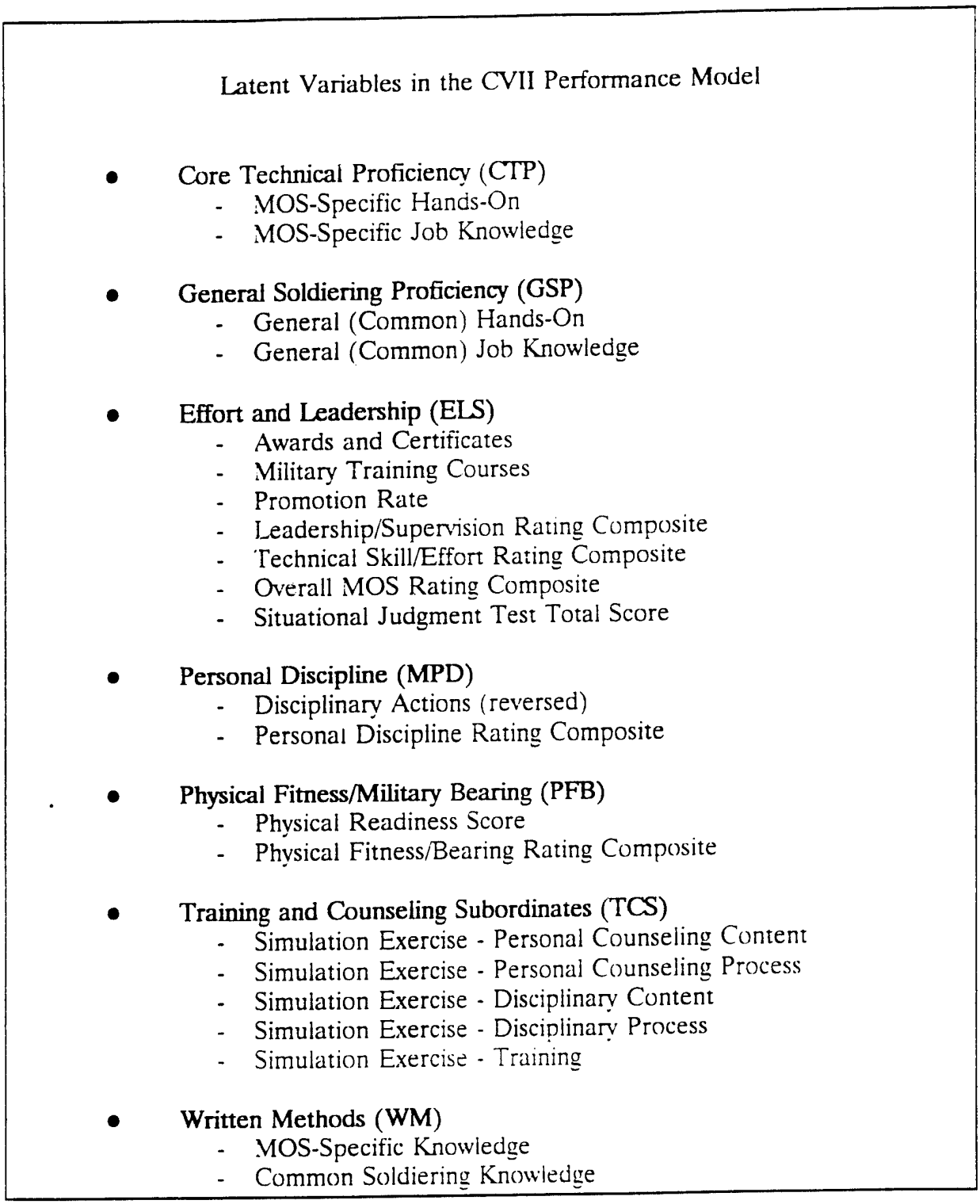


Figure 1.9. Relationship of specific variables to overall factors in the CVII performance model.

Development of Alternative ABLE Factor Composites

As part of Project A, and based on the results of an extensive review of the literature, 10 temperament scales had been developed to form the ABLE. These constructs were selected as the most promising for predicting performance in Army enlisted occupational specialties. In addition, four validity scales were included to detect inaccuracies in self-reports of temperament and a self-report measure of physical condition was also included (see Hough, Eaton, Dunnette, Kamp, & McCloy, 1990, for more information on the development of ABLE). To develop a set of conceptually meaningful construct (composite) scores, Peterson et al. (1990) carried out both exploratory and confirmatory factor analyses on the correlation among the content scale scores.

The resulting seven temperament constructs (composites) and associated ABLE scales are shown in Table 1.11. The constructs of Dependability, Dominance (Surgency), Adjustment, and Cooperativeness have counterparts in the Big Five personality dimensions described by Norman (1963) and Goldberg (1981). Conversely, Achievement and Internal Control are not in the Big Five taxonomy, but were among the strongest predictors of job performance in the Project A review of the temperament domain (see Hough, 1992, for more details on the relationship of ABLE constructs to the Big Five).

Table 1.11
ABLE Rational Composites and Corresponding Content Scales

Composite	ABLE Scale
Achievement Orientation	Self-Esteem Work Orientation Energy Level
Leadership Potential	Dominance
Dependability	Traditional Values Conscientiousness Nondelinquency
Adjustment	Emotional Stability
Cooperativeness	Cooperativeness
Internal Control	Internal Control
Physical Condition	Physical Condition

As noted above, a rational/theoretical approach was the primary method used in developing ABLE. An alternative empirical procedure emphasizes the internal covariance structure of a set of items and uses factor analytic methods. Consequently, during year two, internal scale construction methods were used to increase, through homogeneous keying, the internal consistency of ABLE composites and to decrease their intercorrelations.

Results from factor analyses of 199 items were used to form seven preliminary composites. These composites contained 99 items. Next, correlations between the remaining content-type items (excluding the validity scale items) and the preliminary factor composites were examined and each remaining item was assigned to the composite with which it had the highest correlation. The seven factor composites resulting from this procedure used 168 items and are called the ABLE-168 composites. In all, 125 items were assigned in the same way on the ABLE-168 composites and the ABLE rational composites.

As a second alternative, an item was retained only if it correlated at least .33 with the scale for which it was assigned and had a higher correlation with its own composite (by .03) than any other. In addition, several items that added only minimally to internal consistency were dropped. The resulting set of composites had a total of 114 items and is called the ABLE-114 composites. Eighty-nine of these items were assigned in the same way on ABLE-114 and the ABLE rational composites.

The three scoring methods converged to yield seven similar temperament constructs. The composites measuring the same constructs were very highly correlated ($\underline{r}$ = .88 to 1.0).

ABLE-114 composites had greater discriminant validity than either the ABLE-168 factor composites or the ABLE rational composites. The average correlation among the composites (off-diagonal elements) was .40 for ABLE-114, and .47 for the ABLE rational composites and ABLE-168.

Table 1.12 shows the distribution of items on ABLE-168 and ABLE-114 for each of the ABLE content scales. Items outside the shaded areas were assigned differently on the rational and factor composites. There is much overlap between the rational and factor composites. However, approximately 25 percent of item assignments for the factor composites were different from those used for the rational composites. Most of these are consistent with results from previous research and/or can be understood on the basis of item content.

In sum, there are three alternative sets of ABLE composites measuring seven temperament constructs. The 114-item form is shorter and has higher discriminant validity than the other two sets of composites, with little apparent loss of reliability. Subsequent analyses in the Career Force Project examine the criterion-related validities of these alternative sets of composites.

Table 1.12
Distribution of ABLE Scale Items on ABLE-168 and ABLE-114 Factor Composites

ABLE Scale	No. of Items	ABLE Factor Composite						
		Achievement Orientation	Leadership Potential	Dependability	Adjustment	Cooperativeness	Internal Control	Physical Condition
Self-Esteem	12 (6)		10 (6)		2 (0)			
Work Orientation	19 (15)	18 (14)	1 (1)					
Energy Level	21 (9)	13 (6)			6 (1)			2 (2)
Dominance	12 (12)		12 (12)					
Traditional Values	11 (7)			5 (4)			5 (3)	
Conscientiousness	15 (11)	9 (8)		6 (3)				
Nondelinquency	20 (13)			20 (13)				
Emotional Stability	17 (11)				17 (11)			
Cooperativeness	18 (11)			2 (1)		16 (10)		
Internal Control	16 (11)	2 (0)			2 (1)		12 (10)	
Physical Condition	6 (6)							6 (6)
Poor Impression	2 (2)				2 (2)			
Total	169 (114)	42 (28)	23 (19)	33 (21)	29 (15)	16 (10)	17 (13)	8 (8)

Note. ABLE 114 items are shown in parentheses. Shaded areas indicate convergence between the rational and factor composites.

Prediction of Performance in Training

The objectives of analyses of the end-of-training (EOT) data were to:

(1) Compute the validities for ASVAB and Experimental Battery predictors against rating measures and also paper-and-pencil test measures of training performance.

(2) Compare the validities of four alternative sets of ASVAB scores.

(3) Compare the validities of three alternative sets of ABLE scores.

(4) Assess the incremental validities for the Experimental Battery predictors over ASVAB.

Procedure

The EOT validation analysis consisted of the following steps:

A) Multiple correlations between each set of predictor scores and each set of criterion scores were computed separately by MOS and then averaged across the Batch A MOS and across all MOS.

1) The ASVAB predictor set was represented by:

a) The nine ASVAB subtest scores
b) The four ASVAB factor scores
c) The Armed Forces Qualification Test (AFQT)
d) The MOS-appropriate Aptitude Area composite score

2) The ABLE predictor set was represented by three sets of scores:

a) The seven rational scales
b) Seven empirical scales that retained 168 items
c) Seven empirical scales that retained only 114 items

3) Each of the other predictor sets (i.e., spatial, computer, AVOICE, JOB) was represented as in previous analyses.

All results were adjusted for shrinkage and corrected for multivariate range restriction.

B) Incremental validity was computed for each set of Experimental Battery predictors over the ASVAB.

C) Multiple correlations were computed between each set of predictor scores and a "Peer 1" rating, a "Peer 2" rating, a supervisor rating, and various combinations.

Results

To summarize the principal findings, multiple correlations for six predictor sets are shown in Table 1.13; the incremental validities are summarized in Table 1.14. In general, ASVAB shows high validity against the school knowledge measures and the relative validities for the four ratings factors are as would be expected on the basis of the factors. The ABLE does not predict the "will do" factors quite as well as it did in CVI but it predicts the "can do" factors somewhat better.

These results indicate that the level of validity of the ASVAB factors for predicting the School Knowledge (SK) test scores was extremely high, especially for the Technical (SK-Tech) and Total (SK-Total) scores. Likewise, the spatial composite and the computer battery produced high validities for these criteria.

Results from other analyses indicate that peer ratings of training performance are more accurately predicted than supervisor ratings of training performance. This suggests that peer ratings may be more valid training measures than supervisor ratings, presumably because, in training, peers generally have greater opportunity to observe ratees than do supervisors. This comparison is confounded, however, by the greater reliability of the peer ratings that is, at least in part, due to the fact that they are based on more raters per ratee than are the supervisor ratings. Yet analyses at the 1-rater level corroborate the notion that the peer ratings have more utility than the supervisor ratings for assessing training performance.

Further analysis showed that the average multiple correlations for the four different sets of ASVAB scores differed only slightly in validity, except that the peer ratings of Physical Fitness (PFB) were better predicted by the nine subtests and the four factors. However, the school knowledge test scores were predicted somewhat better (about three to five points) by the ASVAB subtests and factors than by the AFQT or Aptitude Area composites.

Both ABLE and AVOICE predicted the knowledge-based scores quite well. The largest incremental validities were for ABLE over ASVAB when predicting Personal Discipline, Fitness and Bearing, and Leadership.

Finally, there were virtually no differences in validities for the three alternative sets of ABLE scores although the ABLE-114 validities were consistently slightly higher.

Table 1.13
Mean of Multiple Correlations Computed Within Job for End-of-Training Sample for ASVAB Factors, Spatial, Computer, JOB, ABLE Rational Composites, and AVOICE

Criterion[a]	MOS	No. of MOS[b]	ASVAB Factors [4]	Spatial [1]	Computer [8]	JOB [3]	ABLE Comp. [7]	AVOICE [8]
Peer-ETS	Batch A	11	41 (07)	35 (05)	36 (05)	24 (06)	19 (09)	22 (07)
	All MOS	22	43 (13)	37 (10)	33 (14)	23 (11)	23 (12)	23 (10)
Peer-MPD	Batch A	11	25 (04)	22 (05)	21 (05)	09 (07)	19 (05)	11 (07)
	All MOS	22	26 (11)	22 (08)	15 (10)	12 (10)	22 (10)	09 (09)
Peer-PFB	Batch A	11	14 (09)	05 (06)	11 (05)	05 (05)	29 (06)	07 (07)
	All MOS	22	19 (14)	10 (11)	12 (09)	09 (12)	26 (11)	10 (10)
Peer-LEAD	Batch A	11	30 (10)	24 (07)	28 (07)	18 (09)	22 (09)	17 (10)
	All MOS	22	30 (16)	26 (12)	25 (16)	20 (14)	22 (12)	16 (14)
Supv-ETS	Batch A	11	21 (06)	18 (05)	17 (10)	10 (08)	09 (10)	11 (10)
	All MOS	22	27 (15)	22 (11)	18 (13)	10 (10)	11 (12)	10 (10)
Supv-MPD	Batch A	11	13 (09)	12 (07)	11 (08)	06 (06)	05 (06)	06 (06)
	All MOS	22	16 (16)	14 (11)	10 (13)	06 (08)	05 (07)	04 (06)
Supv-PFB	Batch A	11	11 (07)	09 (05)	09 (08)	06 (05)	11 (09)	07 (07)
	All MOS	22	16 (15)	13 (12)	11 (15)	05 (07)	11 (11)	05 (06)
Supv-LEAD	Batch A	11	15 (10)	14 (08)	13 (10)	08 (08)	10 (11)	08 (09)
	All MOS	22	19 (17)	17 (11)	12 (12)	11 (09)	11 (12)	07 (09)
SK-Basic	Batch A	9	68 (06)	57 (06)	57 (06)	38 (05)	30 (07)	37 (05)
	All MOS	20	67 (08)	58 (07)	55 (14)	36 (10)	31 (14)	37 (11)
SK-Tech	Batch A	11	76 (05)	63 (05)	61 (05)	41 (07)	33 (05)	44 (07)
	All MOS	22	75 (06)	62 (08)	59 (06)	38 (11)	33 (13)	40 (12)
SK-Total	Batch A	11	78 (03)	65 (04)	64 (03)	43 (07)	34 (05)	45 (06)
	All MOS	22	77 (05)	65 (07)	62 (07)	40 (11)	35 (14)	42 (13)

Note: Corrected for range restriction and adjusted for shrinkage (Rozeboom, 1978, formula 8). Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Decimals omitted.

[a] ETS = Effort and Technical Skill; MPD = Maintaining Personal Discipline; PFB = Physical Fitness and Military Bearing; LEAD = Leadership Potential; SK = School Knowledge.

[b] Number of MOS for which validities were computed.

Table 1.14
Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for End-of-Training Sample for Spatial, Computer, JOB, ABLE Rational Composites, and AVOICE

Criterion[a]	MOS	No. of MOS[b]	A4 ASVAB Factors [4]	A4+ Spatial [5]	A4+ Computer [12]	A4+ JOB [7]	A4+ ABLE Comp. [11]	A4+ AVOICE [12]
Peer-ETS	Batch A	11	41 (07)	42 (07)	42 (06)	41 (07)	44 (06)	41 (07)
	All MOS	22	43 (13)	42 (14)	40 (16)	42 (13)	45 (11)	41 (14)
Peer-MPD	Batch A	11	25 (04)	25 (05)	24 (05)	25 (05)	34 (06)	24 (07)
	All MOS	22	26 (11)	25 (11)	22 (12)	25 (12)	33 (11)	22 (11)
Peer-PFB	Batch A	11	14 (09)	13 (09)	17 (07)	15 (09)	31 (09)	15 (09)
	All MOS	22	19 (14)	18 (14)	16 (12)	19 (17)	30 (14)	18 (11)
Peer-LEAD	Batch A	11	30 (10)	30 (10)	31 (08)	30 (11)	35 (09)	29 (13)
	All MOS	22	30 (16)	30 (17)	28 (18)	31 (18)	34 (15)	28 (18)
Supv-ETS	Batch A	11	21 (06)	21 (07)	19 (09)	20 (06)	19 (12)	17 (12)
	All MOS	22	27 (15)	26 (15)	24 (15)	25 (15)	25 (19)	22 (16)
Supv-MPD	Batch A	11	13 (09)	12 (09)	11 (09)	11 (09)	13 (11)	11 (10)
	All MOS	22	16 (16)	16 (16)	12 (17)	14 (17)	16 (16)	11 (14)
Supv-PFB	Batch A	11	11 (07)	11 (07)	10 (08)	10 (07)	16 (09)	10 (09)
	All MOS	22	16 (15)	15 (14)	12 (15)	14 (13)	18 (13)	11 (13)
Supv-LEAD	Batch A	11	15 (10)	14 (10)	14 (11)	14 (10)	16 (13)	13 (12)
	All MOS	22	19 (17)	19 (17)	15 (15)	19 (16)	20 (17)	15 (15)
SK-Basic	Batch A	9	68 (06)	69 (06)	68 (06)	68 (06)	68 (07)	68 (06)
	All MOS	20	67 (08)	68 (08)	65 (16)	67 (09)	66 (11)	66 (10)
SK-Tech	Batch A	11	76 (05)	77 (05)	77 (05)	76 (05)	76 (05)	76 (05)
	All MOS	22	75 (06)	75 (06)	75 (05)	75 (06)	75 (07)	74 (07)
SK-Total	Batch A	11	78 (03)	79 (03)	79 (03)	78 (03)	79 (03)	78 (04)
	All MOS	22	77 (05)	77 (05)	77 (05)	77 (05)	77 (06)	76 (06)

Note: Corrected for range restriction and adjusted for shrinkage (Rozeboom, 1978, formula 8). Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Multiple Rs for ASVAB Factors alone are underlined. Underlined numbers in boldface denote multiple Rs greater than for ASVAB Factors alone. Decimals omitted.

[a] ETS = Effort and Technical Skill; MPD = Maintaining Personal Discipline; PFB = Physical Fitness and Military Bearing; LEAD = Leadership Potential; SK = School Knowledge.

[b] Number of MOS for which validities were computed.

Development of Basic Scores for the Longitudinal Validation (LVI) Performance Measures

In 1988 and 1989, first-tour criterion measures were administered to the Longitudinal Validation sample (LVI). This data collection was conducted concurrently with the administration of second-tour criterion measures to the Concurrent Validation sample (CVII). Before the LVI performance model development and subsequent validation analyses could begin, it was necessary to derive basic scores for each of the individual first-tour job performance measures. Dealing with all the individual scores from each task test, each rating scale, and each administrative index was simply not feasible or desirable. There were too many, and the reliabilities of the individual items or scales preserved too much measurement error with very little gain in total information. Consequently, the full array of scale scores was aggregated into a smaller set of basic scores for each measure.

Table 1.15 lists the individual measures that were administered.

Table 1.15
Measures Administered to Soldiers in LVI Sample

MOS in Batch A:	Background Information Form
	Job Knowledge Tests
	Hands-On Tests
	Army-Wide Rating Scales
	MOS-Specific Rating Scales
	Combat Performance Prediction Scales (males only)
	Personnel File Form
	Army Job Satisfaction Questionnaire
	Job History Questionnaire
	Physical Requirements Survey[a]
MOS in Batch Z:	Background Information Form
	School Knowledge Test
	Army-Wide Rating Scales
	Combat Performance Prediction Scales (males only)
	Personnel File Form
	Army Job Satisfaction Questionnaire
	Physical Requirements Survey[a]

Note. Rating scale data were collected from both supervisors and peers.

[a] The Physical Requirements Survey is not a Career Force or Project A measure.

Differences Between CVI and LVI Performance Measures

The 3-year time period between CVI and LVI raised the issue that for the job knowledge and hands-on measures, equipment and/or procedural changes would require test revisions, and changes in MOS responsibilities had the potential of making some tasks obsolete.

Project staff identified relevant changes so that the appropriate revisions could be made. In a few cases where an entire task was obsolete, the task was dropped without replacement. In many cases, revisions were simply a matter of replacing outdated terminology. Updated criterion measures were forwarded to the MOS proponents for a currency review and additional revisions were made on the basis of this review.

While there was considerable interest in keeping the Combat Performance Prediction Scales, project staff and the Scientific Advisory Group agreed that the version used in CVI was too lengthy. New scales were field tested in conjunction with the second-tour criterion measure field tests. The decision was made to retain the original summated scale format, but the total number of items was reduced from 40 to 19.

The self-report form for gathering information on administrative records was updated by reviewing its contents with officers and NCOs representing the Army Personnel Command (PERSCOM). The form was altered to allow soldiers to report an M19 qualification in the event that an M16 qualification was not applicable. Also, three awards were dropped per guidance from PERSCOM.

Task-level ratings were deleted from the array of Batch A first-tour criterion measures used in CVI. The Army-wide and MOS-specific rating scales were retained in their original form.

The development of the basic scores for each measure was based on the performance data collected from individuals in the Batch A and Batch Z MOS that were included in the administration of first-tour criterion measures in 1988 and 1989. The Batch A MOS were the same as those studied in the Concurrent Validation, except for the addition of 19K (M1 Armor Crewman).

As in CVI, the Batch A MOS differed from the Batch Z MOS in the comprehensiveness of the MOS-specific criterion measures that were available for administration. MOS-specific rating scales, hands-on tests, and job knowledge tests were administered to Batch A soldiers. The only MOS-specific measure available for administration to the Batch Z soldiers was the school knowledge test that had been developed for administration at the end of training. The school knowledge test was administered to the Batch Z examinees as a surrogate for a job knowledge test.

Score Development for Administrative Indices

Five scores were computed from the LVI Personnel File Form: (a) awards and memoranda/certificates of achievement, (b) Physical Readiness Test, (c) M16 qualification, (d) Articles 15 and flag actions (disciplinary actions), and (e) promotion rate.

The first score was a composite of (a) awards and decorations; (b) memoranda of appreciation, commendation, or achievement; and (c) certificates of appreciation, commendation, or achievement. The last score, promotion rate, was derived from data available in the Army's computerized personnel records. It was the residual of pay grade regressed on time in service, adjusted by MOS.

Basic Score for the Combat Performance Prediction Ratings

Principal components analyses of the LVI/CVII Combat Scale data indicated the presence of two factors. The second factor, however, was defined by the three negatively worded items. Given that the second factor was probably not substantively distinct from the first, a single total score (with the negatively worded items reverse-scored) was calculated for the Combat Scale ratings. (The two factors were essentially the same as those found in CVI, where two Combat Scale scores were derived.)

Development of Basic Scores for the First-Tour Performance Rating Scales

The Army-wide rating scales include 12 dimensions of soldier effectiveness that are important regardless of soldiers' MOS. MOS-specific rating scales were developed for each of the nine Batch A MOS, and these rating scales include between 7 and 13 dimensions of MOS-specific performance.

Principal factor analyses with varimax rotation were conducted on the Army-wide ratings (across all MOS), for supervisor and peer ratings separately *and* pooled together. The pooled ratings were computed by averaging the mean peer rating and one supervisor rating for those soldiers who had at least one peer rating and one supervisor rating. Because previous analyses (using the CVI sample) showed that a single factor was sufficient to account for the majority of the variance in the MOS-specific ratings, factor analyses were not conducted for the MOS-specific rating data.

Table 1.16 shows the three-factor, rotated solutions for the pooled peer/supervisor ratings. These data demonstrate the remarkable similarity of the rotated factor structures for the CVI and LVI samples. It is worth noting that these same three factors were also obtained in factor analyses of performance rating data for a sample of 950 second-tour soldiers, which was collected using a set of rating scales very similar to those used to collect the present data (Campbell & Zook, 1990).

Table 1.16
Comparison of LVI and CVI Army-Wide Factor Analysis[a] Results: Pooled Peer/ Supervisor Ratings[b]

Dimension	Factor Loadings (LVI/CVI)		
	1	2	3
Technical Knowledge/Skill	.67/.71	.30/.28	.38/.30
Leadership	.65/.69	.34/.30	.44/.37
Effort	.66/.69	.47/.43	.32/.26
Self-Development	.52/.57	.42/.38	.46/.38
Maintaining Equipment	.50/.54	.41/.34	.41/.35
Following Regulations	.39/.41	.73/.69	.31/.30
Self-Control	.19/.22	.65/.63	.20/.20
Integrity	.44/.50	.66/.59	.30/.28
Military Bearing	.31/.32	.35/.32	.57/.57
Physical Fitness	.24/.21	.16/.15	.49/.49
Percent Common Variance	37.7/44.9	36.6/32.7	25.6/22.4

Note. Sample size is 7,919 for LVI and 8,642 for CVI.

[a] Principal factor analysis, varimax rotation.
[b] Computed by averaging the mean peer rating and the mean supervisor rating.

For both the Army-wide and MOS-specific rating scales, the mean, variability, and reliability of the peer, supervisor, and pooled peer/ supervisor ratings appear quite acceptable and are comparable to what was found in the CVI research. Factor analyses of the Army-wide ratings showed that the three-factor CVI solution was replicated in the present data. Accordingly, the three composites shown in Table 1.17, along with the overall effectiveness rating, were used as the basic scores for the Army-wide rating data.

Table 1.17
Composition and Definition of LVI Army-Wide Rating Composites

Factor Name and Definition	Percent Common Variance Accounted For by Relevant Factor[a] (LVI/CVI)	Dimensions Included
1. Technical Skills and Job Effort: Exerting effort over the full range of job tasks; engaging in training or other development activities to increase proficiency; persevering under dangerous or adverse conditions; and demonstrating leadership and support toward peers.	37.8/44.9	Technical Knowledge/ Skill Leadership Effort Self-Development Maintaining Equipment
2. Personal Discipline: Adhering to Army rules and regulations; exercising self-control; demonstrating integrity in day-to-day behavior; and not causing disciplinary problems.	36.6/32.7	Following Regulations Self-Control Integrity
3. Physical Fitness/Military Bearing: Maintaining an appropriate military appearance and bearing and staying in good physical condition.	25.6/22.4	Military Bearing Physical Fitness

[a] Factor analysis of pooled peer/supervisor ratings.

Final Array of LVI Basic Performance Scores

A summary list of the basic performance scores produced by the analyses summarized above is given in Figure 1.11. These are the scores that were put through the final editing and score imputation procedures for the LVI data file. The scores that formed the basis for the confirmatory tests of the LVI model of first-tour job performance were also drawn from this array.

Development of Basic Scores for Hands-On Performance and Job Knowledge Measures

As the first step in replicating the CVI procedures for constructing the basic scores, tasks were clustered into Functional Categories as described in the Project A annual report for 1986 (Campbell, 1987b).

Following the procedures developed with the CVI data, tasks were also sorted into six higher level groups referred to as Task Factors (Communication, Vehicles, Basic Techniques, Identify Targets, Technical, and Safety/ Survival) and known as CVBITS. Tasks were also combined into just two groups: General (i.e., Army-wide) and MOS-specific.

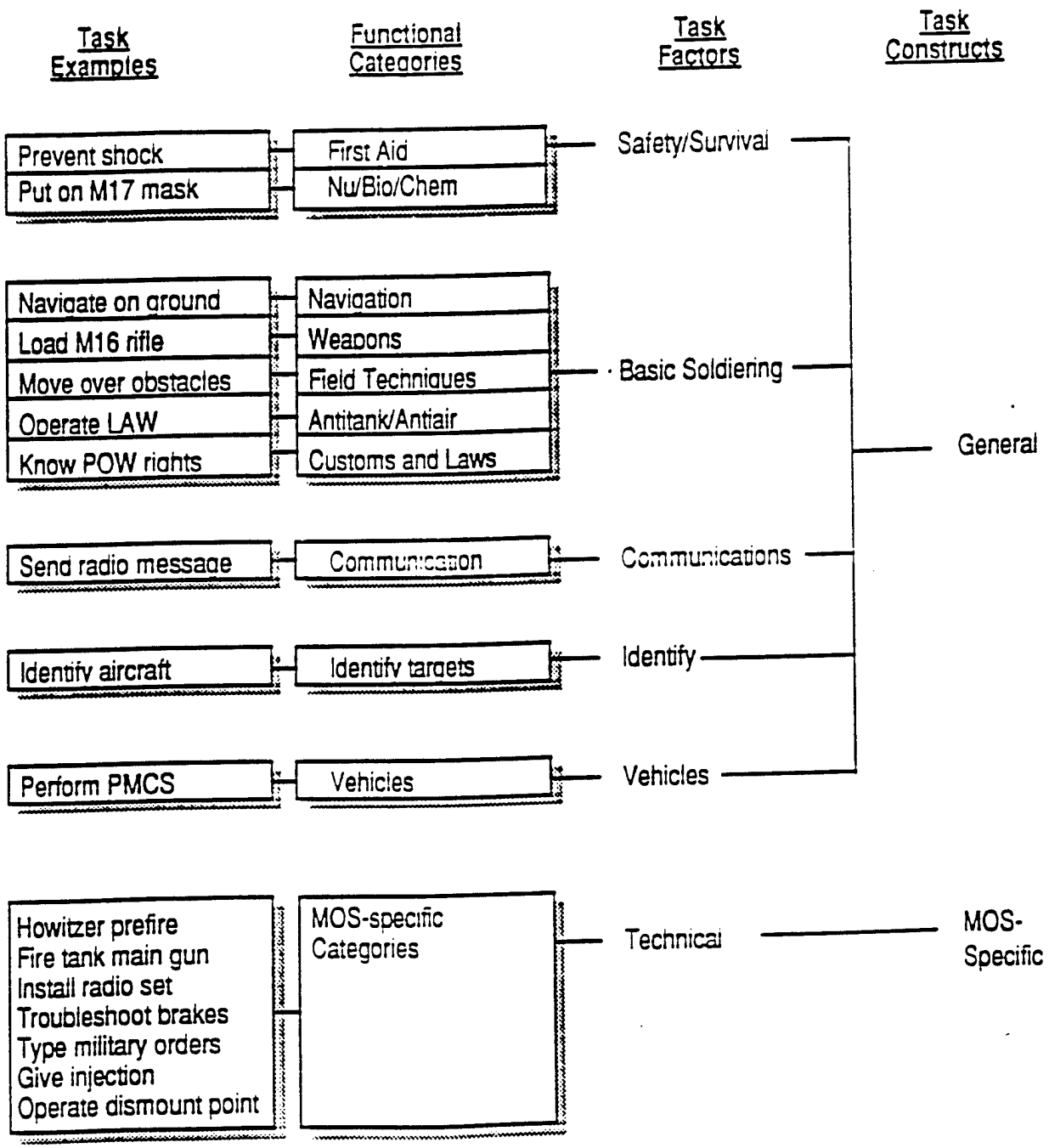
In general, the grouping schemes are hierarchical: Tasks (the lowest level) are placed in Functional Categories, the Functional Categories (level two) are aggregated to form the six Task Factors (level three), and Task Factors are then aggregated to form the two Task Constructs (level four), as diagrammed in Figure 1.10.

For the LVI data, confirmatory factor analyses were conducted to assess the fit of alternative levels of score aggregation. These analyses served two purposes: They were used to assess the relative merits of each model and to corroborate the CVI decision to use the six task factor scores (CVBITS). The analysis required the computation of separate tests of goodness of fit for hands-on and job knowledge test data, for each of the 10 MOS, on each of three competing models. The three models tested were a one-factor model, postulating the existence of a single factor in the data; a two-factor model, proposing the Basic and the Technical Task Constructs; and a three-to-six-factor model (the number of factors varying among MOS and test method), using the Task Factors. Examination of the results from LVI argues for the retention of the six Task Factor scores for both the Hands-On and Job Knowledge measures.

The LVI Data File: Final Data Editing and Score Imputation

The Longitudinal Validation First-Tour (LVI) data were collected from 11,266 soldiers in 21 MOS -- 6,815 from Batch A MOS and 4,451 from Batch Z MOS. Extensive efforts were made to collect complete information from each examinee for all instruments. However, as with all data collection exercises, circumstances precluded complete success. The final counts of soldiers for whom data were analyzed for each instrument are given in Tables 1.18 and 1.19 for Batch A and Batch Z MOS, respectively.

Data for each performance measure were processed individually. After processing was completed for these individual measures, they were combined so that all LVI data for each examinee were included in a single file. The data were combined separately by MOS. When the data were combined, basic scores were calculated for the individual performance measures. Table 1.20 shows the amount of missing data for the final set of basic criterion scores.



Note. The Task Factors correspond to the six task groups known as CVBITS. The Task Constructs termed General and MOS-Specific refer to the same constructs that have previously been called Basic and Technical, or Common and Technical.

Figure 1.10. Hierarchical relationships among Functional Categories, Task Factors, and Task Constructs.

Hands-On Performance Test

1. Safety/survival performance score
2. General (common) task performance score
3. Communication performance score
4. Vehicles performance score
5. MOS-specific task performance score

Job Knowledge Test

6. Safety/survival knowledge score
7. General (common) task knowledge score
8. Communication knowledge score
9. Identify targets knowledge score
10. Vehicles knowledge score
11. MOS-specific task knowledge score

Army-Wide Rating Scales

12. Overall effectiveness rating
13. Technical skill and effort composite
14. Personal discipline composite
15. Physical fitness/military bearing composite

MOS-Specific Rating Scales

16. Overall MOS composite

Combat Performance Prediction Scales

17. Overall Combat Prediction scale composite (available for males only)

Personnel File Form

18. Awards and Certificates
19. Disciplinary Actions (Articles 15 and Flag Actions)
20. Physical Readiness
21. M16 Qualification
22. Promotion Rate

Figure 1.11. Summary list of LVI basic criterion scores.

Table 1.18
LVI Sample Sizes for Performance Measures for Batch A MOS

MOS		N	Hands-On	Job Knowledge	Army-Wide Ratings	MOS Ratings	Combat Ratings	Personnel File	Combined Criteria[a]
11B	Infantryman	909	890	895	899	899	898	906	907
13B	Cannon Crewmember	916	773	810	897	897	897	916	916
19E	M60 Armor Crewman	249	243	248	241	241	241	249	249
19K	M1 Armor Crewman	824	749	812	782	778	782	819	825
31C	Single Channel Radio Operator	529	446	504	497	481	442	527	529
63B	Light-Wheel Vehicle Mechanic	752	624	723	728	719	666	750	752
71L	Administrative Specialist	678	641	664	634	626	199	675	678
88M	Motor Transport Operator	682	588	674	666	663	479	680	682
91A	Medical Specialist	824	794	798	807	797	670	818	824
95B	Military Police	452	444	446	451	450	366	452	452
Total		6,815	6,192	6,574	6,602	6,547	5,640	6,792	6,814

[a] Combined Criteria include Hands-On, Job Knowledge, Army-Wide Ratings, MOS Ratings, and Personnel File Form.

Table 1.19
LVI Sample Sizes for Performance Measures for Batch Z MOS

MOS		N	Job Knowledge	Army-Wide Ratings	Combat Ratings	Personnel File
12B	Combat Engineer	841	840	827	827	838
16S	MANPADS Crewman	472	471	468	468	472
27E	Tow/Dragon Repairer	90	90	89	84	90
29E	Comm.-Electronics Radio Repairer	112	111	106	101	111
51B	Carpentry/Masonry Specialist	213	212	193	190	212
54B	NBC Specialist	499	498	492	462	498
55B	Ammunition Specialist	279	279	269	243	279
67N	Utility Helicopter Repairer	197	194	193	192	197
76Y	Unit Supply Specialist	788	788	734	616	787
94B	Food Service Specialist	832	932	818	717	931
96B	Intelligence Analyst	128	128	122	103	128
Total		4,451	4,443	4,311	4,003	4,443

Table 1.20
LVI Combined Criteria Data: Percentage of Missing Data for Basic Scores by MOS

Criteria	11B	13B	19E	19K	31C	63B	71L	88M	91A	95B
Hands-On - Task Factors										
C - Communications	1.87	15.61	2.41	9.21	15.69	--	--	--	--	--
V - Vehicles	--	--	--	--	15.69	17.02	--	13.78	--	1.77
B - Basic Soldiering	1.87	15.61	2.41	9.21	15.69	17.02	5.46	13.78	3.64	1.77
I - Identify Targets	--	--	--	--	--	--	--	--	--	--
T - Technical	-	15.61	2.41	9.21	15.59	17.02	5.46	--	3.64	1.77
S - Safety/Survival	1.87	15.61	2.41	9.21	15.69	17.02	5.46	13.78	3.64	1.77
Job Knowledge - Task Factors										
C - Communications	2.65	12.01	.40	1.94	6.62	--	--	--	--	1.55
V - Vehicles	--	--	--	--	6.62	5.05	--	1.91	4.98	1.55
B - Basic Soldiering	2.65	12.01	.40	1.94	6.62	5.05	2.36	1.91	4.98	1.55
I - Identify Targets	2.65	12.01	.40	1.94	6.62	--	--	1.91	4.98	1.55
T - Technical	--	12.01	.40	1.94	6.62	5.05	2.36	--	4.98	1.55
S - Safety/Survival	2.65	12.01	.40	1.94	6.62	5.05	2.36	1.91	4.98	1.55
Army-Wide Ratings										
Overall Effectiveness	1.10	2.95	3.21	5.33	6.99	3.32	7.67	2.79	2.31	.22
Technical Skill and Effort	.88	2.07	3.21	5.21	6.05	3.19	6.93	2.35	2.06	.22
Personal Discipline	.88	2.07	3.21	5.21	6.05	3.19	6.93	2.35	2.06	.22
Physical Fitness/Bearing	.88	2.07	3.21	5.21	6.05	3.19	6.93	2.35	2.06	.22
MOS Ratings										
MOS Composite Rating	1.32	5.35	3.21	6.67	9.83	4.65	9.73	4.55	6.43	1.55
Personnel File Form										
Awards and Certificates	2.43	3.60	2.01	3.76	4.91	3.19	2.65	2.79	2.91	1.77
Articles 15 and Flag Actions	1.21	1.53	.00	1.82	1.51	1.06	1.18	.73	1.94	.44
Physical Readiness Score	4.63	5.46	3.21	5.21	9.45	11.44	9.00	9.09	6.55	5.31
M16 Qualification	2.65	4.04	29.32	18.30	2.65	3.19	1.77	2.93	3.88	3.98
Promotion Rate	1.76	1.86	.80	4.00	5.10	4.79	5.01	3.96	2.67	0.88

Note. -- indicates that the particular score was not calculated for that MOS.

In addition to the performance data, missing Longitudinal Validation predictor data were also imputed. For a complete description of the editing process used on the predictor data, see the 1990 annual report. The bulk of the editing process was accomplished during FY90, but additional work was done during FY91. The amounts of missing data for each score on each paper-and- pencil and each computerized measure are shown in Tables 1.21 and 1.22.

An imputation procedure known as PROC IMPUTE (Wise & McLaughlin, 1980) was developed that used existing data to estimate values for missing data. This procedure was also used in the CVI analyses (Wise, McHenry, & Young, 1986). The decision rules used in the CVI analyses were replicated in the LVI analyses as closely as possible.

PROC IMPUTE uses regression estimates to predict missing values. Each missing value is predicted from other values for the subject in question so that individual differences are retained. The regression coefficient and intercept vary from item to item so that differences in item difficulty are also reflected in the predicted values. PROC IMPUTE also adds a random variable with variance equal to the error of estimate for predicting the missing value.

The results of the imputation were examined at two levels. First, after each PROC IMPUTE run, the program output was inspected. Second, the pre-imputed and the post-imputed data sets were compared for each MOS (a) after the hands-on score level imputation, and (b) after the criterion construct level imputation.

The means and variances of the pre- and post-imputation results for the hands-on data for each MOS were found to be virtually identical. Imputation also made virtually no difference in the magnitude of the intercorrelations among the criterion scores that were used to create the performance factor scores in the validation analyses. These results are similar to those obtained earlier from the CVI imputation (Wise et al., 1986).

Development of the LVI First-Tour Performance Model

A latent factor model of first-tour performance, developed using data from the Project A Concurrent Validation (CVI) sample, has been described by Campbell, McHenry, and Wise (1990). This model included the now familiar five performance factors--Core Technical Proficiency (CTP), General Soldiering Proficiency (GSP), Effort and Leadership (ELS), Maintaining Personal Discipline (MPD), and Physical Fitness and Military Bearing (PFB)-- and two measurement method factors, a Ratings method factor and a Paper-and-Pencil Test method factor. During year two, the CVI model was subjected to a confirmatory analysis, using first-tour performance data collected from the Longitudinal Validation (LVI) sample. Additionally, comparative analyses aimed at evaluating more parsimonious models of first-tour performance were carried out.

Table 1.21
LVI Predictor Data: Amount of Missing Data for Paper-and-Pencil Scale Scores

Score	Not Missing	Missing
Assembling Objects - Number Correct	49,042	366
Map - Number Correct	49,047	361
Maze - Number Correct	49,052	356
Object Rotation - Number Correct	49,103	305
Orientation - Number Correct	49,072	336
Reasoning - Number Correct	49,103	305
JOB Scale 1 - Pride	46,525	2,883
JOB Scale 2 - Job Security/Comfort	46,634	2,774
JOB Scale 3 - Serving Others	46,295	3,113
JOB Scale 4 - Job Autonomy	46,037	3,371
JOB Scale 5 - Routine	45,975	3,433
JOB Scale 6 - Ambition	46,058	3,350
ABLE Scale 1 - Emotional Stability	44,264	5,144
ABLE Scale 2 - Self-Esteem	44,247	5,161
ABLE Scale 3 - Cooperativeness	44,258	5,150
ABLE Scale 4 - Conscientiousness	44,199	5,209
ABLE Scale 5 - Nondelinquency	44,228	5,180
ABLE Scale 6 - Traditional Values	44,190	5,218
ABLE Scale 7 - Work Orientation	44,260	5,148
ABLE Scale 8 - Internal Control	44,254	5,154
ABLE Scale 9 - Energy Level	44,217	5,191
ABLE Scale 10 - Dominance	44,246	5,162
ABLE Scale 11 - Physical Condition	44,264	5,144
AVOICE Scale 1 - Clerical/Administrative	45,477	3,931
AVOICE Scale 2 - Mechanics	45,941	3,467
AVOICE Scale 3 - Heavy Construction	45,851	3,557
AVOICE Scale 4 - Electronics	45,922	3,486
AVOICE Scale 5 - Combat	45,939	3,469
AVOICE Scale 6 - Medical Services	45,545	3,863
AVOICE Scale 7 - Rugged Individualism	45,944	3,464
AVOICE Scale 8 - Leadership/Guidance	45,508	3,900
AVOICE Scale 9 - Law Enforcement	45,958	3,450
AVOICE Scale 10 - Food Service Professional	45,916	3,492
AVOICE Scale 11 - Firearms Enthusiast	45,942	3,466
AVOICE Scale 12 - Science/Chemical	45,970	3,438
AVOICE Scale 13 - Drafting	45,976	3,432
AVOICE Scale 14 - Audiographics	45,452	3,956
AVOICE Scale 15 - Aesthetics	45,279	4,129
AVOICE Scale 16 - Computers	45,554	3,854
AVOICE Scale 17 - Food Service Employee	45,965	3,443
AVOICE Scale 18 - Mathematics	45,691	3,717
AVOICE Scale 19 - Electronic Communications	45,602	3,806
AVOICE Scale 20 - Warehousing/Shipping	45,963	3,445
AVOICE Scale 21 - Fire Protection	45,972	3,436
AVOICE Scale 22 - Vehicle Operator	45,971	3,437

Table 1.22
LVI Predictor Data: Amount of Missing Data for Computer-Administered Scale Scores

Score	Not Missing	Missing
Target Identification - Mean of Clipped Decision Time	38,401	513
Target Identification - Proportion Correct	38,404	510
Number Memory - Mean of Clipped Operation Means	38,324	590
Number Memory - Proportion Correct	38,353	561
Target Track 1 - Mean Log (Distance+1)	38,825	89
Target Track 2 - Mean Log (Distance+1)	38,793	121
Cannon Shoot - Mean Absolute Time Discrepancy	38,603	311
Target Shoot - Mean Log (Distance+1)	37,477	1,437
Mean of Median Movement Times across 5 tests	37,863	1,051
Simple Reaction Time - Median Decision Time	38,747	167
Simple Reaction Time - Proportion Correct	38,747	167
Choice Reaction Time - Median Decision Time	38,856	58
Choice Reaction Time - Proportion Correct	38,856	58
Perceptual Speed/Accuracy - Mean of Clipped Decision Time	38,703	211
Perceptual Speed/Accuracy - Proportion Correct	38,734	180
Short-Term Memory - Mean of Clipped Decision Time	38,483	431
Short-Term Memory - Proportion Correct	38,490	424

An earlier section summarized how each of the major sets of performance measures was reduced from a large number of item, task, or individual scale scores to a smaller set of factor or category scores. The results of this first level of aggregation have been referred to as the "basic" array of criterion scores, summarized in Figure 1.11. These included the scores that were used in the modeling analyses described below.

Altogether, the LVI first-tour performance measures were reduced to 20 basic scores. However, because MOS differ in their task content, not all 20 variables were scored in each MOS, and there was some slight variation in the number of variables used in the subsequent analyses.

To test the fit of the different models to the LVI data, confirmatory factor-analytic techniques were applied to each MOS individually, using LISREL 7 (Jöreskog & Sörbom, 1989). The first alternative five-factor model was developed using CVI data. After the fit of the five-factor model was assessed in each MOS, four reduced models (all nested within the five-factor model) were examined. Finally, as had been done in the original CVI analyses, the five-factor model was applied to the Batch A MOS simultaneously (using LISREL's multigroups option). The fit statistics (e.g., root mean-square residuals [RMSRs]) of the five-factor model for each MOS in the LVI and CVI samples were very similar. In fact, for three of the MOS (11B, 13B, and 71L), the RMSRs for the LVI data were smaller than those for the CVI data. These results indicate that the model developed using the CVI data does fit the LVI data quite well.

Four reduced models were also examined using the LVI data. For the four-factor model, the Core Technical Proficiency and General Soldiering Proficiency performance factors were collapsed into a single "can do" performance factor. The three-factor model retained the "can do" performance factor of the four-factor model, but also collapsed the Effort and Leadership and Maintaining Personal Discipline performance factors into a "will do" performance factor. For the two-factor model, the "can do" performance factor was retained; however, the Physical Fitness and Military Bearing performance factor became part of the "will do" performance factor. Finally, for the one-factor model, the "can do" and "will do" performance factors, or equivalently, the five original performance factors, were collapsed into a single performance factor.

The chi-square statistics and RMSRs, respectively, for the four reduced models, as well as for the five-factor model, indicate that the four- and five-factor models fit the LVI data well, while the one-, two-, and three- factor models fit less well. The results also indicated that the parameter estimates for the five-factor model were generally similar across the 10 MOS. The final step was to determine whether the variation in some of these parameters could be attributed to sampling variation. To do this (as described earlier), the following were specified to be invariant across jobs: (a) the correlations among performance factors, (b) the loadings of all the Army-wide measures on the performance factors and on the rating method factor, (c) the loadings of the MOS-specific score on the rating method factor, and (d) the uniqueness coefficients for the Army-wide measures.

The results indicated that the fit of the five-factor model is not as good when the parameters listed above are constrained to be equal across the 10 jobs. Still, the root mean-square residuals associated with the across-MOS model are not substantially greater than those for the within-job analyses. (The average RMSR for the across-MOS model is .0676; the average for the within-MOS models is .0585.)

To create criterion construct scores for use in validation analyses, the scoring procedures were based on the five-factor model. Although the four-factor model has the advantage of greater parsimony, the five-factor model offered the advantage of corresponding to the criterion constructs generated in the CVI validation analyses. Table 1.23 shows the mapping of the basic scores on the five performance factors. As with the CVI data, five residual scores, corresponding to the five criterion constructs, were also created.

The five "raw" criterion construct scores, the five residual criterion construct scores, the total rating and job knowledge scores, and the total score derived from the hands-on test were used to generate a 13 x 13 matrix of criterion intercorrelations for each MOS in Batch A. The averages of these correlations are reported in Table 1.24. These results are very similar to the correlations that were reported by J. P. Campbell et al. (1990) for the CVI sample.

Basic Validation Results for the LVI Sample

The LVI validation results were based on two different sample editing strategies. The first required complete data for all predictor composites, as well as for the ASVAB, and for each performance factor; this sample is referred to as the "listwise deletion" sample. In the alternative strategy, called setwise deletion, a separate validation sample was identified for each set of predictors in the Experimental Battery.

The number of soldiers with complete predictor and criterion data in each MOS is reported in Table 1.25 for both the CVI and LVI data sets.

Table 1.23
Mapping of LVII Performance Measures Onto Latent Performance Factors

	Performance Factors					Method Factors	
Criterion Score[a]	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Maintaining Personal Discipline	Physical Fitness/ Military Bearing	Written Knowledge Tests	Rating Scales
HO Technical	X						
HO Communication		X					
HO Vehicles		X					
HO General Soldier		X					
HO Safety/Survival		X					
JK Technical	X					X	
JK Communication		X				X	
JK Vehicles		X				X	
JK General Soldier		X				X	
JK HD Threat/Target		X				X	
JK Safety/Survival		X				X	
AWB Skill/Effort Composite			X				X
AWB Discipline Composite				X			X
AWB Fitness Composite					X		X
AWB Overall Composite			X	X			X
MOS Rating Composite			X			X	
PFF Awards/Certificates			X				
PFF Physical Readiness					X		
PFF Articles UP/Flags				X			
PFF Promotion Rate				X			

[a] AWB = Army-Wide Rating Scales; HO = Hands On; JK = Job Knowledge; PFF = Personnel File Form.

Table 1.24
Mean Intercorrelations Among 13 Summary Criterion Scores for the Batch A MOS in the LVI Sample

Summary Criterion Score[a]	CTP Raw	GSP Raw	ELS Raw	MPD Raw	PFB Raw	CTP Res	GSP Res	ELS Res	MPD Res	PFB Res	PRT	HOT	JKT
CTP (raw)	1.00												
GSP (raw)	.57	1.00											
ELS (raw)	.25	.26	1.00										
MPD (raw)	.16	.18	.58	1.00									
PFB (raw)	.06	.06	.48	.36	1.00								
CTP (residual)	.88	.41	.30	.20	.07	1.00							
GSP (residual)	.40	.88	.32	.23	.06	.45	1.00						
ELS (residual)	.41	.42	.70	.43	.26	.40	.42	1.00					
MPD (residual)	.20	.22	.28	.88	.17	.20	.23	.46	1.00				
PFB (residual)	.07	.07	.20	.21	.90	.04	.03	.29	.21	1.00			
Perf. Rating Total	.22	.24	.88	.72	.58	.27	.28	.40	.35	.24	1.00		
Hands-On Total	.72	.76	.26	.15	.08	.81	.85	.41	.18	.09	.23	1.00	
Job Knowledge Total	.74	.80	.25	.19	.04	.40	.46	.40	.23	.04	.22	.47	1.00

[a] CTP = Core Technical Proficiency; GSP = General Soldiering Proficiency; ELS = Effort and Leadership; MPD = Maintaining Personal Discipline; PFB = Physical Fitness and Military Bearing.

Table 1.25
Soldiers in CVI and LVI Data Sets With Complete Predictor and First-Tour Criterion Data by MOS

MOS		CVI	LVI (Listwise Deletion Sample)
11B	Infantryman	491	235
13B	Cannon Crewmember	464	553
19E[a]	M60 Armor Crewman	394	73
19K	M1 Armor Crewman	---	446
31C	Single Channel Radio Operator	289	172
63B	Light-Wheel Vehicle Mechanic	478	406
71L	Administrative Specialist	427	252
88M	Motor Transport Operator	507	221
91A	Medical Specialist	392	535
95B	Military Police	597	270
Total		4,039	3,163

[a] MOS 19E not included in LVI validity analyses.

The analysis procedure consisted of the following major steps:

A) Using the listwise deletion sample, multiple correlations between each set of predictor scores and the five substantive factor scores, their five residual factor scores, the two method factor scores, and the total scores from the hands-on and job knowledge tests were computed separately by MOS and then averaged.

B) Using the listwise deletion sample, incremental validities for each set of Experimental Battery predictors (e.g., AVOICE composites or computer composites) over the four ASVAB factor composites were computed against the same criteria used to compute the validities in Step A. Once again, the results were computed separately by MOS and then averaged.

C) Using the setwise deletion samples, multiple correlations and incremental validities (over the four ASVAB factor composites) between each set of Experimental Battery predictors and the criteria used in the first two steps were computed separately by MOS and then averaged. All results to this point were corrected for range restriction and adjusted for shrinkage using the Rozeboom formula.

D) Finally, once again using the listwise deletion sample, multiple correlations and incremental validities (over the four ASVAB factors) were computed for each set of predictors in the Experimental Battery, this time adjusting the results for shrinkage with the Claudy (1978) instead of the Rozeboom formula. This step was conducted to allow comparisons between the first-tour validity results associated with the longitudinal sample and those that had been reported for the concurrent sample (for which only the Claudy formula was used, e.g., McHenry, Hough, Toquam, Hanson, & Ashworth, 1990).

Multiple Correlations and Incremental Validities Based on Listwise Deletion Samples

Multiple correlations for the four ASVAB factor composites, the single spatial composite, the eight computer composites, the three JOB composites, the seven ABLE composites, and the eight AVOICE composites are reported in Table 1.26.

Incremental validity results for the Experimental Battery predictors over the ASVAB factors are reported in Table 1.27. The results indicate that the spatial composite added slightly to the prediction of the raw and residual Core Technical and General Soldiering performance factors, as well as to the written method factor and the hands-on and job knowledge total scores. They also show that the seven ABLE composites contributed substantially to the prediction of the raw and residual Personal Discipline and Physical Fitness performance factors.

Multiple Correlations and Incremental Validities Based on the Setwise Deletion Samples

Multiple correlations for the spatial composite, the eight computer composites, the three JOB composites, the seven ABLE composites, and the eight AVOICE composites based on the setwise deletion samples described above are reported in Table 1.28. These multiple correlations were very similar to those computed with the listwise sample. However, there was a consistent difference between the two sets of results: specifically, the multiple correlations based on the setwise samples were generally one to three validity points higher.

Incremental validity results associated with the setwise deletion samples can be found in Table 1.29. The incremental validity results based on the setwise samples were practically identical to those based on the listwise sample. Again, the primary difference between the two sets of results was that the level of validities was sometimes one or two points lower for the listwise sample than for the setwise samples.

Table 1.26
Mean of Multiple Correlations Computed Within Job for LVI Listwise Deletion Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE

Criterion[a]	No. of MOS[b]	ASVAB Factors [4]	Spatial [1]	Computer [8]	JOB [3]	ABLE Comp. [7]	AVOICE [8]
CTP (Raw)	9	62 (13)	57 (11)	47 (16)	29 (13)	21 (09)	38 (08)
GSP (Raw)	8	66 (07)	64 (06)	55 (08)	29 (13)	23 (14)	37 (07)
ELS (Raw)	9	37 (12)	32 (08)	29 (15)	18 (14)	13 (11)	17 (15)
MPD (Raw)	9	17 (13)	14 (11)	10 (16)	06 (13)	14 (11)	05 (10)
PFB (Raw)	9	16 (06)	10 (04)	07 (07)	06 (06)	27 (07)	05 (09)
CTP (Res)	9	46 (17)	42 (15)	29 (22)	17 (12)	08 (11)	28 (12)
GSP (Res)	8	51 (10)	51 (08)	41 (10)	18 (11)	12 (12)	26 (09)
ELS (Res)	9	46 (18)	41 (13)	37 (20)	23 (15)	21 (15)	24 (16)
MPD (Res)	9	18 (13)	14 (12)	08 (16)	07 (11)	13 (11)	06 (10)
PFB (Res)	9	20 (10)	12 (08)	09 (11)	07 (06)	28 (10)	09 (11)
Written	9	54 (13)	49 (12)	43 (18)	29 (16)	23 (12)	29 (14)
Ratings	9	12 (09)	09 (07)	07 (09)	06 (09)	03 (05)	02 (07)
HO-Total	9	50 (14)	48 (11)	38 (15)	18 (13)	11 (11)	28 (09)
JK-Total	9	71 (08)	65 (07)	58 (10)	36 (14)	31 (08)	41 (08)

Note. Corrected for range restriction, and adjusted for shrinkage (Rozeboom formula 8). Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Decimals omitted.

[a] CTP = Core Technical Proficiency; GSP = General Soldiering Proficiency; ELS = Effort and Leadership; MPD = Maintaining Personal Discipline; PFB = Physical Fitness and Military Bearing; HO = Hands-On; JK = Job Knowledge.

[b] Number of MOS for which validities were computed.

Table 1.27
Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for LVI Listwise Deletion Samples for Spatial, Computer, JOB, ABLE Composites, and AVOICE

Criterion	No. of MOS[a]	ASVAB Factors (A4) [4]	A4+ Spatial [5]	A4+ Computer [12]	A4+ JOB [7]	A4+ ABLE Comp. [11]	A4+ AVOICE [12]
CTP (Raw)	9	*62* (13)	**63** (13)	61 (14)	61 (13)	61 (13)	62 (13)
GSP (Raw)	8	*66* (07)	**68** (07)	66 (07)	66 (07)	66 (07)	66 (07)
ELS (Raw)	9	*37* (12)	36 (13)	35 (13)	36 (13)	34 (17)	33 (16)
MPD (Raw)	9	*17* (13)	16 (14)	16 (15)	14 (15)	**23** (14)	10 (15)
PFB (Raw)	9	*16* (06)	13 (08)	09 (08)	**17** (08)	**30** (06)	12 (10)
CTP (Res)	9	*46* (17)	**47** (17)	44 (18)	45 (18)	43 (19)	46 (19)
GSP (Res)	8	*51* (10)	**53** (09)	51 (10)	50 (10)	50 (10)	50 (10)
ELS (Res)	9	*46* (18)	**47** (18)	44 (21)	45 (21)	45 (22)	44 (21)
MPD (Res)	9	*18* (13)	15 (14)	15 (14)	14 (14)	**22** (14)	12 (13)
PFB (Res)	9	*20* (10)	18 (12)	13 (11)	20 (11)	**34** (10)	18 (13)
Written	9	*54* (13)	**55** (13)	51 (18)	54 (13)	54 (12)	52 (17)
Ratings	9	*12* (09)	11 (08)	09 (10)	09 (10)	09 (08)	05 (08)
HO-Total	9	*50* (14)	**52** (13)	49 (14)	49 (15)	48 (14)	49 (15)
JK-Total	9	*71* (08)	**72** (08)	71 (09)	71 (08)	71 (08)	71 (08)

Note. Corrected for range restriction, and adjusted for shrinkage (Rozeboom formula 8). Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Multiple Rs for ASVAB Factors alone are in italics. Underlined numbers denote multiple Rs greater than for ASVAB Factors alone. Decimals omitted.

[a] Number of MOS for which validities were computed.

Table 1.28
Mean of Multiple Correlations Computed Within Job for LVI Setwise Deletion Samples for Spatial, Computer, JOB. ABLE Composites, and AVOICE

Criterion	No. of MOS[a]	Spatial [1]	Computer [8]	JOB [3]	ABLE Composites [7]	AVOICE [8]
CTP (Raw)	9	58 (11)	49 (16)	31 (13)	21 (09)	39 (07)
GSP (Raw)	8	65 (06)	55 (08)	32 (13)	24 (14)	38 (07)
ELS (Raw)	9	33 (08)	30 (15)	19 (14)	12 (11)	20 (12)
MPD (Raw)	9	14 (11)	10 (16)	06 (13)	15 (11)	05 (11)
PFB (Raw)	9	08 (04)	13 (07)	07 (06)	28 (07)	09 (09)
CTP (Res)	9	43 (15)	31 (22)	17 (12)	10 (11)	29 (09)
GSP (Res)	8	51 (08)	40 (10)	21 (11)	14 (12)	28 (09)
ELS (Res)	9	41 (13)	36 (20)	24 (15)	21 (15)	26 (06)
MPD (Res)	9	13 (12)	10 (16)	06 (11)	15 (11)	07 (13)
PFB (Res)	9	11 (08)	10 (11)	09 (06)	30 (10)	12 (10)
Written	9	51 (11)	46 (16)	31 (17)	25 (11)	32 (15)
Ratings	9	09 (08)	09 (09)	07 (08)	04 (06)	03 (07)
HO-Total	9	50 (11)	38 (15)	20 (13)	13 (11)	30 (07)
JK-Total	9	66 (07)	60 (10)	38 (14)	30 (08)	43 (08)

Note. Corrected for range restriction and adjusted for shrinkage (Rozeboom formula 8). Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Decimals omitted.

[a] Number of MOS for which validities were computed.

Table 1.29
Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for LVI Setwise Deletion Samples for Spatial, Computer, JOB, ABLE Composites, and AVOICE

Criterion	No. of MOS[a]	ASVAB Factors (A4) + Spatial [5]	A4+ Computer [12]	A4+ JOB [7]	A4+ ABLE Composites [11]	A4+ AVOICE [12]
CTP (Raw)	9	64 (10)	61 (11)	63 (11)	61 (12)	64 (11)
GSP (Raw)	8	69 (06)	66 (07)	67 (07)	66 (08)	66 (07)
ELS (Raw)	9	37 (10)	36 (14)	37 (11)	36 (13)	36 (11)
MPD (Raw)	9	15 (13)	15 (15)	12 (13)	24 (13)	11 (14)
PFB (Raw)	9	15 (08)	17 (05)	17 (07)	32 (04)	15 (10)
CTP (Res)	9	48 (12)	45 (14)	46 (14)	45 (14)	47 (14)
GSP (Res)	8	54 (06)	50 (08)	51 (08)	50 (07)	50 (07)
ELS (Res)	9	47 (12)	43 (20)	46 (15)	46 (15)	46 (14)
MPD (Res)	9	14 (13)	13 (15)	13 (13)	22 (12)	11 (14)
PFB (Res)	9	20 (11)	18 (11)	20 (10)	36 (08)	21 (11)
Written	9	57 (13)	53 (17)	58 (12)	55 (13)	54 (18)
Ratings	9	10 (09)	11 (11)	11 (09)	11 (07)	06 (09)
HO-Total	9	53 (09)	49 (11)	50 (12)	49 (11)	50 (11)
JK-Total	9	73 (08)	71 (09)	72 (08)	71 (09)	71 (09)

Note. Corrected for range restriction and adjusted for shrinkage (Rozeboom formula 8). Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Underlined numbers denote multiple Rs greater than for ASVAB Factors alone. Decimals omitted.

[a] Number of MOS for which validities were computed.

Comparison of Validity Research in LVI and CVI Samples

The final set of results concern the comparison between the validity estimates associated with the longitudinal data (i.e., LVI) and those reported for the concurrent validation data (CVI). Table 1.30 reports the multiple correlations for the ASVAB factors and each set of experimental predictors as computed for the listwise sample in both data sets.

The results in Table 1.30 demonstrate that the patterns and levels of validities are very similar across the two sets of analyses. Still, there are some differences worth pointing out. Specifically, in comparison to the results of the CVI analyses: (a) The LVI

validities of the "cognitive" predictors (i.e., ASVAB, spatial, computer) for predicting the "will do" performance factors (ELS, MPD, and PFB) are higher; (b) the LVI validities of the ABLE composites for predicting the "will do" performance factors are somewhat lower; and (c) the LVI validities of the AVOICE composites for predicting the "can do" performance factors (CTP and GSP) are higher.

Table 1.30
Comparison of Mean Multiple Correlations Computed Within Job for LVI and CVI Listwise Deletion Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE

		ASVAB Factors		Spatial		Computer		JOB		ABLE Comp.		AVOICE	
Criterion	No. of MOS[a]	LV [4]	CV [4]	LV [1]	CV [1]	LV [8]	CV [6]	LV [3]	CV [3]	LV [7]	CV [4]	LV [8]	CV [6]
CTP (Raw)	9	63	63	57	56	50	53	31	29	27	26	41	35
GSP (Raw)	8	67	65	64	63	57	57	32	30	29	25	40	34
ELS (Raw)	9	39	31	32	25	34	26	22	19	20	33	25	24
MPD (Raw)	9	22	16	14	12	15	12	11	11	22	32	11	13
PFB (Raw)	9	21	20	10	10	17	11	12	11	31	37	15	12
CTP (Res)	9	48	47	42	37	35	37	20	21	18	22	33	28
GSP (Res)	8	53	49	51	48	44	41	22	22	19	21	31	26
ELS (Res)	9	48	46	41	41	40	38	25	27	26	31	29	32
MPD (Res)	9	23	19	14	15	14	13	12	10	21	28	13	15
PFB (Res)	9	24	21	12	11	17	14	11	10	32	35	16	14
Written	9	56	62	49	55	47	54	31	28	29	21	33	32
Ratings	9	16	15	09	07	17	08	10	08	09	18	09	09

Note. Corrected for range restriction and adjusted for shrinkage (Claudy formula). Numbers in brackets are the numbers of predictor scores entering prediction equations. Decimals omitted.

[a] Number of MOS for which validities were computed.

Further Exploration of ELS and ABLE

As shown in the data reported above, the largest difference between the CVI and LVI validation results was in the prediction of the Effort and Leadership (ELS) performance factors with the ABLE basic scores. Corrected for restriction of range and for shrinkage, the validity of the four ABLE composite scores in CVI was .33 for ELS and the validity of the seven ABLE factor scores in LVI was .20. When cast against the variability in results across studies in the extant literature, such a difference may not seem all that large or very unusual. However, since the obtained results from CVI, CVII, and LVI have been so consistent, in terms of the expected convergent and divergent results, we subjected this particular difference to a series of additional analyses in an attempt to determine the reason for the discrepancy.

First, the discrepancy does not seem to arise from any general deterioration in the measurement properties of either the ABLE or the ELS composite in the LVI sample. For example, while the correlation of the ABLE with ELS and MPD went down, the ABLE's correlations with CTP and GSP went up slightly. Similarly, a decrease in the validity with which ELS was predicted was characteristic only of the ABLE. The validities of the cognitive measures, the JOB, and AVOICE for predicting ELS actually increased by varying amounts. Consequently, the decrease in validity seems to be specific to the ABLE/ELS correlation and, to a lesser extent, the ABLE/MPD correlation.

The followup analyses were also able to rule out two possible additional sources of the CVI/LVI validity differences. First, differences in the composition and number of ABLE basic scores from CVI to LVI did not account for the differences in patterns of validity. Second, differences in the composition of the Effort/Leadership factor score from CVI to LVI did not account for differences in validity.

Rather, the somewhat lower correlation of ABLE with Effort/Leadership in LVI seems due to the joint effects of two influences. First, the determinants of ELS scores seem to favor ability slightly more and motivation slightly less in LVI versus CVI, perhaps because their true score variances were different across the two cohorts. Second, the greater influence of the social desirability response tendency in LVI seems to produce more positive manifold (i.e., higher intercorrelations for the LVI ABLE basic scores), as contrasted with CVI. This could also lower the correlation of the regression-weighted ABLE composite with ELS, whereas it might not have the same effect with the Core Technical and General Soldiering factors.

Another potential explanation that cannot be ruled out is a specific pattern of changes in true scores on the ABLE which produce the LVI pattern of correlations. One such specific change is suggested by a separate analysis of the ABLE carried out by White and Moss (1995). The items on the ABLE which had an Army frame of reference were shown to have higher item validities in CVI than in LVI.

Summary of LVI Validation

Generally speaking, the ASVAB was the best predictor of performance. However, the composite of spatial tests provided a small amount of incremental validity for the "can do" criteria (1-3 points), and the ABLE provided larger increments (7-20 points) for two of the three "will do" criteria (Maintaining Personal Discipline, and Physical Fitness and Bearing). Estimates of incremental validity were somewhat higher when the results were not corrected for range restriction.

With regard to ASVAB scoring options, results indicate a very slight edge for using multiple regression equations based on the four ASVAB unit-weighted factor scores. In the test of ABLE scoring options, the method using factor scores computed from a subset of all the ABLE items (ABLE-114) proved to have consistently slightly higher validities.

Perhaps the most interesting finding is derived from the comparisons between the Longitudinal Validation results and those from the Concurrent Validation. Generally speaking, the pattern and level of the validity coefficients were highly similar across the two samples. The correlation between the CV and LV coefficients in Table 1.30 was .962 and the root mean-squared difference between the two sets of coefficients was .046. However, the correlation is not 1.00. As noted above, the longitudinal validities were higher for cognitive predictors against "will do" criteria and lower for ABLE composites against "will do" criteria. Some of the possible explanations for those differences include changes in the nature of predictor scores when administered in a longitudinal versus concurrent design, changes in criterion or predictor scores due to cohort differences, and changes in the true relationship between abilities and performance as persons gain more experience and training in an organization and job. These and other possible explanations will be explored in future analyses.

Results of the Concurrent Sample Second-Tour Validation (CVII)

The CVII validation results are based on the CVII sample, which was assessed on the criterion measures of second-tour performance at the same time that the LVI performance data were collected from the first-tour longitudinal sample. The predictor set is limited to ASVAB and ABLE because only a small proportion (approximately 12%) of the CVII sample had been assessed with the Experimental Predictor Battery. ASVAB scores, taken 5-6 years earlier, were available from the Enlisted Master File. The ABLE was administered concurrently during the CVII data collection to approximately 45 percent of the total sample (i.e., those individuals who had no peers in the sample to rate and thus had time to take the ABLE). Everyone in the sample was assessed on the full set of second-tour performance measures. By design, the MOS in the CVII sample were limited to the MOS in Batch A. Because of the generally small samples for individual MOS, results for most analyses are reported for the combined sample.

The CVII data collection and data presentation are described in the first annual report for Building the Career Force (Campbell & Zook, 1990; see Chapters 5 and 6). After final editing, the total N for CVII was 1,053. The total sample was distributed across the Batch A MOS as shown in Table 1.31.

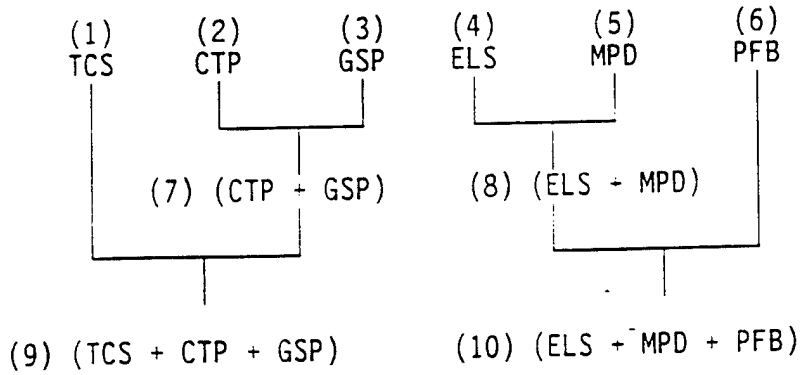
Table 1.31
CVII Sample Sizes by MOS

MOS	N
11B Infantryman	127
13B Cannon Crewmember	162
19E M60 Armor Crewman	33
19K M1 Armor Crewman	10
31C Single Channel Radio Operator	103
63B Light-Wheel Vehicle Mechanic	116
71L Administrative Specialist	112
88M Motor Transport Operator	144
91A Medical Specialist	146
95B Military Police	141
Total	1,053

Because of some missing data, the sample sizes varied depending on the specific analysis being reported. For example, for the reasons cited above, ABLE scores were available for only 477 individuals. All the analyses that require a common covariance matrix for ABLE and ASVAB were based on this reduced sample.

The development of the CVII performance measures, and the analysis and modeling of CVII performance, all have been described previously (Campbell & Zook, 1990) and are summarized in a previous section of the present chapter. The solution that yielded the best fit consisted of six substantive factors and two methods factors. The two methods factors were defined to be orthogonal to the substantive factors, but the correlations among the substantive factors were not so constrained. The six substantive factors and two methods factors, and the variables that are scored on each, were shown in Figure 1.9.

The complete basic validation analyses utilized a total of 10 scores for the performance factors, as shown below.



That is, all 10 scores were used as criterion measures. All higher order composite scores were obtained by standardizing the component scores and then taking the simple sum.

Procedure

The CVII validation analysis procedure consisted of the following steps:

(1) The ASVAB and ABLE were correlated with the six performance factor scores, their five residual scores (there was no residual for TCS), the higher order factor composites, the two methods factor scores, and the total score from the hands-on tests, the job knowledge tests, and the Situational Judgment Test. ASVAB was represented by the AFQT, a regression-weighted composite of the four factors, and a regression-weighted composite of the nine subtests. ABLE was represented by the three alternative sets of scores described previously. Both corrected (for multivariate restriction of range) and uncorrected estimates were computed, and both regression weights and unit weights (applied to standardized scores) were used. When multiple regression weights were used, the Rozeboom correction (Rozeboom, 1978) was used to account for the fitting of error.

(2) As in CVI, incremental validities for the ABLE composites over the ASVAB composites were also computed against each criterion score.

(3) A hierarchical regression analysis, stopping at six predictors, was run against each performance factor, factor composite, and individual criterion score (i.e., hands-on, job knowledge, and Situational Judgment Test).

(4) A hierarchical regression analysis was also carried out on selected criterion variables for the combined samples from three MOS clusters. The clusters were based on the results of an MOS clustering within the Synthetic Validation Project (Wise, Peterson, Hoffman, Campbell, & Arabian, 1991) and on the results of the validity generalization analysis for the Batch A MOS in the CVI sample (Wise, McHenry, & Campbell, 1990).

(5) The final step consisted of using the optimal six variable equations from the hierarchical regression analyses described above to develop a picture of the degree of differential prediction across performance factors and across the three MOS clusters.

Results

The basic multiple correlations for ASVAB (four factors vs. nine subtests) and ABLE (seven theoretically based composites vs. seven "purified" empirical factors) are given in Table 1.32. Several things are worth noting. ASVAB, taken at time of entry, is still a highly valid predictor of Core Technical and General Soldiering Proficiency and has respectable validity for Effort/Leadership. For ASVAB, the four factors and the nine

Table 1.32
Multiple Correlations for ASVAB Factors, ASVAB Subtests, ABLE Composites, and ABLE-114 Scores Against 19 CVII Criterion Variables (All MOS), With Unit Weights

Variable	ASVAB Factors [4]	ASVAB Subtests [9]	ABLE Composites [7]	ABLE-114 [7]
Core Technical (Raw)	43 (42)	43 (43)	15 (14)	20 (15)
General Soldiering (Raw)	56 (54)	57 (55)	22 (16)	26 (18)
Effort/Leadership (Raw)	38 (38)	39 (38)	37 (32)	41 (32)
Personal Discipline (Raw)	00 (11)	00 (11)	20 (21)	18 (22)
Physical Fitness (Raw)	13 (16)	06 (16)	32 (23)	34 (21)
Training/Counseling (Raw)	06 (13)	00 (12)	27 (19)	23 (18)
Core Technical (Res)	29 (29)	28 (30)	00 (12)	07 (13)
General Soldiering (Res)	42 (42)	43 (42)	14 (15)	18 (16)
Effort/Leadership (Res)	25 (26)	27 (25)	38 (31)	41 (30)
Personal Discipline (Res)	00 (09)	00 (09)	16 (20)	15 (19)
Physical Fitness (Res)	16 (20)	09 (20)	34 (21)	35 (18)
ELS - No Situational Judgment	24 (22)	23 (22)	34 (31)	38 (30)
Criterion Composite CTP/GSP	57 (55)	58 (56)	22 (17)	27 (19)
Criterion Composite ELS/MPD	29 (30)	29 (29)	34 (32)	37 (32)
Criterion Factor 1 CTP+GSP+TCS	50 (50)	50 (50)	29 (22)	32 (23)
Criterion Factor 2 ELS+MPD+PFB	14 (16)	12 (15)	34 (35)	35 (34)
Hands-On Average	39 (40)	38 (40)	12 (12)	18 (13)
Job Knowledge Total	59 (56)	59 (57)	25 (14)	28 (16)
Situational Judgment	42 (43)	42 (43)	27 (20)	31 (21)

Note. N = 412. Adjusted (Rozeboom formula). Numbers in brackets are the number of predictor scores entering prediction equations. Validities of unit-weighted composites are in parentheses. Decimals omitted.

subtests provide virtually the same level of predictive accuracy. However, for ABLE the reduced factor scores (114 items) are consistently the best predictor set. ABLE predicts Effort/Leadership and Physical Fitness very well and has reasonable correlations with General Soldiering and Training/Counseling.

In general, after adjustments, regression weights and unit weights for ASVAB yield about the same level of validity. However, regression weights are somewhat better than unit weights for the seven empirical ABLE factors. There is not as much positive manifold among the ABLE factors as there is among the ASVAB subtests.

Table 1.33 contains the same type of incremental analyses that were done in CVI (Campbell & Zook, 1991). ABLE does not add to the prediction of Core Technical and General Soldiering Proficiency, but it adds about the same amount to the prediction of Effort/Leadership as it did in CVI. However, the overall level of prediction for ELS is higher in CVII than it was in CVI ($\underline{R}$ = .50 vs. .43).

The hierarchical procedure asked for the optimal six-variable equation. For any specific criterion measure the first four variables were never all from ASVAB or all from ABLE. It appears that ABLE, most frequently the Dependability scale, does play a role in predicting CTP and GSP. This contribution is masked when the non-hierarchical procedure is used.

Table 1.33
Multiple Correlations for ASVAB Factors Plus ABLE Composites and Plus ABLE-114 Scores, and for ASVAB Subtests Plus ABLE Composites and Plus ABLE-114 Scores Against 19 CVII Criterion Variables, All MOS

Variable	4 ASVAB Factors + 7 ABLE Comp (K=11)	4 ASVAB Factors + 7 ABLE-114 (K=11)	9 ASVAB Subtests + 7 ABLE Comp (K=16)	9 ASVAB Subtests + 7 ABLE-114 (K=16)
Core Technical (Raw)	.42	.43	.42	.43
General Soldiering (Raw)	.56	.57	.58	.58
Effort/Leadership (Raw)	.49	.49	.49	.50
Personal Discipline (Raw)	.16	.13	.09	.03
Physical Fitness (Raw)	.34	.35	.32	.33
Training/Counseling (Raw)	.26	.20	.24	.17
Core Technical (Res)	.24	.26	.24	.25
General Soldiering (Res)	.42	.42	.44	.44
Effort/Leadership (Res)	.43	.43	.43	.43
Personal Discipline (Res)	.09	.07	.00	.00
Physical Fitness (Res)	.36	.37	.34	.34
ELS - No Situational Judgment	.39	.41	.38	.41
Criterion Composite CTP/GSP	.57	.57	.58	.58
Criterion Composite ELS/MPD	.40	.40	.40	.40
Criterion Factor 1: CTP+GSP+TCS	.54	.54	.54	.54
Criterion Factor 2: ELS+MPD+PFB	.35	.35	.34	.34
Hands-On Average	.37	.37	.37	.37
Job Knowledge Total	.60	.60	.60	.60
Situational Judgment	.45	.44	.45	.44

Note. N = 412. Corrected for range restriction and adjusted (Rozeboom formula).

Generalizability

A descriptive picture of the generalizability of prediction equations across performance factors (for the combined sample) is shown in Table 1.34. All entries are multiple correlations and the diagonals represent estimates based on optimal weights. Estimates of what happens when less than optimal weights are used to predict the same criterion are obtained by looking across the rows. Estimates of what happens when a particular set of weights is applied to other criterion measures or other MOS are obtained by looking down the columns. All estimates are based on the corrected covariance matrix. The diagonals are adjusted for shrinkage using the Rozeboom formula with k = 6. The off-diagonals are not adjusted because the weights were not computed against that particular dependent variable.

As shown in Table 1.34, within MOS there is very little differential validity for Core Technical vs. General Soldiering Proficiency. Either set of weights works about as well. However, the same is not the case for the other four performance factors. Better prediction is always achieved by using the equation developed for each factor.

Summary of LVII Validity Estimates

In general, in spite of the small samples for each MOS and the necessity of regarding all mean criterion differences as error (i.e., standardizing criterion scores within MOS), the validities for ASVAB and ABLE were as high, or higher, for predicting second-tour performance as for predicting first-tour performance. While unit weights did not weaken the validities for ASVAB, they did constrain the predictive accuracy of ABLE.

A consistent finding from the hierarchical analysis is that for the Core Technical Proficiency, General Soldiering Proficiency, and Effort/Leadership criteria, the optimal predictor battery is never composed of only ASVAB or only ABLE factor scores. For example, the Dependability factor from the ABLE is a consistent predictor of the "can do" component of performance.

Finally, based on the above analyses, there appears to be more differential validity across MOS for the second-tour samples than was found during the analyses of the first-tour data in CVI.

All of these issues can be analyzed more rigorously when the larger samples and fuller set of predictor measures from the second-tour longitudinal (LVII) validation are analyzed.

Table 1.34
Multiple Correlations for 10 Sets of Criterion Composite Weights, All MOS

	Raw CTP Weights	Raw GSP Weights	Raw ELS Weights	Raw MPD Weights	Raw PFB Weights	Raw TCS Weights	CTP+GSP Weights	ELS+MPD Weights	Criterion 1 Weights	Criterion 2 Weights
Core Technical (Raw CTP)	**.451** **(.429)**	.436	.331	.165	.193	.195	.446	.298	.434	.154
General Soldiering (Raw GSP)	.553	**.571** **(.557)**	.422	.173	.232	.288	.568	.367	.561	.192
Effort/ Leadership (Raw ELS)	.368	.370	**.500** **(.482)**	.375	.046	.358	.372	.489	.404	.422
Personal Discipline (Raw MPD)	.083	.069	.169	**.226** **(.169)**	.057	.130	.075	.197	.094	.188
Physical Fitness (Raw PFB)	.171	.163	.037	.100	**.401** **(.375)**	.055	.168	.059	.135	.235
Training/ Counseling (Raw TCS)	.119	.139	.197	.159	.038	**.275** **(.231)**	.131	.196	.175	.178
CTP+GSP	.572	.574	.429	.193	.242	.276	**.578** **(.564)**	.379	.567	.197
ELS+MPD	.272	.265	.403	.359	.061	.293	.270	**.412** **(.387)**	.301	.366
Criterion 1[a]	.514	.524	.431	.223	.180	.339	.524	.390	**.533** **(.517)**	.235
Criterion 2[b]	.129	.128	.319	.314	.222	.245	.129	.336	.167	**.378** **(.350)**

Note. Rows are criteria; columns are weights corrected for range restriction; multiple R for optimal weights in bold; Rozeboom adjustments in parentheses.

[a] Criterion Factor 1 = CTP+GSP+TCS.

[b] Criterion Factor 2 = ELS+MPD+PFB.

Prediction of Second-Tour Performance From the Trial Battery and From First-Tour Performance

The fundamental research designs for Project A and Career Force include the concept of combining successive pieces of information from (a) predictor tests administered at entry, (b) measures of performance during training, and (c) measures of first-tour job performance to predict individual performance in the second tour of duty. Pending fuller analyses as larger samples become available from later data collections, a preliminary explanation was conducted with the small samples from CVII available at this stage.

These analyses of CVI and CVII data examine the relationship of ASVAB scores (from tests given at the time recruits entered the Army), the CVI predictor scores (i.e., the Project A CVI Trial Battery, the preliminary version of the Experimental Predictor Battery, given during the first tour), and first-tour job performance scores to second-tour CVII job performance scores. Two complications with these initial analyses were that available sample sizes were extremely small, and it was unclear exactly how to account for range restriction for a sample of this type.

There were 121 soldiers in Batch A MOS who had been assessed on at least a subset of measures during the CVI and CVII data collections. Not all 121 soldiers had complete CVI and CVII data; the lowest number available for a given combination of CVI and CVII measures was 102. Table 1.35 shows the number of soldiers who had CVI and CVII data, by MOS.

Table 1.35
Numbers of Soldiers With CVI and CVII Data by MOS: Initial Sample

MOS		N
11B	Infantryman	8
13B	Cannon Crewmember	26
19E	M60 Armor Crewman	4
31C	Single Channel Radio Operator	8
63B	Light-Wheel Vehicle Mechanic	25
71L	Administrative Specialist	15
88M	Motor Transport Operator	7
91A/B	Medical Specialist	15
95A	Military Police	13
	Total	121

Measures

The second-tour performance criterion CVII measures used in the exploratory analysis were the raw and residual scores for the five constructs first identified during the first-tour Concurrent Validation, and confirmed by the CVII modeling analysis. Predictor measures came from the ASVAB, from the Project A CVI Trial Battery, and from first-tour job performance measures. Because of the extremely limited sample sizes, the least-squares weights developed for the CVI criterion constructs were used rather than developing new weights for CVII criterion constructs.

Analysis and Results

CVI predictor scores were correlated with the CVII criterion scores in two ways: (a) Correlations were computed within each MOS and these values were averaged (weighted by N), and (b) correlations were computed across the total sample. Correlations with CVII criteria were computed separately for the ASVAB, spatial, computer-administered, ABLE, AVOICE, and JOB composites and for the CVI criterion scores. Correlations were also computed for the ASVAB plus each of the other predictor sets from the Trial Battery and the CVI criteria. When the CVI criteria were combined with any of the other predictor scores, they were standardized within MOS (using the larger CVI samples to compute standard scores) and summed to achieve equal weighting between ASVAB/Trial Battery and CVI criterion scores.

Because of the number of different points at which additional range restriction could occur, there are a number of different "populations" to which the CVII sample could be corrected. If the problem is to select second-tour soldiers from experienced first-tour personnel, then the set of all persons who are nearing completion of the first tour seems the most appropriate population.

The correlations of scores on the first-tour criteria with scores on second-tour criteria in the combined sample are shown in Table 1.36. The correlations are not corrected for restriction of range. The note for the table shows the mean of the diagonal correlations, which contains the correlations of the same criteria across first and second tour--that is, the correlation of Core Technical between first and second tour, and so on. This mean is an index of convergent validity for the set of criterion constructs. The note also shows the mean of the off-diagonal correlations--that is, the correlations between different criterion constructs across first and second tour. The difference between the mean diagonal and mean off-diagonal correlation can be thought of as an indicator of discriminant validity.

The correlations of predicted scores based on CVI weights for ASVAB and Trial Battery composites and CVI criterion scores with CVII criteria are shown in Table 1.37. On the whole, of all the predictors, the CVI criterion scores have the highest correlations with CVII criterion scores. However, adding the ASVAB and the ASVAB plus Trial Battery composite scores to CVI scores does increment the CVI validity coefficients.

Table 1.36
Uncorrected Correlations Between CVI and CVII Raw Criterion Composites Computed Across MOS: Initial Sample

CVI Criterion Composite	CVII Criterion Composite				
	CTP	GSP	ELS	MPD	PFB
Core Technical Proficiency	.47	.48	.22	.10	.08
General Soldiering Proficiency	.47	.43	.36	.13	.17
Effort and Leadership	.19	.07	.30	.19	.13
Maintaining Personal Discipline	.06	.14	.16	.26	.19
Physical Fitness and Military Bearing	.00	-.04	.15	.15	.48

Note. Ns = 102-121. Correlations between matching variables are underlined.
Mean diagonal value = .39; mean off-diagonal value = .17.

The ASVAB validities follow the familiar pattern of predicting the two "can do" criteria, but not predicting the "will do" criteria very well. The JOB unexpectedly did the best job of predicting Maintaining Personal Discipline.

In sum, the results with this small initial sample provide evidence that ASVAB scores, weighted on the basis of regression estimates for predicting first-tour performance, predict second-tour "can do" performance with substantial validity. The results also provide evidence of convergent and discriminant validity of the first-tour job performance for predicting second-tour job performance criteria.

Future analyses of the LVI Experimental Predictor Battery and LVII criterion scores will provide better indications of the new predictors' relationships with second-tour performance.

Table 1.37
Correlations Between CVI Weighted Predictor Composites, CVI Criterion Composites, and CVII Criterion Composites for Raw Scores, Computed Across MOS: Initial Sample

Predictor and CVI Criterion Composites and Combinations	CVII Criterion Composite				
	CTP	GSP	ELS	MPD	PFB
ASVAB	.33	.42	.11	-.05	.11
CVI Performance	.47	.43	.30	.26	.48
ASVAB+CVI Performance	.51	.51	.33	.26	.47
Computer Tests	.23	.13	-.01	-.04	.10
ASVAB+Computer Tests	.37	.41	.13	.05	.12
ASVAB+Comp. Tests+ CVI Performance	.52	.51	.33	.27	.46
AVOICE	.15	.16	.06	-.02	.06
ASVAB+AVOICE	.43	.44	.14	.00	.13
ASVAB+AVOICE+CVI Performance	.54	.52	.33	.27	.46
JOB	.12	.00	.19	.30	.12
ASVAB+JOB	.33	.41	.16	.20	.16
ASVAB+JOB+CVI Performance	.51	.51	.34	.31	.48
Spatial	.47	.41	.14	-.01	.04
ASVAB+Spatial	.41	.43	.10	-.06	.11
ASVAB+Spatial+CVI Performance	.52	.51	.33	.26	.46
ABLE	.10	.01	.21	.15	.29
ASVAB+ABLE	.34	.41	.22	.12	.25
ASVAB+ABLE+CVI Performance	.51	.52	.36	.30	.47

Note. Ns = 102-121. Correlations are uncorrected for range restriction. Coefficients do not require shrinkage adjustments. CVI criterion scores and predictor composites were summed.

SUMMARY OF PROJECT EFFORTS FOR YEAR THREE

As described in the third annual report (Campbell & Zook, 1994a), the Project had four main objectives during FY92:

(1) Complete the Longitudinal Second-Tour (LVII) data collection.

(2) Prepare the LVII data file for analysis.

(3) Analyze the LVII criterion data to develop the basic performance scores.

(4) Model the covariance structure of the LVII performance scores.

In general, because the LVII results could be viewed as a major replication of CVII, much effort was devoted to using the LVII sample data in a confirmatory way. That is, when possible, the CVII results were used as a hypothesis to be tested in LVII.

Longitudinal Validation Second-Tour Data Collection

The LVII data collection administered second-tour criterion measures to soldiers in the longitudinal validation sample who had reenlisted for a second tour and who were available for assessment at a specified set of data collection locations. Although this data collection involved substantially fewer soldiers than the CVI or LVI data collections, it posed a number of special challenges. Having to locate and test particular individual soldiers, especially when there were relatively few to begin with, was a difficult task. However, despite a major deployment of U.S troops to Southwest Asia (Operation Desert Shield/Storm), LVII data were collected from 1,577 soldiers.

A list of the instruments administered in the LVII data collection is provided in Table 1.38. Most of the instruments served as second-tour performance criterion measures; several other instruments (e.g., the Background Information Form) provided supplemental data for the project.

The original project plan called for the LVII data collection to take place July-December 1991. Second-tour criterion data were to be collected from at least 150 soldiers in each of nine MOS (the Batch A group designated in previous collections).

Even before the deployment of troops to Southwest Asia, the anticipated data collection problems included difficulty projecting future location of soldiers targeted for testing because of frequent reassignments, and limited access due to training or alert status, leave, and so forth. The problems were compounded by a tasking system which requires that Troop Support Requests (TSRs) be submitted well in advance of data collection (135 days up to 500 days for U.S. Forces Command, less for the U.S. Training and Doctrine Command). Moreover, before detailed data collection planning activities began, the Army was starting to respond to directives to downsize and to reduce the proportion of troops stationed in Germany.

These concerns led to the following strategy for maximizing the number of LVII subjects. In May 1990, analyses were conducted to determine the number of Project A soldiers who were still in the Army and their locations. It was clear that sample size requirements would not be met if only soldiers having predictor and first-tour criterion data were tested. Accordingly, the decision was made to test soldiers for whom predictor and/or first-tour criterion data were available. Only soldiers with Project A data were eligible. An additional consideration was the fact that, shortly after the LVI data collection ended, MOS 31C began declining in strength because certain radio equipment was being phased out. The collection of hands-on data is inordinately resource-intensive for small numbers of examinees; consequently, hands-on tests were dropped from the performance measures for the 31C soldiers.

Table 1.38
LVII Data Collection Instruments

Performance Criterion Instruments

- Job Knowledge Tests
- Hands-On Tests
- Performance Rating Scales (completed by supervisors)
  - Army-Wide Booklet
  - MOS-Specific Booklet
  - Combat Performance Prediction Scales
  - Combat Performance Questionnaire (Operation Desert Shield/Storm), administered if applicable
- Personnel File Form (PFF)
- Situational Judgment Test (SJT)
- Supervisory Simulation Exercises
  - Personal Counseling
  - Disciplinary Counseling
  - Training

Supplemental Instruments

- Background Information Form
- MOS-Specific Job History Questionnaire
- Supervisory Experience Questionnaire
- Army Job Satisfaction Questionnaire (AJSQ)
- Assessment of Background and Life Experiences (ABLE)
- Leader and Unit Attitudes Questionnaire

The May 1990 analyses also indicated that appreciable concentrations of Project A soldiers were stationed in locations other than those identified in the original research plan. Accordingly, requests for troop support were written to include some of these new sites. Then, lists of all soldiers eligible for testing were electronically matched with each installation's own personnel files, to obtain the most accurate identification of soldiers qualified for testing at each location.

Data Collection Schedule

The original research plan, calling for LVII data collection July-December 1991, was adjusted to accommodate the interests of supporting commands. It was agreed that test sites could be scheduled to conduct testing as early as May 1991 and as late as February 1992.

This data collection strategy had been established before hostilities involving U.S. troops in Southwest Asia arose. The U.S. Forces Command, which was to provide the majority of LVII soldiers, invoked a moratorium on research support in September 1990. The moratorium was lifted in April 1991, and the data collection schedule was again modified. The final schedule is shown in Table 1.39. Four data collection teams were sent to Germany, whereas one team of data collectors was sent to each of the other test sites. The first LVII data collection occurred in June 1991 and the last in July 1992.

Composition of the teams, in terms of project staff, varied from location to location. Generally however, each test site was staffed with a team comprised of the following personnel:

1 Test Site Manager (TSM)
1-2 Hands-on Managers (HOMs)
3 Test Administrators (TAs)

All of these positions were filled by permanent employees of the contractor consortium; often a representative of the Army Research Institute was present during the testing. The Army installations also provided personnel to help support the data collection activities. In addition to the test site POC, each test site provided eight senior NCOs for each MOS (except 31C) to administer and score the hands-on tests, and two to four NCOs to fill general supporting roles (e.g., to track down soldiers who failed to report for testing and to handle problems with defective equipment).

Data Collection Team Training

One day of classroom training and considerable follow-up on-the-job training were provided to TAs for the written test and supervisor rating procedures. One to two days of additional training were provided to each TA for each subordinate role a TA was responsible for playing in the Situational Judgment Test. Two documents were developed to support TA training: the Test Administrator's Manual and the Supervisory Role-Play Exercises Administration Manual.

NCO hands-on scorers were trained the day before the administration of the hands-on tests to soldiers in a given MOS. The training followed the same basic procedures as those that had been used in the CV and LVI/CVII data collections (R. Campbell, 1985).

Various procedures and documents were used to handle completed data collection instruments before they were shipped to the facility where they would be processed and keypunched. Test site personnel checked measures for completeness and legibility, and documented explanations for data that were incomplete or otherwise anomalous. Transmittal documents were used to help ensure that data could be tracked once it left the test site.

After data collection at a given location was completed, the TSM prepared and submitted a report to ARI.

Table 1.39
LVII Data Collection Schedule

Command	Location	Test Dates
		1991
USAREUR	Germany	7 June - 27 June
USAREUR	Germany	5 July - 2 August
USAREUR	Germany	5 July - 3 August
Eighth Army	Republic of South Korea	5 July - 9 August
USAREUR	Germany	September - October
HSC	Fort Sam Houston, TX	October
FORSCOM	Fort Lewis, WA	9 December - 19 December
		1992
FORSCOM	Fort Drum, NY	13 January - 24 January
TRADOC	Fort Bliss, TX	20 January - 31 January
MDW & AMC	Fort Belvoir, VA	February
TRADOC	Fort Knox, KY	2 March - 6 March
FORSCOM	Fort Bragg, GA	16 March - 3 April
TRADOC	Fort Benning, GA	31 March - 3 April
FORSCOM	Fort Riley, KS	6 April - 10 April
FORSCOM	Fort Hood, TX	4 May - 15 May
FORSCOM	Fort Campbell, KY	11 May - 15 May
FORSCOM	Fort Carson, CO	1 June - 5 June
FORSCOM	Fort Stewart, GA	15 June - 23 June
TRADOC	Fort Polk, LA	13 July - 16 July

USAREUR	U.S. Army Europe
HSC	Health Services Command
FORSCOM	Forces Command
TRADOC	Training and Doctrine Command
MDW	Military District of Washington
AMC	Army Materiel Command

Development of Basic Scores for LVII Performance Measures

The LVII performance criterion measures have been described in detail elsewhere (Campbell, 1988; Campbell & Zook, 1990). They were originally administered to second-tour soldiers in the CVII sample and were subsequently revised in preparation for administration to the LVII sample.

Analyses of the data from the LVII sample had three major objectives: (a) examine and evaluate the psychometric properties of the LVII measures, (b) compare the psychometric properties of the LVII scores with the CVII scores, and (c) develop the basic criterion scores to be used in modeling second-tour performance.

Job Knowledge and Hands-On Tests

In earlier research a set of 28-30 tasks had been selected for performance measurement in each MOS. All tasks were assessed using a written job knowledge test format. Performance on a subset (14-17) of the tasks was assessed using a hands-on performance test format.

The full set of tasks included (a) common tasks, basic soldiering tasks that all soldiers are expected to know how to perform (e.g., first aid, personal weapons, map reading), and (b) MOS-specific tasks, central to the jobs of the soldiers working in a given MOS and typically unique to that MOS.

Some tasks are performed differently depending upon the type of equipment a soldier uses (e.g., an M16A1 rifle versus an M16A2 rifle). To deal appropriately with such situations, tracked (i.e., parallel) tests were prepared for tasks where the equipment varied.

Before the CVII measures could be used again, technical currency reviews were conducted. Revisions were made to test items and to supporting graphics and handouts as necessary.

Scoring Adjustments. Specifications for the basic scores for the LVII job knowledge and hands-on measures depended largely on previous work in CVI, CVII, and LVI. Job knowledge and hands-on task scores were calculated as percent-correct (or percent-GO) scores at all score levels. The data for tracked tests were examined for evidence of level and dispersion differences between tracks and no anomalous differences were found.

As with the previous data collections, hands-on test scores were standardized by site at the task level to control for site differences. One adjustment affected only the job knowledge tests; that is, between one and four items per MOS were dropped because of doctrinal changes subsequent to CVII.

Table 1.40 shows the overall number of items in the job knowledge component for each MOS and the range of items per task test. Table 1.41 shows the overall number of steps in the hands-on component for each MOS and the range of steps per task test.

After data editing, four levels of scores (Tasks, Functional Categories, Task Factors, and Task Constructs) were constructed. They are as depicted in Figure 1.10 in an earlier section of this chapter.

Table 1.40
Number of LVII Job Knowledge Tasks and Items by MOS

MOS		No. of Tasks	Items Dropped	Total Items	Items Per Task	Average Items Per Task
11B	Infantryman[a]	29	2	128	2-12	4.4
13B	Cannon Crewmember[a]	30	3	119-120	2-8	4.0
19K	M1 Armor Crewman	28	4	142	3-12	5.1
31C	Single Channel Radio Operator[a]	30	1	111-112	3-5	3.7
63B	Light Wheel Vehicle Mechanic	27	2	102	2-6	3.8
71L	Administrative Specialist[a]	30	2	125	2-12	4.2
88M	Motor Transport Operator	30	1	119	3-12	4.0
91A/B	Medical Specialist	30	3	113	2-6	3.6
95B	Military Police	29	4	109	2-7	3.8

[a] One or more task tests were tracked; tracked tests do not necessarily have the same number of items.

Table 1.41
Number of LVII Hands-On Tasks and Steps by MOS

MOS		No. of Tasks	Total Steps	Steps Per Task	Average Steps Per Task
11B	Infantryman	9	121	5-31	13.4
13B	Cannon Crewmember[a]	12	258-259	7-67	21.5-21.6
19K	M1 Armor Crewman	10	167	8-37	16.7
63B	Light Wheel Vehicle Mechanic[a]	8	142	7-44	17.8
71L	Administrative Specialist[b]	14	140-146	2-44	10.0-10.4
88M	Motor Transport Operator[a]	10	193-195	4-44	19.3-19.5
91A/B	Medical Specialist[a]	13	216	6-44	16.6
95B	Military Police[a]	10	223-227	7-37	22.3-22.7

[a] One or more task tests were tracked; tracked tests do not necessarily have the same number of steps.
[b] One task was scored on a continuous scale; it is not included in calculating total steps, steps per task, or average steps per task.

At each level of aggregation, hierarchical scores were computed using task-level data. That is, each category, factor, and construct score was computed by calculating the mean percentage of items correct (or percentage of steps passed) across all constituent tasks.

Final Basic Scores for Job Knowledge and Hands-On Measures. The descriptive statistics calculated across MOS for both the Task Construct and Task Factor scores do not differ much from the results for the CVII soldiers tested (reported in Campbell & Zook, 1990).

Task Factor (otherwise known as CVBITS) scores had been used in the performance modeling exercises conducted for CVI and LVI; however, Task Construct scores (i.e., MOS-Specific and General) were used for this purpose in CVII. Although Task Factors preserve somewhat more information than the more highly aggregated Task Construct scores, they have the disadvantage of differing across MOS as to the availability of each of the six scores. This problem is compounded by the considerably smaller sample sizes available for the two second-tour data collections relative to the two first-tour data collections. Moreover, in both CVI and LVI, the Technical Task Factor score invariably loaded on the Core Technical Proficiency performance construct while the other five Task Factor scores invariably loaded on the General Soldiering Proficiency performance construct. Therefore, the two Task Construct scores were selected for use in the LVII performance modeling exercise.

Performance Rating Scales

As reported previously (Campbell, 1988), the second-tour performance rating scales (with the exception of the Combat Performance Questionnaire) were developed by revising and adapting rating scales developed for first-tour soldiers. Based on results of the CVII data analyses, additional minor modifications were made to these three sets of scales: the Army-Wide ratings, the MOS-Specific ratings, and the Combat Performance Prediction scales.

Army-Wide Ratings. The Army-Wide rating booklet included 12 behavior-based dimensions, seven task-based leadership dimensions, a rating of overall effectiveness, and a rating of senior NCO potential.

Raters in the CVII sample had tended to make frequent use of the highest rating scale values, suggesting that the scale behavioral anchors may have been too lenient for more experienced soldiers. In the LVII sample, the behavioral anchors for most rating dimensions were revised to make the scale values reflect a slightly higher level of performance.

MOS-Specific Ratings. The MOS-Specific rating booklets included from 7 to 14 technically oriented behavior-based dimensions and a rating of overall MOS effectiveness. A set of scales suitable for second-tour MOS 19K soldiers were developed by adapting the second-tour MOS 19E scales that had been used in CVII. In five of the nine MOS, one or two of the MOS-specific dimensions measured some aspect of leadership (e.g.,

Leading the Team for MOS 11B). As with the Army-wide rating dimensions, the CVII behavioral anchors for most MOS-specific rating dimensions were revised to reflect slightly higher levels of performance.

Combat Performance Prediction Scales. The Combat Performance Prediction scales consisted of 14 items which depict examples of soldier behaviors under combat conditions. The rater's task was to estimate the likelihood that the ratee would behave as described in the behavioral example. Ratings were made on a 7-point scale ranging from very likely to very unlikely. The items were a subset of the 40 items that appeared on the original CVI version of the Combat Performance Prediction scales. Unlike the LVI/CVII data collections, LVII Combat Performance Prediction scale ratings were collected for both male and female soldiers.

Summary of Ratings Data. Across all nine MOS, two or more ratings were obtained for 75 percent of the soldiers (1,194 of 1,595) and at least one rating was obtained for 94 percent of the sample (1,494 of 1,595). The soldiers who received ratings averaged 1.82 raters per ratee.

Substantive analyses for the Army-wide and Combat Performance Prediction scale ratings were carried out on the total sample; MOS-specific ratings were, of course, analyzed separately by MOS. The analyses for the Army-wide and MOS-specific rating scales focused first on the distributions of the individual ratings and reliability estimates. This was followed by principal factor analyses with varimax rotation to determine the composition of basic scores.

Overall, the LVII rating distributions were as expected. The means were generally lower than in CVII and the variability was similar. The interrater reliability for the LVII ratings was almost exactly the same as that found in the LVI and CVII research.

Several factor analyses were conducted on the LVII sample. Army-wide ratings on the nine second-tour nonleadership dimensions were intercorrelated and factor analyzed so that the LVI and LVII factor structures could be compared. The same procedure was followed for all 19 of the Army-wide dimensions. For this analysis, the factor structure obtained in the LVII sample was compared to the factor structure obtained in the CVII sample.

The similarity of the rotated factor structures for LVI vs. LVII for the nine nonleadership/supervision dimensions that are common to the first-tour and second-tour rating scales was striking.

The four-factor rotated solutions obtained in the LVII and CVII samples for all 19 scales are shown in Table 1.42. Both include three factors that are quite similar to the three LVI factors, plus a separate leadership/supervision factor.

Table 1.42
Comparison of LVII and CVII Army-Wide Factor Analysis[a] Results: All Dimensions

Dimension	Factor Loadings (LVII/CVII) 1	2	3	4
1. Technical Knowledge/Skill	.47/.41	.23/.24	.26/.22	.56/.65
2. Effort	.45/.39	.34/.31	.26/.27	.58/.68
3. Supervising	.63/.57	.22/.21	.24/.28	.42/.53
4. Following Regs/Orders	.32/.29	.63/.63	.29/.30	.31/.36
5. Integrity	.38/.32	.58/.66	.24/.22	.34/.37
6. Training/Development	.60/.52	.20/.24	.27/.27	.38/.52
7. Maintain Equipment	.36/.32	.27/.33	.32/.25	.38/.50
8. Physical Fitness	.17/.20	.14/.18	.53/.60	.16/.19
9. Self-Development	.48/.41	.29/.27	.41/.44	.32/.48
10. Consideration for Subord	.61/.47	.40/.44	.16/.26	.28/.40
11. Military Bearing	.26/.30	.32/.34	.62/.63	.12/.22
12. Self-Control	.16/.17	.57/.56	.20/.18	.07/.09
13. Role Model	.51/.53	.37/.40	.56/.51	.25/.31
14. Communication	.61/.62	.39/.34	.22/.23	.26/.35
15. Personal Counseling	.74/.72	.24/.31	.27/.26	.11/.19
16. Monitoring	.68/.63	.18/.31	.30/.22	.30/.41
17. Organizing Missions/Operations	.66/.70	.22/.26	.27/.20	.30/.36
18. Personnel Administration	.65/.63	.28/.20	.22/.24	.17/.29
19. Performance Counseling	.72/.72	.22/.20	.23/.29	.24/.32
Percent Common Variance	45.3/37.6	25.4/20.3	18.2/16.9	16.9/25.3

Note. Sample size is 1,388 for LVII and 823 for CVII. CVII analyses based on supervisor ratings only.

[a] Principal factor analysis, varimax rotation.

Basic Scores. Factor analyses of the Army-wide ratings suggest that the four-factor CVII solution is appropriate for LVII. Accordingly, the four composites shown in Table 1.43 and the overall effectiveness rating were used as basic scores for the LVII Army-wide rating data.

Table 1.43
Composition of LVII Army-Wide Rating Composites

Percent Common Variance Accounted for by Relevant Factor	Composite Name	Dimensions Included
45.3	1. Leading/Supervising	Supervising Training/Development Consideration for Subord Communication Personal Counseling Monitoring Organizing Missions/Opers Personnel Administration Performance Counseling
25.4	2. Personal Discipline	Following Regs/Orders Integrity Self-Control
16.9	3. Technical Skill/Effort	Technical Knowledge/Skill Effort Maintain Equipment
18.2	4. Physical Fitness/ Military Bearing	Military Bearing Physical Fitness

Note. Two dimensions were not included in any composites: Acting as a Role Model and Self-Development.

Because the factor analysis results did not indicate multiple factors for any of the MOS-specific rating analyses, a unit-weighted composite of all dimension ratings for each MOS was used as the basic score. This is identical to the scoring system used in CVII, and yielded comparable reliability estimates.

Results of the principal component analysis for the Combat Performance Prediction scales on the combined LVII sample confirmed the findings that were obtained in LVI and CVII. Specifically, two factors were identified; however, the second

factor was simply a reflection of the first (i.e., it was comprised of the negatively worded items). The 14 items were summed to form a single basic score.

The mean LVII Combat Performance Prediction scale composite score and standard deviation were virtually identical to the mean and standard deviation of the supervisor ratings of soldiers in the CVII sample.

Administrative Measures: The Personnel File Form

The LVII Personnel File Form was used to gather self-reports of archival/ administrative information dealing with personnel actions reflective of individual performance. The first-tour versions (CVI and LVI) of the PFF requested information regarding (a) evidence of exemplary performance, including awards and memoranda/ certificates of appreciation, commendation, and achievement; (b) receipt of disciplinary actions (i.e., Articles 15 and flag actions); and (c) test results, including Physical Readiness test scores, individual weapon qualification scores, and Skill Qualification Test scores.

The original second-tour version of the PFF developed for CVII included these same types of variables and added others. The additional items were related to education (military training and civilian college courses) and promotions (e.g., how often recommended for accelerated promotion, number of promotion board points received). Another modification was to distinguish between awards, memoranda, and disciplinary actions received while in grades E-1 through E-3 and those received while in grades E-4 and above.

Before being administered to the LVII sample, the second-tour PFF was revised in several minor ways. Most of these revisions were intended to increase the interpretability/accuracy of responses and to reduce the amount of missing data. For example, the PFF response format was changed so that soldiers could indicate if they had earned more than one Army Achievement Medal.

Means and standard deviations for the administrative indices of performance are presented in Table 1.44. The corresponding descriptive statistics for CVII are not comparable for the Awards and Certificates score because of response format differences between the CVII and LVII instruments. Otherwise, the means and standard deviations for the LVII and CVII scores are very similar.

Subsequent analyses suggested that the basic scores tentatively derived for the PFF satisfactorily captured the useful information on that form. Consequently, they were made available for use in the second-tour performance modeling exercise.

Situational Judgment Test (SJT)

The SJT was designed to evaluate the effectiveness of NCO judgments concerning what the NCO should do in difficult supervisory situations. Thus, the SJT can be viewed

Table 1.44
Administrative Indices Descriptive Statistics for LVII and CVII

Measure		N	Mean	SD	Range
Awards and Certificates[a]	CVII	928	10.53	5.63	0-44
	LVII	1,577	14.69	6.79	0-40
Disciplinary Actions	CVII	930	.42	.87	0-8
	LVII	1,577	.37	.76	0-6
Physical Readiness Score	CVII	998	250.11	30.68	121-300
	LVII	1,522	248.81	31.27	23-288
Weapon Qualification	CVII	1036	2.52	.67	1-3
	LVII	1,565	2.58	.67	1-3
Promotion Rate	LVII	1,513	100.00	7.79	61-128
Promotion Rate (CVII Scoring)	CVII	901	100.14	8.09	67-121
	LVII	1,513	99.98	7.48	57-121

[a] Differences in LVII and CVII results reflect differences in PFF response format.

as a job knowledge test pertaining to the leadership/supervision components of second-tour positions.

As reported previously (Campbell, 1988), development of the SJT involved asking groups of soldiers similar to the target soldiers (i.e., at the level of SP4/SP5/Sergeant) to describe a large number of difficult but realistic situations that Army first-line supervisors face on their jobs. After a large number of these situations had been generated, a wide variety of possible actions (i.e., response alternatives) for each situation were gathered, and ratings of the effectiveness of each of these actions were collected from both experts (senior NCOs) and the target group. These effectiveness ratings were used to select situations and response alternatives to be included on the SJT.

The initial version of the SJT, which was administered to the CVII sample, consisted of 35 items. Because the CVII data analysis results indicated that the SJT was a promising measure of supervisory performance, this test was lengthened to 49 items for the LVII data collection to increase the internal consistency reliability and make it easier to identify SJT subscores.

Because the SJT had proven to be rather difficult for the CVII sample, an effort was made to select relatively less difficult additional items to include in the lengthened test for the LVII administration.

Analysis Procedure. Procedures for scoring the SJT were identical to those used in CVII. Five different scores were computed. The most straightforward was a simple number correct score. The second score involved weighting each response alternative by the mean effectiveness rating given to that response alternative by the expert group.

Two analogous scoring procedures based on respondents' choices for the least effective response to each situation were also used.

The fifth score involved combining the choices for the most and the least effective response alternatives into one overall score. For each item, the mean effectiveness of the response alternative each soldier chose as the least effective was subtracted from the mean effectiveness of the response alternative they chose as the most effective. Because it is actually better to indicate that less effective response alternatives are the least effective, this score can be seen as not a "difference" score but a simple sum.

Each of these scores was computed twice for the LVII soldiers, once using all 49 SJT items and once including only the 35 SJT items that had been administered to the CVII sample as well.

Descriptive statistics and internal consistency reliabilities were computed for each of the five scoring procedures for both the 49-item and the 35-item versions of the SJT. Intercorrelations were computed among the five scores generated by the five different scoring procedures for the 49-item SJT only. Finally, item analyses were conducted for each of the scoring procedures. (See Campbell & Zook, 1994b.)

Development of Subscales. Efforts to identify distinct SJT subscores in the CVII sample had not been particularly successful. However, the LVII version of the SJT contained almost 40 percent more items, and it was possible that a more interpretable solution would be found using the LVII data. In addition, a content analysis of the SJT items conducted by Hanson and Borman (in press) revealed some promising new subscales. Consequently, the dimensionality of the SJT for the LVII sample was investigated both rationally and empirically, with the primary goal to develop a set of more homogeneous SJT subscores.

The item-level responses from the LVII sample were well distributed across the response alternatives for each item, which suggests that the correct responses to SJT items were not obvious. All of the scoring procedures resulted in a reasonable amount of variability in both the LVII and CVII samples. The most reliable score for both samples is M-L Effectiveness, probably because this score contains the most information.

For the 49-item SJT, in which the maximum possible M-Correct score is 49, the mean in the LVII sample is only 25.84, indicating that this longer version of the SJT was also relatively difficult.

Based on the full array of descriptive statistics, the M-L Effectiveness score appeared to provide the best summary of the information contained in the SJT

responses. Thus, all remaining analyses focused on the M-L Effectiveness score, which became the SJT Total Score.

The rational/empirical analysis of the item covariances resulted in six factor-based subscales that contained between six and nine items each. Six remaining items were not included in any subscale. Definitions of these factor-based subscales are presented in Table 1.45. These subscales have potential for more clearly delineating the leadership/ supervision aspects of the second-tour soldier job. They are included in one of the major alternative models of second-tour performance to be evaluated in subsequent confirmatory analyses.

Supervisory Simulation Exercises

The supervisory simulation measures were designed to assess areas of second-tour job performance that deal with specific components of supervisor/subordinate interaction. These areas included personal counseling, disciplinary counseling, and one-on-one training. A trained evaluator (role player) acted out the role of a subordinate to be counseled or trained and the examinee assumed the role of a first-line supervisor who was to conduct the counseling or training. In each exercise, evaluators scored the examinees on a number of rating scales.

The subordinate and supervisor roles were essentially the same as those used in the CVII data collection. The role players who assumed the role of the subordinate in each of these exercises were trained to play the roles in a standardized fashion.

The rating system used to evaluate LVII examinees was modified in several ways from CVII. First, the CVII analyses identified the scales that (a) appeared to be difficult to rate, (b) conceptually redundant, and/or (c) not correlated with other rated behaviors in meaningful ways. These behavior ratings were dropped and anchors for some of the remaining scales were changed. The scale itself was expanded from 3 to 5 points. The overall effectiveness rating was retained, but the overall affect and fairness rating scales were eliminated. Thus, examinees were rated on each exercise on from 7 to 11 behavioral scales and on one overall effectiveness scale.

Another important difference between the CVII and LVII measures was the background of the evaluators. The smaller size of the LVII data collection allowed for the selection and training of role players evaluators who had been formally educated as personnel researchers and who were employed full-time by organizations in the project consortium. In contrast, the scope of the LVI/CVII data collection had required the hiring of a number of temporary employees to serve as role players.

Descriptive analyses were conducted, followed by a series of factor analyses. Maximum likelihood factor analyses with oblique rotations were performed within each exercise. The purpose of the factor analyses was to identify the content of basic criterion scores for each of the simulation exercises.

Table 1.45
Situational Judgment Test: Definitions of Factor-Based Subscales

1. Discipline soldiers when necessary (Discipline). This subscale is made up of items on which the most effective responses involve disciplining soldiers, sometimes severely, and the less effective responses involve either less severe discipline or no discipline at all. (Six items.)

2. Focus on the positive (Positive). This subscale is made up of items on which the more effective responses involve focusing on the positive aspects of a problem situation (e.g., a soldier's past good performance, appreciation for a soldier's extra effort, the benefits the Army has to offer). (Six items.)

3. Search for underlying reasons (Search). This subscale is made up of items on which the more effective responses involve searching for the underlying causes of soldiers' performance or personal problems rather than reacting to the problems themselves. (Eight items.)

4. Work within the chain of command and with supervisor appropriately (Chain/Command). For a few items on this subscale the less effective responses involve promising soldiers rewards that are beyond a direct supervisor's control (e.g., "comp" time). The remaining items involve working through the chain of command appropriately. (Six items.)

5. Show support/concern for subordinates and avoid inappropriate discipline (Support). This subscale is made up of items where the more effective response alternatives involve helping the soldiers with work-related or personal problems and the less effective responses involve not providing needed support or using inappropriately harsh discipline. (Eight items.)

6. Take immediate/direct action (Action). This subscale is composed of items where the more effective response alternatives involve taking immediate and direct action to solve problems and the less effective response alternatives involve not taking action (e.g., taking a "wait and see" approach) or taking actions that are not directly targeted at the problem at hand. (Nine items.)

The median and the range of the scale means and the median and the range of the scale standard deviations, for each exercise, indicated that (a) there is a reasonable amount of variation in the scale ratings, (b) none of the scale ratings show a floor effect, and (c) a reasonable number of the ratings do not show a ceiling effect.

Personal Counseling Exercise. Table 1.46 presents the pattern of matrices resulting from the factor analyses of the standardized and raw score Personal Counseling exercise ratings that specified two factors. The two-factor structure was preferred over the one- or three- (or more) factor structures based on the superior simple structure

Table 1.46
LVII Personal Counseling Exercise Scales and Factor Analysis Results[a]

Scale	Factor 1	Factor 2	h^2
Communication/Interpersonal Skill			
1. States the purpose of the counseling session clearly and concisely.[1]	.45	-.04	.18
2. Gives the subordinate positive feedback for his/her overall good past performance.[1]	.74	-.10	.48
3. Explains what the solider did wrong and why it was or can be a problem.[1]	.38	-.06	.12
7. Maintains eye contact during the interview.[2]	.30	.14	.16
8. Behaves in a manner that demonstrates support and concern for subordinate.[OMIT]	.52	.30	.54
9. Conducts the counseling session in a professional manner.[2]	.47	.12	.29
10. Maintains open communication.[2]	.13	.45	.27
Diagnosis/Prescription			
4. Asks open-ended. fact -finding questions that uncover important and relevant information.[1]	.01	.78	.61
5. Provides advice to the subordinate concerning actions that should be taken to solve problems.[1]	-.04	.87	.73
6. Sets a time or date to follow-up with the subordinate.[1]	.01	.52	.27
Omitted Item			
11. Does not interrupt the subordinate when he/she is talking.[2]	.08	.17	.05
Eigenvalue[b]	6.73	1.39	

Note. The underline indicates which composite the scale was assigned to for the construction of simulation exercise basic scores; h^2 = Communality.

[a] Maximum likelihood factor analysis with an oblique rotation. From analysis of standardized scale ratings.
[b] Eigenvalues of the first two unrotated factors.

[1] A similar (or the same) scale was assigned to the Personal Counseling - Content composite score in CVII.
[2] A similar (or the same) scale was assigned to the Personal Counseling - Process composite score in CVII.
[OMIT] A similar scale was not assigned to a composite score in the CVII analyses.

and interpretability of the rotated two-factor pattern matrix. Factor 1 was labeled "Communication/Interpersonal Skill," and Factor 2 was labeled "Diagnosis/Prescription."

As indicated by the notations in Table 1.46, the factor analysis results for LVII did not exhibit the same pattern as that obtained in CVII.

Disciplinary Counseling Exercise. Table 1.47 presents the pattern of matrices resulting from the factor analyses of the standardized and raw scale Disciplinary Counseling exercise ratings that specified three factors. The three-factor structure was preferred over the one-, two-, or four- (or more) factor structures based on the superior simple structure and interpretability of the rotated three-factor pattern matrix. Factor 1 was labeled "Structure." Factor 2 was labeled "Interpersonal Skill," and Factor 3 was labeled "Communication."

Training Exercise. Table 1.48 presents the pattern matrices resulting from the factor analyses of the standardized and raw scale Training exercise ratings that specified two factors, which were labeled "Structure" and "Motivation Maintenance."

Basic Scores. Scales were assigned to composite scores based on the patterns of their relative factor loadings in the factor structure for each exercise. This procedure resulted in empirically derived basic scores for each exercise that seemed to have considerable substantive meaning.

Across all exercises, each basic composite score was generated by (a) standardizing the ratings on each scale within each evaluator, (b) scaling each standardized rating by its raw score mean and standard deviation, and (c) calculating the mean of the transformed (i.e., standardized and scaled) ratings that were assigned to that particular basic criterion composite. The ratings were standardized within evaluator because (a) each evaluator rated examinees in only some MOS and (b) there was more variance in mean ratings across evaluators than there was in mean ratings across MOS. The standardized ratings were scaled with their original overall means and standard deviations so that each scale would retain its relative central tendency and variability.

Summary of Basic Criterion Scores

The analyses of the LVII performance measures resulted in an array of basic criterion scores which were available for the performance modeling analyses. These scores are summarized in Figure 1.12.

Table 1.47
LVII Disciplinary Counseling Exercise Scales and Factor Analysis Results[a]

Scale	Factor 1	Factor 2	Factor 3	h^2
Structure				
1. Remains focused on the immediate problems (i.e., the subordinate's absences and/or lying).[1]	.38	.12	-.08	.17
2. Determines an appropriate corrective action.[1]	.57	.04	-.02	.33
3. States the exact provisions of the punishment.[1]	.57	-.01	.07	.34
Interpersonal Skill				
6. Conducts the counseling session in a professional manner.[2]	.07	.72	-.02	.53
7. Defuses rather than escalates potential arguments.[2]	-.04	.67	-.02	.44
Communication				
4. Explains the ramifications of the soldier's actions.[OMIT]	.01	-.03	.82	.66
5. Allows the subordinate to present his/her view of the situation.[2]	.14	.08	.29	.14
Eigenvalue[b]	2.62	1.52	1.02	

Note. The underline indicates which composite the scale was assigned to for the construction of simulation exercise basic scores; h^2 = Communality.

[a] Maximum likelihood factor analysis with an oblique rotation. From analysis of standardized scale ratings.
[b] Eigenvalues of the first three unrotated factors.

[1] A similar (or the same) scale was assigned to the Disciplinary Counseling - Content score in CVII.
[2] A similar (or the same) scale was assigned to the Disciplinary Counseling - Interpersonal Skills score in CVII.
[OMIT] A similar item was not assigned to a composite score in the CVII analyses.

Table 1.48
LVII Training Exercise Scales and Factor Analysis Results[a]

Scale	Factor 1	Factor 2	h^2
Structure			
2. Organizes and presents the training steps in a logical sequence.	.64	-.03	.39
3. Demonstrates the task steps for the trainee.	.58	.07	.39
4. Identifies and corrects the trainee's errors.	.74	-.16	.41
5. Makes the trainee practice each movement required to perform the task.	.66	-.03	.41
6. Provides specific feedback to the trainee following good performance.	.70	.04	.53
Motivation Maintenance			
7. Provides positive feedback to the trainee following good performance.	-.01	.81	.65
8. Encourages the trainee when mistakes are made.	-.07	.80	.57
Omitted Items			
1. Presents an overview of what will be learned.	.18	.21	.13
9. Speaks in a clear, distinct, and understandable manner.	.28	.26	.25
Eigenvalues[b]	6.12	1.32	

Note. The underline indicates which composite the scale was assigned to for the construction of simulation exercise basic scores. In the CVII analyses scales similar (or identical) to those above were assigned to a single Training Exercise composite score. h^2 = Communality.

[a] Maximum likelihood factor analysis with an oblique rotation. From analysis of standardized scale ratings
[b] Eigenvalues of the first two unrotated factors.

Hands-On Performance Test

1. MOS-specific task performance score
2. General (common) task performance score

Job Knowledge Test

3. MOS-specific task knowledge score
4. General (common) task knowledge score

Army-Wide Rating Scales

5. Overall effectiveness rating
6. Leadership/supervision composite
7. Technical skill and effort composite
8. Personal discipline composite
9. Physical fitness/military bearing composite

MOS-Specific Rating Scales

10. Overall MOS composite

Combat Performance Prediction Scales

11. Overall Combat Prediction scale composite

Personnel File Form

12. Awards and certificates
13. Disciplinary actions (Articles 15 and Flag actions)
14. Physical readiness
15. Promotion rate

Situational Judgment Test

16. Total composite or, alternatively,
17. Discipline soldiers when necessary
18. Focus on the positive
19. Search for underlying causes
20. Work within chain of command
21. Show support/concern for subordinates
22. Take immediate/direct action

Supervisory Simulation Exercises

23. Personal counseling - Communication/Interpersonal skill
24. Personal counseling - Diagnosis/Prescription
25. Disciplinary counseling - Structure
26. Disciplinary counseling - Interpersonal skill
27. Disciplinary counseling - Communication
28. Training - Structure
29. Training - Motivation maintenance

Figure 1.12. Summary list of LVII basic criterion scores.

The LVII Data File: Extent of Missing Data

The LVII data were collected from 1,595 soldiers in the nine MOS designated as Batch A MOS in previous data collections. It was not always possible to collect complete information from each soldier for all instruments. For example, for the hands-on measures, some necessary pieces of equipment might have been unavailable for use, making it impossible to score some or all of the steps of a particular task test, or supervisors may have felt that they were not able to use a particular rating scale because of too few opportunities to observe that aspect of performance. For the Personnel File Form, soldiers may have left questions unanswered if they did not know, or chose not to provide, the requested information.

The number of soldiers who are missing all data on a particular instrument can be determined from Table 1.49. For example, only 341 of the 347 MOS 11B soldiers participated in hands-on testing while all 347 soldiers in the 11B sample participated in the job knowledge test administration.

Various methods were used for the criterion instruments to deal with partially missing data. For the Personnel File Form and Simulation Exercises, missing data were simply left as missing. For the other measures, various strategies were used to treat missing data.

The percentages of assigned values for missing data for each performance instrument are shown in Table 1.50. That is, these are the individuals in the sample who had some missing data but not enough to be dropped from the data set for a particular instrument. Instead, their scores were computed using the rules described in Campbell and Zook (1994b). Note that these percentages are generally very low: almost all are less than one percent except for the MOS Ratings Scales.

Development of the LVII Performance Model

The specific objective was to determine which model (i.e., a particular specification of the number of components and their substantive content), from among several proposed alternative models of the latent structure of basic criterion score intercorrelations, best fits the observed data. Analyses were guided by the same general framework that was used in modeling the covariation among performance measures for first-tour performance (J.P. Campbell et al., 1990).

One alternative was the model developed based on data from the Project A Concurrent Validation second-tour (CVII) sample. This model, referred to as the Training and Counseling model, is described in detail in Campbell and Oppler (1990).

Table 1.49
Number of LVII Soldiers With Complete or Partial Data by Criterion Instrument and MOS

MOS	N	Job Knowledge	Hands-On	Army-Wide Rating Scales	MOS Rating Scales	Combat Prediction	PFF	SJT	Simulation Exercises
11B	347	347	341	333	321	313	347	346	341
13B	180	179	174	167	164	160	179	178	174
19K	168	168	160	156	149	152	161	166	156
31C[a]	70	70	--	65	66	64	68	70	--
63B	194	192	187	182	191	176	193	193	188
71L	157	155	156	153	150	150	155	157	156
88M	89	89	88	86	87	85	88	89	88
91A/B	222	220	215	212	208	205	218	220	214
95B	168	168	168	167	162	163	168	168	167
Total	1,595	1,589	1,489	1,521	1,498	1,468	1,577	1,587	1,485

Note. PFF = Personnel File Form; SJT = Situational Judgment Test.

[a] Hands-On and Supervisory Simulation Exercises data were not collected for MOS 31C.

Table 1.50
Percent of LVII Assigned Values by Type of Instrument and MOS

MOS	Job Knowledge	Hands-On	Army-Wide Rating Scales	MOS Rating Scales	Combat Ratings	Personnel File Form	Situational Judgment Test	Supervisory Simulation Exercises
11B	.00	.88	.19	2.88	.00	.00	.14	.00
13B	.00	1.55	.55	2.00	.00	.00	.17	.00
19K	.00	.00	.03	.68	.00	.00	.06	.00
31C	.00	--[a]	.46	1.79	.00	.00	.15	--[a]
63B	.00	1.92	.54	.66	.00	.00	.12	.00
71L	.00	.92	.85	1.75	.00	.00	.11	.00
88M	.00	.91	.44	3.96	.00	.00	.23	.00
91A/B	.00	.92	.78	8.08	.00	.00	.09	.00
95B	.00	.85	.61	6.67	.00	.00	.09	.00
Total Sample	.00	.98	.47	3.33	.00	.00	.12	.00

[a] Hands-On and Supervisory Simulation Exercises data were not collected for MOS 31C.

Sample and Procedure

The sample used in the LVII modeling analyses included soldiers from eight of the nine Batch A MOS for which a full set of criterion measures had been developed (C.H. Campbell et al., 1990). Because complete data on the entire array of basic criterion scores were required and because soldiers from MOS 31C did not have hands-on performance scores, these soldiers were excluded from all of the present analyses.

As a result of these considerations, a total sample of 1,144 soldiers with complete data was available for the initial modeling analyses. The MOS breakdown is shown in Table 1.51.

Table 1.51
Number of LVII Soldiers With Complete Array of Basic Criterion Scores (Excluding Combat Performance Prediction Scales) by MOS

MOS		Number With Complete Data
11B	Infantryman	281[a]
13B	Cannon Crewmember	117
19K	M1 Armor Crewman	105
31C	Single Channel Radio Operator	0
63B	Light-Wheel Vehicle Mechanic	157
71L	Administrative Specialist	129
88M	Motor Transport Operator	69
91A/B	Medical Specialist	156
95B	Military Police	130
Total Sample		1,144

[a] These soldiers do not have general soldiering scores for the hands-on or job knowledge tests.

As a first step, several alternative models of second-tour soldier performance were hypothesized. The fit of these alternative models was then assessed using the LVII data and compared with the fit of the CVII Training and Counseling model. Second, because the Combat Performance Prediction Scales were not included in this initial modeling, key analyses were rerun with these scales included to confirm that the Combat scales fit the models as expected and to determine whether including them would affect the degree of fit. Once a best fitting model was identified, subsequent analyses were conducted to determine whether the model fit equally well across MOS and across demographic subgroups. Finally, based on the results of these analyses, a set of criterion construct scores to be used in the LVII validation analyses was specified.

To generate alternative hypotheses for the latent structure, definitions of the LVII basic criterion scores used in the modeling exercise were circulated to the project staff, and a variety of hypotheses concerning the nature of the underlying structure of second-tour soldier performance were obtained. These hypotheses were consolidated into one principal central alternative model, several variations on this model, and a series of more parsimonious models that involved collapsing two or more of the substantive factors.

The central alternative, the Consideration/Initiating Structure model presented in Table 1.52, differs from the CVII Training and Counseling model primarily in that it includes two leadership factors. Based on staff judgment, the leadership rating scales and each of the SJT and Supervisory Simulation scores were assigned to one of these two factors.

Because the within-MOS sample sizes in the LVII sample were relatively small (ranging from 69 to 281), initial tests of the models were conducted using the entire LVII sample. Criterion scores were first standardized within each MOS, then the inter-correlations among these standardized basic scores were computed across all MOS. The total sample matrix was used as input for the analyses.

The analysis plan was to first compare the fit of the Consideration/ Initiating Structure model with the variations of this model and with the Training and Counseling model to identify the alternatives that best fit the LVII covariance structure. The next set of analyses involved comparing a series of nested models to determine the extent to which the observed correlations could be accounted for by fewer underlying factors. LISREL 7 was used to estimate the parameters and evaluate the fit of each of the alternative models.

Results

The fit of the Training and Counseling model in the LVII sample was remarkably similar to the fit of this same model in the CVII sample, especially considering that the performance data were collected several years apart using somewhat different measures.

Tests of the Consideration/Initiating Structure model and the variations on this model resulted in a very poor fit to the data (e.g., RMSR values greater than .09) and the program encountered a variety of problems in estimating the parameters for these models.

To determine whether there were other reasonable alternative models of second-tour soldier performance, the LVII total sample was randomly divided into two subsamples: 60 percent for model development and 40 percent for cross-validation/ confirmation.

Table 1.52
Consideration/Initiating Structure Model

Latent Variable	Scores Loading on Latent Variables
Core Technical Proficiency (CT)	MOS-Specific Hands-On MOS-Specific Job Knowledge
General Soldiering Proficiency (GP)	General Hands-On General Job Knowledge
Achievement and Effort (AE)	Awards and Certificates Promotion Rate Army-Wide Ratings: Technical Skill/Effort Composite Overall Effectiveness Rating MOS Ratings: Overall Composite Combat Prediction: Overall Composite
Personal Discipline (PD)	Disciplinary Actions (reversed) Army-Wide Ratings: Personal Discipline Composite
Physical Fitness/Military Bearing (PF)	Physical Readiness Score Army-Wide Ratings: Physical Fitness/ Bearing Composite
Leadership: Initiating Structure (IS)	Army-Wide Ratings: Leading/Supervising Composite SE - Disciplinary Structure SE - Counseling Diagnosis/Prescription SE - Training Structure SJT - Disciplining SJT - Immediate/Direct Action SJT - Chain of Command
Leadership: Consideration (LC)	SE - Disciplinary Communication SE - Disciplinary Interpersonal Skill SE - Counseling Communication/Interpersonal Skills SE - Training Motivation Maintenance SJT - Support SJT - Search for Reasons SJT - Focus on the Positive
Written Methods	Technical Knowledge Basic Job Knowledge All Six SJT Scores
Ratings Methods	All Four Army-Wide Ratings Composites Overall Effectiveness Rating MOS Ratings: Overall Composite Combat Prediction: Overall Composite
Disciplinary Simulation Exercise Methods	All Three SE - Disciplinary Counseling Scores
Counseling Simulation Exercise Methods	Both SE - Personal Counseling Scores
Training Simulation Exercise Methods	Both SE - Training Scores

The matrix of intercorrelations among the basic criterion scores for the developmental subsample was examined by project staff and several alternative models were suggested. A number of alternatives tried different arrangements of the supervisory simulation, SJT, and rating scale basic scores, while still preserving two leadership factors. None of the these alternatives resulted in a good fit. However, a model that collapsed the Consideration and Initiating Structure factors into a single Leadership factor, included a single Simulation Exercise method factor, and moved the promotion rate variable to the new Leadership factor did result in a considerably better fit to the data in both the developmental and holdout samples.

The "Leadership Factor" model that was developed based on these exploratory analyses is shown in Table 1.53. The fit of the new Leadership Factor model to the LVII data is, for all practical purposes, identical to the fit of the Training and Counseling model to these same data. The 90 percent confidence intervals for the RMSEAs overlap almost completely.

Because these models have an equally good fit to the data and because the Leadership Factor model does not confound method variance with substantive variance, the Leadership Factor model was chosen as the best representation of the latent structure of second-tour performance.

The Leadership Factor model was tested again with the Combat Performance Prediction Scales included. For one comparison, the Combat Prediction Score was constrained to load only on the Leadership factor and the Rating Method factor. For the second, the Combat Prediction score was constrained to load on the Achievement and Effort and the Rating Method factors only.

The second assignment (i.e., the Combat Prediction Score assigned to the Achievement and Effort factor) produced a much better fit.

Nested Models. Next, the Leadership Factor model was used as the starting point to develop a series of more parsimonious nested models, similar to those tested in the LVI sample by Oppler, Childs, and Peterson (1994). The first was identical to the full Leadership Factor model except that the Achievement and Effort factor was collapsed with the Leadership factor.

Similarly, the second nested model was identical to the model just described except that, in addition, the Core Technical and General Soldiering Proficiency factors were replaced with a single "can do" factor. Third, the Personal Discipline factor and the new Achievement/Leadership factor were also collapsed. The fourth model involved adding the variables from the Physical Fitness factor to this Achievement/Leadership/ Personal Discipline factor, resulting in a single "will do" factor. The final model collapsed all of the substantive factors into a single overall performance factor.

Table 1.53
Leadership Factor Model

Latent Variable	Scores Loading on Latent Variables
Core Technical Proficiency (CT)	MOS-Specific Hands-On MOS-Specific Job Knowledge
General Soldiering Proficiency (GP)	General Hands-On General Job Knowledge
Achievement and Effort (AE)	Awards and Certificates Army-Wide Ratings: Technical Skill/Effort Composite Overall Effectiveness Rating MOS Ratings: Overall Composite Combat Prediction: Overall Composite
Personal Discipline (PD)	Disciplinary Actions (reversed) Army-Wide Ratings: Personal Discipline Composite
Physical Fitness/Military Bearing (PF)	Physical Readiness Score Army-Wide Ratings: Physical Fitness/Bearing Composite
Leadership (LD)	Promotion Rate Army-Wide Ratings: Leading/Supervising Composite SE - Disciplinary Structure SE - Disciplinary Communication SE - Disciplinary Interpersonal Skill SE - Counseling Diagnosis/Prescription SE - Counseling Communication/Interpersonal Skills SE - Training Structure SE - Training Motivation Maintenance SJT - Total Score
Written Method	Job-Specific Knowledge General Job Knowledge SJT - Total Score
Ratings Method	Four Army-Wide Ratings Composites Overall Effectiveness Rating MOS Ratings: Total Composite Combat Prediction: Overall Composite
Simulation Exercise Method	All Seven Simulation Exercise Scores

Because these more parsimonious models are nested within each other, the significance of the loss of fit could be tested by comparing the chi-square values for the various models.

In the first nested model, which involved collapsing the Leadership factor with the Achievement and Effort factor, the resulting decrement in fit was very small. Similarly, collapsing the two "can do" factors resulted in a very small reduction in model fit. Based on these results, a model with only four substantive factors (and three method factors) can account for the data almost as well as the full Leadership Factor model.

Collapsing additional factors beyond this level resulted in larger decrements in model fit.

Retrospective Re-Analysis of the CVII Data. One final approach to confirming the Leadership Factor model was to assess the fit of this new model to the CVII data. These results were virtually identical to those obtained in the LVII data.

LVII Criterion Construct Scores

The basic criterion construct scores for use in validation analyses are based on the full Leadership Factor model, with six substantive factors. The nested model with four factors (with a single Achievement/Leadership factor and a single "can do" factor combining Core Technical and General Soldiering Proficiency) fits the data almost as well and has the advantage of greater parsimony. However, it is still plausible that all six performance factors have somewhat different antecedents and could be related to different predictor constructs. Therefore, for the initial validity analyses the model that incorporates the six criterion construct scores was retained.

Results of the nested analyses were used to form more parsimonious sets of criterion construct scores as well. This was done by first standardizing each of the six construct scores described above (based on the full Leadership model). These were then added together in the order shown in Figure 1.13 to form sets of five, four, three, two and finally one criterion composite construct score.

Concluding Comments

In general, results of the LVII modeling analyses showed that both the Training and Counseling model and the Leadership Factor model fit the LVII data quite well. Further, retrospective reanalysis of the CVII data showed that these two models had a similarly good fit in the CVII sample.

The new six-factor Leadership Factor model of second-tour performance is also consistent with the CVI/LVI model of first-tour soldier performance. In addition to including performance factors that are parallel to those identified for first-tour soldiers, the LVII second-tour model includes a Leadership factor that contains all measures that were in fact targeted at the leadership/supervision aspects of the job. This is consistent with the results of the second-tour job analyses which indicated that second-tour soldiers perform many of the same tasks as the first-tour soldiers in addition to their supervisory responsibilities. In sum, the Leadership Factor model provides the starting point for the LVII validity analyses and further enhances our understanding of second-tour soldier performance.

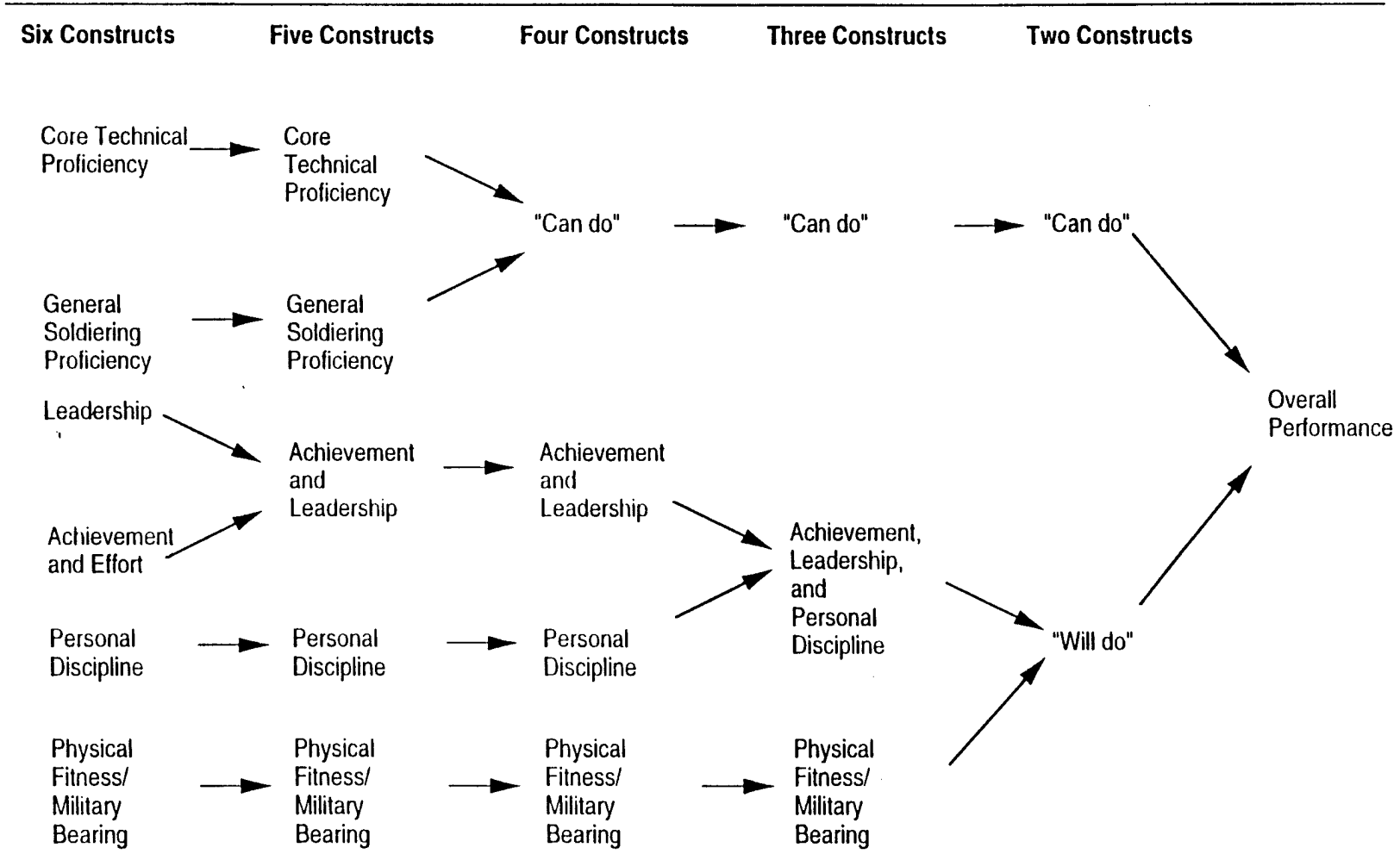


Figure 1.13. Final LVII Criterion and Alternate Criterion Constructs based on more parsimonious models.

SUMMARY OF PROJECT EFFORTS FOR YEAR FOUR

The overall objectives for the fourth annual report (Campbell & Zook, 1994c) were to report the results of a number of validation analyses related to the broad spectrum of criterion measures that have been used in Project A and Career Force. These range from an analysis to support the evaluation of the Enhanced Computer Administered Test (ECAT) battery project to an analysis of the Army Job Satisfaction Questionnaire. The analyses reported for FY93 included the following:

(1) Identifying "optimal" subsets of Project A Experimental Battery tests corresponding to the ECAT battery that maximize (a) absolute validity and (b) discriminant validity, and minimize (c) subgroup differences.

(2) Completing basic validation of the Experimental Predictor Battery for the LVII sample.

(3) Validating training performance (EOT) and current job performance as predictors of future performance.

(4) Predicting first-term attrition.

(5) Validating job satisfaction as a predictor of turnover and attrition.

Identification of Optimal Predictor Batteries Using a Subset of the Project A/Career Force Experimental Predictor Battery

These analyses were completed to assist a subcommittee of the Manpower Accession Policy Working Group in its deliberations about possible revisions to the ASVAB. The committee had previously identified a battery of experimental tests, designated as the Joint Services Enhanced Computer Administered Test battery (ECAT), that would constitute the array of new predictors to be considered as potential additions to ASVAB. Our analyses were limited to the experimental predictors in the ECAT battery that were also in the Project A/Career Force Experimental Battery.

The problem was to identify which subset of 17 tests (the ten ASVAB subtests plus seven tests from the Project A/Career Force predictor battery) should be included in an operational test battery when the objective is to maximize certain specific indices of battery quality. A number of indices can be used to evaluate the performance of a test battery. The indices considered included (a) the validity of the battery, (b) the battery's capacity to maximize classification efficiency, (c) the subgroup differences that could result from use of the test battery relative to multiple subgroups (i.e., Blacks, Hispanics, and Females), (d) the performance of the battery relative to these first three parameters across multiple MOS, and (e) the amount of time required to administer the battery. Again, the objective was to analyze the part of the Project A battery that corresponded most closely to the ECAT battery so as to aid the evaluation of ECAT.

Indices for Evaluating the Value of a Test Battery

Absolute Validity. Absolute validity was defined as the multiple correlation associated with the regression of a criterion on a set of predictors and was used as the index of selection validity.

Classification Efficiency: Differential Validity. Classification efficiency can be viewed as the increase in the average level of job performance that would result from replacing selection with some method of classification (differential job assignments). A necessary condition for classification efficiency is differential validity that results in intra-individual variation in predicted performance scores across jobs.

Our index of differential validity is referred to as discriminant validity and is defined as the difference between mean absolute validity and mean generalizability validity. Mean absolute validity is the average multiple correlation resulting from the regression of the criterion on a set of predictors in each of multiple jobs. That is, the predicted performance scores for individuals in each job would be computed using regression weights that are specific to that job. The predicted performance scores from the equation using the least squares weights computed within each job can also be correlated with performance scores in each of the other jobs. The mean of these correlations is the mean generalizability validity. If there are (m) jobs, this mean is based on m(m-1) validity estimates. Thus, our index of differential validity is:

Discriminant Validity (DV) = mean absolute validity - mean generalizability validity

Classification Efficiency: Brogden Index of Classification Efficiency. Brogden's mean predicted performance (MPP) (Brogden, 1959) is obtained by calculating

$$MPP = R(1-r)^{1.2} f(A)$$

where R is the mean of the absolute validities across jobs, r is the mean of the correlations among the predicted scores for each job, and f(A) is a multiplier that increases with the selection ratio and with the number of jobs for which classification occurs (see Peterson, Oppler, Sager, & Rosse, 1993). The advantages of the equation are that it (a) simultaneously considers indicators of absolute validity and differential validity, and (b) takes into account two important contextual selection variables - the number of jobs and the selection ratio. However, it assumes that there is (a) no variation in R across jobs (the least squares regression equations yield the same absolute validities across all jobs), and (b) no variation in r across jobs (the same degree of similarity exists among all pairs of equations).

For the purposes of this report, we used the Brogden index of classification efficiency

$$R(1-r)^{1.2}$$

which removes the term (f(A)) that is invariant for the Project A/Career Force data (because the number of jobs and the selection ratio never vary).

Subgroup Differences. Subgroup differences for White versus Black (W-B), White versus Hispanic (W-H), and Male versus Female (M-F) were defined as the mean of the predicted scores for individuals in the referent subgroup (e.g., Whites) minus the mean of the predicted scores for individuals in the non-referent subgroup (e.g., Blacks) divided by the standard deviation of the referent subgroup's predicted scores.

Summary of Issues. The overall goal of the analyses was to identify an "optimal" combination of tests for predicting three criteria (technical/school knowledge, on-the-job core technical proficiency, and on-the-job hands-on test performance) at various time lengths for test battery administration. "Optimal" was defined as the appropriate trade-off among three battery indices: mean absolute validity, mean discriminant validity, and mean adverse impact.

A New Approach

The general analysis procedure was to calculate all the indices of test battery performance for every possible combination of subtests that fell within a given test administration time interval. The combinations of subtests could then be rank-ordered according to each index of test battery performance. For instance, the top 20 test batteries ranked on the basis of maximum absolute validity can be compared to the top 20 test batteries ranked on the basis of minimal M-F adverse impact. This method provides the information necessary to evaluate trade-offs among absolute validity discriminant validity, and adverse impact.

Sample

The sample data were from nine of the Longitudinal Validation (LV) first-tour Batch A MOS (19E was excluded). Recall that these data were collected using a longitudinal validation design with approximately 2 years between the time when predictor tests were admininstered and the time when performance criteria data were collected. The end-of-course training performance technical knowledge test was administered 3-6 months after the individual started basic training.

Criteria

Three criteria were used: Technical/School Knowledge test score (an end-of-training criterion), Core Technical Proficiency (CTP), and Hands-On test performance (HO).

Predictors

The analyses were designed to identify optimal subsets of 15 candidate predictors to be administered along with the Arithmetic Reasoning and Word Knowledge subtests of the ASVAB, which were designated as the base condition to be included in each of the potential test batteries. The 15 predictors considered included all the 8 remaining ASVAB subtests and 7 Experimental Battery subtests under serious consideration for supplementing the ASVAB. Specifically, the tests were:

- Eight ASVAB subtests other than Arithmetic Reasoning (ASAR) and Word Knowledge (ASWK):
  - Auto and Shop Information (ASAS)
  - Coding Speed (ASCS)
  - Electronics Information (ASEI)
  - General Science (ASGS)
  - Mechanical Comprehension (ASMC)
  - Mathematical Knowledge (ASMK)
  - Numerical Operations (ASNO)
  - Paragraph Comprehension (ASPV)

- Three Project A/Career Force paper-and-pencil spatial tests:
  - Assembling Objects (SPAO)
  - Orientation (SPOR)
  - Reasoning (SPRS)

- Three Project A/Career Force computerized tests:
  - Target Identification (Decision Time) (CMTI)
  - One-Hand Tracking (CM1T)
  - Two-Hand Tracking (CM2T)

- One Project A/Career Force computerized test composite:
  - Short-Term Memory (CMST)

Test Battery Time Intervals

Time intervals were defined by total minutes allowed for "start-up" time, instruction time, and test-taking time. The three time intervals were: 74-104 minutes, 134-164 minutes, and 194-224 minutes.

Results

The results are summarized in Table 1.54. The separate sections of the table present the general summary statistics and information about the performance of the prediction equations across all possible combinations of predictors (that each include Arithmetic Reasoning and Word Knowledge) for each time interval and for each of the criterion measures.

Results listing the top 20 potential test batteries identified for each of the six test battery performance indices are shown in Appendix A of the FY93 annual report. These results are reported separately for each time interval within each criterion/time interval combination: each table has six lists of equations rank-ordered by a performance index.

Table 1.54
Summary of Results for Predicting Each Criterion at Each Time Interval

General Summary Statistics	74-104 Minute Batteries		134-164 Minute Batteries		194-224 Minute Batteries	
Total Number of Combinations	210		10,152		6,509	
Mean Testing Time in Minutes (SD)	97.5	(5.8)	151.0	(8.7)	205.7	(8.4)
Mean Number of Tests (SD)	4.6	(.7)	8.2	(1.0)	11.6	(1.0)
Technical/School Knowledge						
Mean Validity Coefficient (SD)	.739	(.009)	.761	(.007)	.771	(.005)
Mean Discriminant Validity Index (SD)	.013	(.008)	.025	(.007)	.031	(.004)
Mean Brogden Index (SD)	.102	(.026)	.145	(.018)	.164	(.010)
Mean W-B Difference (SD)	1.426	(.054)	1.457	(.045)	1.465	(.028)
Mean W-H Difference (SD)	1.053	(.042)	1.080	(.034)	1.088	(.023)
Mean M-F Difference (SD)	.262	(.089)	.315	(.069)	.321	(.044)
On-the-Job Core Technical Proficiency						
Mean Validity Coefficient (SD)	.593	(.010)	.615	(.009)	.623	(.006)
Mean Discriminant Validity Index (SD)	.012	(.007)	.022	(.007)	.027	(.005)
Mean Brogden Index (SD)	.123	(.018)	.163	(.015)	.186	(.010)
Mean W-B Difference (SD)	1.339	(.063)	1.378	(.047)	1.392	(.029)
Mean W-H Difference (SD)	.993	(.051)	1.019	(.043)	1.023	(.030)
Mean M-F Difference (SD)	.342	(.114)	.407	(.092)	.411	(.057)
On-the-Job Hands-On Test Performance						
Mean Validity Coefficient (SD)	.455	(.017)	.490	(.012)	.500	(.007)
Mean Discriminant Validity Index (SD)	.017	(.007)	.027	(.007)	.030	(.006)
Mean Brogden Index (SD)	.132	(.014)	.167	(.012)	.187	(.010)
Mean W-B Difference (SD)	1.182	(.103)	1.197	(.065)	1.193	(.035)
Mean W-H Difference (SD)	.888	(.083)	.899	(.055)	.895	(.034)
Mean M-F Difference (SD)	.526	(.175)	.636	(.125)	.656	(.074)

W-B = White-Black; W-H = White-Hispanic; M-F = Male-Female.

"Optimal" Test Batteries

Many different rules could be used to create an "optimal" test battery that somehow considers all the discussed indices of battery performance. As a demonstration, two such rules were implemented using the "top twenty" potential test batteries.

The first "rule" implemented was "Maximize Validity/Minimize Average Subgroup Differences." According to this rule, the optimal combination of tests for each criterion and time interval was that battery of tests in the top 20 mean absolute validities list that also had the lowest average across the three types of subgroup differences. The optimal equations according to this rule for each criterion and time interval are presented in Table 1.55.

The second "rule" implemented was "Maximize Discriminant Validity/Minimize Average Subgroup Differences." According to this rule, the optimal subtest combination for each criterion and time interval was that battery of tests in the top 20 mean discriminant validities list that also had the lowest average across the three types of subgroup differences. The optimal equations according to this rule for each criterion and time interval are presented in Table 1.56.

The results in Table 1.56 are in most respects very similar to the results in Table 1.55. A major difference is that the computer tests occur more often in these "optimal" equations, especially for the longer intervals and for those equations predicting On-the-Job Hands-On test performance. This result suggests that the computer tests contribute more to maximizing discriminant validity. The problem is, however, that they also appear to contribute to M-F differences. There was a general tendency for the estimates of M-F differences in Table 1.56 to be of greater magnitude than the corresponding estimates in Table 1.55.

Summary

This method of analysis does not identify the combination of subtests that will simultaneously optimize all the test battery performance parameters. The results of the analyses convincingly make the point that a single, optimal test battery does not exist. Examination of the test batteries ordered by each of the indices makes some trade-offs apparent.

The more dramatic trade-offs are (a) the maximization of absolute validity versus the minimization of all three types of subgroup differences, (b) the maximization of classification efficiency versus the minimization of all three types of subgroup differences, (c) the minimization of W-B differences versus the minimization of M-F differences, and (d) the minimization of W-H differences versus the minimization of M-F differences.

Predictably, subtests in the ASVAB battery show a substantial tendency to contribute to absolute validity and a moderate tendency to contribute to

Table 1.55
"Optimal" Test Batteries for Each Criterion and Time Interval According to the "Maximize Validity/Minimize Average Subgroup Difference" Rule

Criterion (Time Interval)	V	DV	BI	W-B	W-H	M-F	T (min.)	Test Battery[a]
TSK (74-104 mins.)	.759	.029	.153	1.448	1.037	0.356	103	ASAR,ASWK,ASAS,ASCS,ASNO
TSK (134-164 mins.)	.774	.033	.167	1.453	1.057	0.346	161	ASAR,ASWK,ASAS,ASEI,ASMK,SPAO,CMST,CMTI
TSK (194-224 mins.)	.779	.036	.177	1.447	1.068	0.348	218	ASAR,ASWK,ASAS,ASEI,ASGS,ASMK,ASPC,SPAO, SPOR,CMST,CMTI, CM1T, CM2T
CTP (74-104 mins.)	.610	.006	.101	1.410	1.106	0.352	100	ASAR,ASWK,ASEI,SPRS
CTP (134-164 mins.)	.633	.029	.171	1.413	0.993	0.417	153	ASAR,ASWK,ASAS,ASCS,ASMK,SPAO,CMST
CTP (194-224 mins.)	.633	.034	.191	1.397	0.994	0.439	203	ASAR,ASWK,ASAS,ASCS,ASEI,ASMC,ASMK,ASPC, SPAO,CMST
HOT (74-104 mins.)	.486	.027	.149	1.164	0.857	0.725	101	ASAR,ASWK,ASAS,ASNO,CMST
HOT (134-164 mins.)	.513	.034	.177	1.210	0.873	0.682	164	ASAR,ASWK,ASAS,ASGS,ASMK,SPAO,CMST,CMTI
HOT (194-224 mins.)	.514	.034	.189	1.237	0.933	0.597	200	ASAR,ASWK,ASAS,ASGS,ASMC,ASMK,SPAO,SPRS, CMST,CMTI

TSK = Technical/School Knowledge; CTP = On-the-Job Core Technical Proficiency; HOT = On-the-Job Hands-On Test Performance; V = Validity; DV = Discriminant Validity; BI = Brogden Index of Classification Efficiency; W-B = White - Black Difference; W-H = White - Hispanic Difference; M-F = Male - Female Difference; T = Testing Time for Battery.

[a] See list of tests in the "Predictors" subsection.

Table 1.56
"Optimal" Test Batteries for Each Criterion and Time Interval According to the "Maximize Discriminant Validity/Minimize Average Subgroup Difference" Rule

Criterion (Time Interval)	V	DV	BI	W-B	W-H	M-F	T (min.)	Test Battery[a]
TSK (74-104 mins.)	.759	.029	.153	1.448	1.037	0.356	103	ASAR,ASWK,ASAS,ASCS,ASNO
TSK (134-164 mins.)	.770	.036	.173	1.433	1.071	0.393	157	ASAR,ASWK,ASAS,ASMK,ASNO,ASPC,CMST, CMTI,CM1T
TSK (194-224 mins.)	.778	.036	.177	1.449	1.069	0.348	220	ASAR,ASWK,ASAS,ASEI,ASGS,ASMC,ASMK, ASPC,SPAO,CMST,CMTI,CM1T
CTP (74-104 mins.)	.615	.029	.158	1.386	0.990	0.478	103	ASAR,ASWK,ASAS,ASCS,ASNO
CTP (134-164 mins.)	.621	.039	.200	1.360	0.980	0.519	163	ASAR,ASWK,ASAS,ASCS,ASMC,ASNO,ASPC, CMST,CM1T,CM2T
CTP (194-224 mins.)	.622	.041	.211	1.357	0.990	0.502	203	ASAR,ASWK,ASAS,ASCS,ASMC,ASMK,ASPC, SPOR,CMST,CMTI,CM1T,CM2T
HOT (74-104 mins.)	.461	.029	.161	1.028	0.785	0.596	100	ASAR,ASWK,ASNO,CMST,CMTI,CM2T
HOT (134-164 mins.)	.494	.044	.201	1.139	0.856	0.794	161	ASAR,ASWK,ASAS,ASGS,ASNO,ASPC,CMST, CMTI,CM1T,CM2T
HOT (194-224 mins.)	.509	.043	.210	1.183	0.894	0.668	214	ASAR,ASWK,ASAS,ASGS,ASMK,ASNO,ASPC, SPAO,SPRS,CMST,CMTI,CM1T,CM2T

TSK = Technical/School Knowledge; CTP = On-the-Job Core Technical Proficiency; HOT = On-the-Job Hands-On Test Performance; V = Validity; DV = Discriminant Validity; BI = Brogden Index of Classification Efficiency; W-B = White - Black Difference; W-H = White - Hispanic Difference; M-F = Male - Female Difference; T = Testing Time for Battery.

[a] See list of tests in the "Predictors" subsection.

potential classification efficiency. Compared to the ASVAB subtests, the Project A/Career Force subtests show a relatively more substantial tendency to contribute to classification efficiency and a relatively moderate tendency to contribute to absolute validity. The Project A/Career Force subtests do not seem to show a dramatically greater tendency, compared to the ASVAB subtests, to contribute to minimizing the three types of subgroup differences.

Value judgments will be faced in any situation where a battery of subtests is formed for operational use. We described some of the indices to consider in these situations. If judgments regarding the relative priority of optimizing each of the test battery parameters are not made explicitly, they will have been made implicitly by default.

Basic Validation Results for the LVII Sample

These analyses dealt with the evaluation of the validity of the ASVAB and the Project A Longitudinal Validation (LV) Experimental Battery for predicting second-tour performance in the Army (Oppler, Peterson, & Rose, in press). The results are based on the second-tour performance data collected from the Project A/Career Force longitudinal sample (LVII).

The objectives of the analyses were to:

(1) Compute the basic validities for ASVAB and Experimental Battery predictors against the second-tour performance factors and selected individual performance measures.

(2) Compare the validities of four alternative sets of ASVAB scores (nine ASVAB subtests vs. four ASVAB factors vs. AFQT vs. MOS-appropriate Aptitude Area composites).

(3) Compare the validities of three alternative sets of ABLE scores.

(4) Assess the incremental validities for the Experimental Battery predictors over the four ASVAB factor composites.

(5) Compare the incremental validities of three alternative sets of ABLE scores.

(6) Compare the validities and incremental validities of the Experimental Battery predictors reported for LVI with the validities and incremental validities reported for LVII.

Sample

The results reported here are based on a different strategy for the LVII sample than that used in LVI. Using the soldiers in each MOS in the LVII sample that were able to meet the setwise deletion requirements used in the LVI analyses (i.e., a separate

validation sample identified for each set of predictors) resulted in three MOS (19K, 71L, and 88M) with sample sizes of consistently fewer than 100 members. Consequently, a third sample selection strategy, termed "predictor/criterion setwise deletion," was developed.

Specifically, to be included in the validation sample for a given predictor/criterion set pair, soldiers were required to have complete data for the ASVAB, the predictor composites in the predictor set being examined, and only the specific criterion score being predicted. Table 1.57 reports the number of soldiers in each MOS meeting the predictor/criterion setwise deletion sample requirement for the Core Technical Proficiency criterion. All analyses in this subsection were conducted with samples selected with the predictor/criterion setwise deletion strategy.

This approach provided samples of reasonable size for MOS 19K and 71L, but MOS 88M was still considerably smaller. Therefore, MOS 88M (along with 31C) was not included in this analyses.

Table 1.57
Soldiers in LVII Sample Meeting Predictor/Criterion Setwise Deletion Data Requirements for Validation of ASVAB Operational Scores and Spatial, Computer, JOB, ABLE, and AVOICE Experimental Battery Predictor Composites Against Core Technical Proficiency by MOS

	Predictor Sets					
MOS	ASVAB	Spatial	Computer	JOB	ABLE	AVOICE
11B	333	322	112	301	297	309
13B	170	165	152	159	148	156
19K	156	130	130	122	123	129
63B	169	147	139	136	132	140
71L	147	115	104	105	109	102
88M[a]	84	56	54	51	52	53
91A	205	191	174	185	165	183
95B	160	149	140	142	133	145
Total	1,424	1,275	1,005	1,201	1,159	1,217

[a] MOS 88M was not included in LVII validity analyses.

Predictors and Criteria

The predictor scores were derived from the ASVAB and the Project A LV Experimental Battery. For the ASVAB, four types of scores were examined: the nine

ASVAB subtests, the four ASVAB factor composite scores, the AFQT, and the MOS-appropriate Aptitude Area composite scores.

Three different sets of ABLE scores were used. The first set, labeled the ABLE Composites, were derived along with the other LV predictor composites. The other two sets, labeled ABLE-168 Composites and ABLE-114 Composites, were based on results of more recent factor analyses of the ABLE items. ABLE-168 was scored using 168 of the ABLE items, and ABLE-114 was scored using only 114 items. The development of these scores is described by White (1994).

The model of second-tour performance that emerged from the LVII modeling analysis (labeled the Leadership Factor model) specified six substantive performance factors and three method factors ("written verbal", "ratings", and "simulation exercise"). The six substantive factors and four additional performance measures were used as criteria in the validation analysis. Two of these additional criteria are the total scores from the hands-on and job knowledge tests. The other two are variations of the Leadership factor. The first variation (LDR2) does not include the Situational Judgment Test (SJT) total score that was included in the original Leadership factor, and the second variation (LDR3) does not include either the SJT or the scores from the Supervisory Simulation exercises.

Results

Comparison of Alternative ASVAB Scores. The average multiple correlations (corrected and uncorrected for range restrictions) for the four different sets of ASVAB scores indicated that all four sets had virtually identical multiple Rs with all of the criterion measures. The four ASVAB factors tended to have very slightly higher validities than the other three sets of scores, whereas the AFQT tended to have slightly lower validities.

Comparison of Alternative ABLE Scores. The patterns and levels of multiple correlations were generally very similar across the three sets of ABLE items. However, the ABLE composites were somewhat better predictors of the Leadership factor ($R = .34$) than were ABLE-168 and ABLE-114 sets, and ABLE-168 was the best predictor of Core Technical Proficiency ($R = .30$).

Multiple Correlations by Predictor Type. Multiple correlations for the four ASVAB factors, the single spatial composite, the eight computer-based predictor scores, the three JOB composite scores, the seven ABLE composite scores, and the eight AVOICE composite scores are reported in Table 1.58. Using the predictor/criterion deletion sample, these results were computed separately by MOS and then averaged.

The results in Table 1.58 indicate that the four ASVAB factors were the best set of predictors for all of the criterion performance factors (CTP, GSP, AE, PD, PFB, and LDR), the two additional leadership factors (LDR2 and LDR3), and the Hands-On and Job Knowledge total scores. The highest multiple correlation was between the ASVAB

Table 1.58
Mean of Multiple Correlations Computed Within Job for LVII Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE

Criterion[a]	No. of MOS[b]	ASVAB Factors [4]	Spatial [1]	Computer [8]	JOB [3]	ABLE Comp. [7]	AVOICE [8]
CTP	7	64 (10)	57 (11)	53 (11)	33 (17)	24 (19)	41 (12)
GSP	6	63 (06)	58 (05)	48 (10)	28 (19)	19 (17)	29 (24)
AE	7	29 (14)	27 (13)	09 (11)	07 (12)	13 (17)	09 (15)
PD	7	15 (12)	15 (10)	12 (12)	03 (05)	06 (10)	06 (10)
PFB	7	16 (11)	13 (06)	03 (06)	07 (08)	17 (15)	09 (13)
LDR	7	63 (14)	55 (08)	49 (13)	34 (21)	34 (20)	35 (24)
LDR2	7	51 (16)	46 (12)	35 (19)	26 (21)	25 (22)	25 (24)
LDR3	7	47 (13)	39 (12)	31 (15)	19 (18)	23 (17)	20 (21)
HO-Total	7	46 (13)	41 (14)	33 (21)	24 (11)	12 (15)	21 (18)
JK-Total	7	74 (05)	67 (03)	58 (06)	37 (16)	29 (17)	44 (14)

Note: Predictor/criterion setwise deletion sample. Adjusted for shrinkage (Rozeboom formula 8) and corrected for range restriction. Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Decimals omitted.

[a] CTP = Core Technical Proficiency; GSP = General Soldiering Proficiency;
AE = Achievement and Effort; PD = Personal Discipline;
PFB = Physical Fitness/Military Bearing; LDR = Leadership;
LDR2 = Leadership minus Situational Judgment Test;
LDR3 = Leadership minus Situational Judgment Test and Supervisory Simulation Exercises;
HO = Hands-On; JK = Job Knowledge

[b] Number of MOS for which validities were computed.

factors and the Job Knowledge total score (R = .74), and the lowest correlations were with the Personal Discipline and Physical Fitness scores, (.15 and .6, respectively).

Except for the prediction of PFB with the ABLE composites, the spatial composite was the next best predictor, with patterns highly similar to the ASVAB pattern. The other predictor sets tended to have different patterns of correlations for the different criterion performance factors.

With regard to the ABLE, in general, the highest correlations are with the Leadership factor and the Core Technical factor. Comparatively, the correlations of the ABLE with Achievement and Effort and with Personal Discipline are lower. In large part this reflects the emergence of a separate leadership factor and the fact that the promotion rate index produced a better fit for the LVII model when it was scored as a Leadership component than as a component of the Personal Discipline factor. As in CVII, a faster promotion rate for LVII personnel is more a function of good things that happen rather than an absence of negative events that act to slow an individual's progression, as it was in CVI and LVI.

Incremental Validities for the Experimental Battery Over ASVAB. Incremental validity results for the Experimental Battery predictors over the ASVAB factor composites are reported in Table 1.59 and 1.60. Table 1.59 reports the multiple correlations for the four ASVAB factor composites alone (as computed separately in each of the predictor/criterion setwise deletion samples), whereas Table 1.60 reports the multiple correlations for the four ASVAB factors along with each set of predictors in the Experimental Battery. Numbers underlined in Table 1.60 indicate multiple correlations that are higher than those based on ASVAB alone (Table 1.59).

Table 1.59
Mean of Multiple Correlations Computed Within Job for ASVAB Factors Within Each LVII Predictor/Criterion Setwise Deletion Sample

Criterion	No. of MOS[a]	ASVAB Factors (Spatial Sample) [4]	ASVAB Factors (Computer Sample) [4]	ASVAB Factors (JOB Sample) [4]	ASVAB Factors (ABLE Comp. Sample) [4]	ASVAB Factors (AVOICE Sample) [4]
CTP	7	65 (10)	64 (07)	66 (11)	67 (09)	66 (12)
GSP	6	61 (08)	62 (09)	60 (09)	61 (11)	61 (08)
AE	7	31 (15)	24 (13)	32 (15)	30 (17)	30 (16)
PD	7	17 (12)	21 (17)	17 (13)	15 (11)	16 (12)
PFB	7	15 (10)	17 (14)	16 (10)	13 (13)	18 (11)
LDR	7	63 (13)	62 (16)	63 (14)	64 (12)	62 (13)
LDR2	7	52 (17)	50 (20)	50 (19)	51 (20)	49 (19)
LDR3	7	46 (19)	45 (23)	44 (22)	45 (23)	43 (23)
HO-Total	7	46 (14)	45 (17)	44 (17)	46 (18)	47 (16)
JK-Total	7	74 (05)	74 (06)	74 (06)	75 (05)	74 (06)

Note: Adjusted for shrinkage (Rozeboom formula 8) and corrected for range restriction. Numbers in parentheses are standard deviations. Numbers in brackets are the numbers of predictor scores entering prediction equations in each separate predictor/criterion sample. Decimals omitted.

[a] Number of MOS for which validities were computed.

The results indicate that there were no increments to the prediction of any of the criteria for the computer, JOB, or AVOICE composites. The spatial composite added a point to the prediction of GSP, AE, and JK-Total, and the ABLE composites added five points to the prediction of PFB and one point to the prediction of LDR.

Table 1.60
Mean of Incremental Correlations Over ASVAB Factors Computed Within Job for LVII Samples for Spatial, Computer, JOB, ABLE Composites, and AVOICE

Criterion	No. of MOS[a]	ASVAB Factors (A4) + Spatial [5]	A4 + Computer [12]	A4 + JOB [7]	A4 + ABLE Composites [11]	A4 + AVOICE [12]
CTP	7	65 (11)	63 (10)	65 (11)	65 (11)	64 (13)
GSP	6	62 (08)	62 (10)	60 (11)	60 (10)	57 (12)
AE	7	33 (10)	10 (14)	30 (15)	24 (20)	20 (18)
PD	7	16 (12)	16 (17)	14 (15)	13 (12)	06 (11)
PFB	7	12 (10)	08 (11)	13 (11)	18 (15)	11 (16)
LDR	7	63 (12)	61 (15)	63 (13)	65 (13)	62 (13)
LDR2	7	51 (17)	48 (20)	49 (20)	50 (24)	48 (19)
LDR3	7	45 (20)	41 (21)	42 (23)	45 (22)	40 (24)
HO-Total	7	46 (15)	43 (17)	44 (15)	43 (21)	39 (23)
JK-Total	7	75 (05)	73 (06)	74 (06)	74 (06)	73 (06)

Note: Predictor/criterion setwise deletion samples. Adjusted for shrinkage (Rozeboom formula 8) and corrected for range restriction. Numbers in brackets are the numbers of predictor scores entering prediction equations. Numbers in parentheses are standard deviations. Underlined numbers denote multiple Rs greater than for ASVAB Factors alone. Decimals omitted.

[a] Number of MOS for which validities were computed.

Comparison Between Validity Results Obtained with LVI and LVII Samples. The multiple correlations for the ASVAB factors and each set of experimental predictors as computed for LVI and LVII, respectively, are reported in Table 1.61. These results have been corrected for range restriction and adjusted for shrinkage.

Note that the soldiers included in the LVI validation samples were required to have complete criterion data, but the soldiers in the LVII sample were not. Also, the LVII analyses did not include two MOS (31C and 88M) that were included in the LVI analyses. Finally, as described earlier, there were differences between the components of the Achievement and Effort (AE) and Personal Discipline (PD) factors computed for soldiers in the LVII sample and their corresponding factors in the LVI sample (Effort and Leadership [ELS] and Maintaining Personal Discipline [MPD], respectively).

Table 1.61
Comparison of Mean Multiple Correlations Computed Within Job for ASVAB Factors, Spatial, Computer, JOB, ABLE Composites, and AVOICE Within LVI and LVII Samples

	No. of MOS[b]		ASVAB Factors [4]		Spatial [1]		Computer [8]		JOB [3]		ABLE Comp. [7]		AVOICE [8]	
Criterion[a]	LVI	LVII	LVI	LVII	LVI	LVII	LVI	LVII	LVI	LVII	LVI	LVII	LVI	LVII
CTP	9	7	63	64	58	57	49	53	31	33	21	24	39	41
GSP	8	6	66	63	65	58	55	48	32	28	24	19	38	29
ELS/AE	9	7	34	29	33	27	30	09	19	07	12	13	20	09
MPD/PD	9	7	16	15	14	15	10	12	06	03	15	06	05	06
PFB	9	7	12	16	08	13	13	03	07	07	28	17	09	09
HO-Total	9	7	50	46	50	41	38	33	20	24	13	12	30	21
JK-Total	9	7	73	74	66	67	60	58	38	37	30	29	43	44

Note: LVI setwise deletion samples; LVII predictor/criterion setwise deletion sample. Adjusted for shrinkage (Rozeboom formula 8) and corrected for range restriction. Numbers in brackets are the numbers of predictor scores entering prediction equations. Decimals omitted.

[a] CTP = Core Technical Proficiency; GSP = General Soldiering Proficiency; ELS = Effort and Leadership (LVI); AE = Achievement and Effort (LVII); MPD = Maintaining Personal Discipline (LVI); PD = Personal Discipline (LVII); PFB = Physical Fitness/Military Bearing; HO = Hands-On; JK = Job Knowledge

[b] Number of MOS for which validities were computed.

The results in Table 1.61 demonstrate that the patterns and levels of validities are very similar across the two sets of analyses, especially for the four ASVAB factor composites. For example, the multiple R between the ASVAB and the CTP is .63 and .64 for the LVI and LVII samples, respectively.

The greatest discrepancies between the two sets of results concern the multiple correlations between the ABLE composites and two of the three "will do" criterion factors - [Maintaining] Personal Discipline and Physical Fitness and Military Bearing. Some of the decrease in the ABLE's ability to predict Personal Discipline in LVII may be due to the removal from that factor of the Promotion Rate component. The validity of the other predictors was not similarly affected by this scoring change. Again, the highest correlation for the ABLE in LVII was with the Leadership factor (R = .34). The LVII Leadership factor included the promotion rate index, all scores derived from the supervisory role plays, and the Army-wide BARS (rating scale) leading and supervising factor, which was part of the ELS factor in CVI and LVI. In effect, these differences were expected to decrease the ABLE correlations with the LVII Achievement and Effort factor and increase the ASVAB correlations with this factor, which in LVII is more reflective of technical achievement than was the ELS factor in LVI and CVI. These are

the expected patterns and they lend further support to the construct validity of the performance models.

Summary and Conclusion

The basic validation results for the LVII sample produced results that were largely consistent with those obtained for LVI. In summarizing the prior validation results, Oppler, Peterson, and Russell (1994) concluded that the ASVAB was the best overall predictor of first-tour performance, but that the composite of spatial tests provided a small amount of incremental validity for the "can do" criteria (i.e., Core Technical Proficiency, General Soldiering Proficiency), and the ABLE provided larger increments for two of the three "will do" criteria (Maintaining Personal Discipline, Physical Fitness and Military Bearing). The same pattern of results was found in the present analyses. Furthermore, not only were the LVII results similar in pattern to those of the LVI analyses, they were also similar in magnitude (with the exception of some of the ABLE validities, which were lower in LVII).

Prediction of Future Performance From Current Performance and From Training Performance

The general question of how accurately individual job performance at one level in the organization predicts job performance at another level is virtually a "classic problem" in personnel research. In the Army context it is a question of the extent to which promotion or reenlistment decisions should be based on assessments of prior performance. A related issue is the extent to which initial job assignments should be based on assessments of performance in training.

The data from the Project A/Career Force (Campbell & Zook, 1990, 1991) permit some of the above issues to be addressed. The three samples of interest are from the longitudinal cohort that entered the Army in 1986/87. This group took the Experimental Test Battery at the start of basic training. Their performance in training was assessed at the end of their technical training course and their job performance was assessed approximately 18-24 months after they entered the Army. For those who reenlisted, their second-tour performance as a junior NCO was assessed during the second half of their second tour.

The samples, predictor tests, and performance measures that are used in this analysis have been described in detail elsewhere (Campbell & Zook, 1991). Extensive job analyses, criterion development, and analyses of the latent structure of MOS performance for both first-tour and second-tour MOS have attempted to produce a complete specification of performance at each level. The models of performance for training performance, first-tour performance, and second-tour performance provide some clear predictions about the pattern of convergent and divergent relationships that should be found.

Objectives

The specific questions addressed were the following.

(1) To what degree does an individual's level of performance in a first-tour enlisted position predict performance during his or her second tour?

(2) To what degree does performance in training predict subsequent job performance, both in first-tour enlisted positions and in second-tour positions?

(3) Given that performance is not unidimensional, do the separately measured components of performance exhibit the appropriate patterns of convergent and divergent validity when current performance is used to predict future performance?

(4) As a predictor of future performance, do measures of current performance add variance that is not accounted for by measures of ability, personality, and interests?

Samples

The LV sample consisted of approximately 45,000 new accessions who took the Experimental Predictor Battery during their first three days in the Army. Of the total sample, approximately 30,000 would enter the nine Batch A MOS.

The EOT sample for the Batch A MOS consists of those individuals who completed their Advanced Individual Training and from whom a set of End-of-Training performance measures were obtained during the last two days of the training period (approximately 26,000). The measures consisted of the project-developed EOT achievement test and a set of multiple peer ratings on ten rating scales dealing with technical competence, personnel discipline, effort level, leadership potential, and military bearing. Confirmatory factor analysis techniques were used to develop composites of the individual measures to reflect scores on six latent factors of performance in training.

The LVI performance measurement sample consisted of all the individuals in the original LV predictor sample who were available for performance assessment 18-24 months later at any one of 15 data collection sites in the United States, Europe, and Korea. For the Batch A MOS this was approximately 6,800 individuals. The first-tour job performance measures are described in detail in Campbell and Zook (1991) and summarized earlier in this chapter. The individual measures were subjected to a confirmatory modeling analysis (Oppler et al., 1994) and five simple sum factor scores were subsequently used as the best representation of the latent structure of job performance during the first tour of duty.

The final longitudinal follow-up of the 86/87 cohort focused on the soldiers in the Batch A MOS who reenlisted for a second tour and who could be located at any one of

16 data collection sites approximately 25-30 months after reenlistment. The total sample size across the nine MOS turned out to be approximately 1,550 individuals.

The individual performance measures consisted of a hands-on job sample test, a comprehensive job knowledge test, administrative/archival indices of performance, and multiple rating scales. In addition, the assessment measures for second tour performance included a paper-and-pencil test of situational leadership/supervisory judgment and three role-play simulations designed to measure certain aspect of supervisory/leadership skill having to do with counseling and training subordinates. The LVII array of second-tour performance measures is summarized earlier in this chapter.

As described by Hanson, Campbell, & McKee (in press), the second-tour performance measures were subjected to a confirmatory analysis and a six-factor solution represented the best fit for both LVII and CVII sample data.

In summary, extensive confirmatory analysis at each organizational level yielded a very consistent picture of the latent structure of performance, both across cohorts within levels (i.e., CVI and LVI, and CVII and LVII) and across levels within cohorts (e.g., EOT vs. LVI vs. LVII). The substantive content, or latent structure, of performance showed a strong tendency to be consistent where it should be (i.e., across cohorts) and different where it should be (i.e., across organizational levels).

To examine the patterns of correlations for the Project A/Career Force Project Longitudinal Validation sample, three basic intercorrelation matrices were computed (Campbell, Peterson, & Johnson, in press). Each matrix was calculated by computing the intercorrelations within each MOS and then averaging over MOS. All correlations were corrected for restriction of range by using a multivariate correction that treated the six EOT performance factors as the "implicit" selection variables on the grounds that, in comparison to other incidental selection variables, these factors would have the most to do with whether an individual advanced in the organization.

Results

The correlations of training performance with first-tour performance are shown in Table 1.62; the correlations of first-tour performance with second-tour performance are shown in Table 1.63; and the correlations of training performance and second-tour performance are shown in Table 1.64. In general, there are substantial correlations of performance with performance. Performance in training does predict performance as a first-tour job incumbent, and performance in the first tour of duty does predict performance in the second tour after reenlistment. Performance in training also predicts performance during the second tour, approximately 5-6 years later. This is true both for the variables measured with standardized tests and job samples and for the variables assessed via peer ratings.

Table 1.62
Zero-Order Correlations of Training Performance (EOT) Variables With First-Tour Job Performance (LVI) Variables: Weighted Average Across MOS

LVI Variables	EOT Variables[a] EOT:TECH	EOT:BASE	EOT:ETS	EOT:MPD	EOT:PFB	EOT:LEAD	EOT:ELS	EOT:CAN	EOT:WILL	EOT:TOT
Core Technical Proficiency (CTP)	.482 3857	.380 3582	.217 3843	.153 3843	.049 3843	.180 3843	.208 3843	.475 3582	.180 3843	.485 3535
General Soldiering Proficiency (GSP)	.493 3857	.452 3582	.230 3843	.171 3843	.043 3843	.162 3843	.203 3843	.526 3582	.181 3843	.534 3535
Effort and Leadership (ELS)	.209 3795	.167 3525	.354 3783	.250 3783	.277 3783	.353 3783	.376 3783	.208 3525	.365 3783	.251 3479
Maintain Personal Discipline (MPD)	.174 3908	.136 3633	.310 3894	.355 3894	.214 3894	.272 3894	.307 3894	.170 3633	.340 3894	.211 3586
Physical Fitness and Bearing (PFB)	-.011 3908	-.016 3633	.262 3894	.127 3894	.444 3894	.308 3894	.307 3894	-.015 3633	.330 3894	.031 3586
"Can Do" Performance Composite (CAN)	.530 3857	.451 3582	.245 3843	.177 3843	.050 3843	.187 3843	.226 3843	.545 3582	.197 3843	.555 3535
"Will Do" Performance Composite (WILL)	.167 3795	.128 3525	.386 3783	.302 3783	.373 3783	.389 3783	.413 3783	.163 3525	.427 3783	.216 3479
Total Performance Composite (TOT)	.388 3741	.322 3471	.407 3729	.314 3729	.298 3729	.379 3729	.416 3729	.394 3471	.413 3729	.438 3425
NCO Potential Rating <Supv> (NCO)	.165 3458	.140 3458	.309 3444	.224 3444	.259 3444	.306 3444	.327 3444	.169 3458	.324 3444	.208 3397

Note. Corrected for range restriction. Pairwise Ns are printed below each correlation. Correlations between matching variables are underlined.

[a] TECH = Technical Knowledge Score; BASE = Basic Knowledge Score; ETS = Effort and Technical Skill; LEAD = Leadership Potential.

There is also a reasonable pattern of convergent and divergent validity across performance factors, even without correcting these coefficients for attenuation and thereby controlling for the effects of differential reliability. The one possible exception is the predictability of the leadership performance factor for second-tour personnel. This component of NCO performance is predicted by almost all components of past performance.

To address the issue of whether information about past performance contributes unique variance to the prediction of future performance over that contained in measures of ability, personality, and interests, two types of hierarchical regressions were carried out. The first specified that the order of entry would be past performance first, followed by the four factor scores from the ASVAB, and then followed by the eight composite scores from the AVOICE and then the seven factor scores from the ABLE, in that order. The second sequence was similar to the first except that the order of past performance and ASVAB was reversed.

Table 1.63
Zero-Order Correlations of First-Tour Job Performance (LVI) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS

	LVI Variables								
LVII Variables	LVI:CTP	LVI:GSP	LVI:ELS	LVI:MPD	LVI:PFB	LVI:CAN	LVI:WILL	LVI:TOT	LVI:NCO
Core Technical Proficiency (CTP)	.440 412	.413 412	.249 400	.078 413	.015 412	.449 412	.181 400	.375 397	.230 379
General Soldiering Proficiency (GSP)	.511 412	.569 412	.219 400	.085 413	-.008 412	.578 412	.157 400	.440 397	.220 379
Achievement and Effort (AE)	.103 390	.167 390	.450 377	.280 390	.319 390	.150 390	.464 377	.470 374	.412 353
Leadership (LEAD)	.359 344	.411 344	.379 333	.272 343	.169 342	.421 344	.365 333	.517 332	.378 319
Leadership Minus SJT Score	.264 348	.310 348	.372 337	.249 347	.233 346	.321 348	.380 337	.467 336	.398 322
Achievement, Effort and Leadership	.275 333	.335 333	.471 322	.292 332	.264 331	.336 333	.459 322	.620 321	.444 307
Maintain Personal Discipline (MPD)	-.044 406	.038 406	.116 393	.257 406	.166 406	-.002 406	.211 393	.164 390	.114 370
Physical Fitness and Bearing (PFB)	-.026 392	-.013 392	.220 379	.135 392	.460 392	-.022 392	.333 379	.250 376	.265 356
"Can Do" Performance Composite (CAN)	.520 412	.533 412	.259 400	.097 413	.010 412	.562 412	.193 400	.452 397	.260 379
"Will Do" Performance Composite (WILL)	.141 321	.190 321	.370 310	.295 320	.347 319	.182 321	.433 310	.445 309	.404 296
Total Performance Composite (TOT)	.336 313	.357 313	.381 302	.240 312	.252 311	.375 313	.394 302	.521 301	.423 289

Note. Corrected for range restriction. Pairwise Ns are printed below each correlation. Correlations between matching variables are underlined.

The general findings from these analyses seemed clear. Both past performance and measured abilities add unique variance to the prediction of future performance. Knowledge of past performance adds relatively more, in comparison to trait measures, to the prediction of the "will do" components of future performance than to the task performance based on "can do" components. Conversely, the cognitive ability measure (ASVAB) adds more to the prediction of future performance on the "can do" factors while the ABLE adds relatively more to the prediction of the "will do" components.

Table 1.64
Zero-Order Correlations of Training Performance (EOT) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS

	EOT Variables[a]									
LVII Variables	EOT:TECH	EOT:BASE	EOT:ETS	EOT:MPD	EOT:PFB	EOT:LEAD	EOT:ELS	EOT:CAN	EOT:WILL	EOT:TOT
Core Technical Proficiency (CTP)	.479 1014	.413 960	.215 1056	.147 1056	.080 1056	.174 1056	.204 1056	.484 960	.183 1056	.480 936
General Soldiering Proficiency (GSP)	.488 1014	.429 960	.192 1056	.107 1056	.064 1056	.112 1056	.155 1056	.496 960	.139 1056	.489 936
Achievement and Effort (AE)	.098 946	.151 896	.248 983	.172 983	.189 983	.238 983	.258 983	.141 896	.250 983	.165 874
Leadership (LEAD)	.322 900	.387 854	.294 931	.191 931	.152 931	.254 931	.289 931	.396 854	.264 931	.416 832
Leadership Minus SJT Score	.202 905	.284 905	.293 936	.203 936	.187 936	.276 936	.302 936	.274 905	.284 936	.302 881
Achievement, Effort and Leadership	.238 856	.310 811	.281 886	.197 886	.168 886	.258 886	.285 886	.307 811	.268 886	.328 791
Maintain Personal Discipline (MPD)	.080 1006	.086 953	.210 1045	.260 1045	.162 1045	.210 1045	.224 1045	.091 953	.249 1045	.117 929
Physical Fitness and Bearing (PFB)	-.047 967	-.007 916	.123 1003	.067 1003	.320 1003	.208 1003	.183 1003	-.032 916	.208 1003	-.005 893
"Can Do" Performance Composite (CAN)	.527 1014	.457 960	.228 1056	.145 1056	.082 1056	.160 1056	.201 1056	.534 960	.181 1056	.530 936
"Will Do" Performance Composite (WILL)	.168 823	.221 780	.270 852	.218 852	.235 852	.281 852	.295 852	.215 780	.297 852	.242 761
Total Performance Composite (TOT)	.375 805	.386 762	.301 831	.225 831	.208 831	.273 831	.303 831	.417 762	.297 831	.434 743

Note. Corrected for range restriction. Pairwise Ns are printed below each correlation. Correlations between matching variables are underlined.

[a] TECH = Technical Knowledge Score; BASE = Basic Knowledge Score; ETS = Effort and Technical Skill; LEAD = Leadership Potential.

Prediction of First-Term Military Attrition Using Pre-Enlistment Predictors

These analyses (McCloy & DiFazio, in press) had two main goals. The first was to determine the relationship between first-term attrition and the three non-cognitive predictor measures: the Assessment of Background and Life Experiences (ABLE), the Army Vocational Interest Career Examination (AVOICE), and the Job Orientation Blank (JOB). The second goal was to use all the available predictor data to develop a specific predictor composite for attrition that could be used to select applicants who have higher probabilities for completing their first term of service.

The relationship between first-term attrition and the pre-enlistment predictors was addressed using proportional hazards modeling (Cox. 1972), a form of event history analysis (cf. Allison, 1984). A proportional hazards model allows the relationship between one or more predictor variables and the rate at which events occur over time to be determined. The two principal functions used in event history analyses are the survivor function and the hazard function.

The survivor function, S(t), describes the probability that an individual will survive at least until time t without experiencing the event in question (e.g., attrition). S(t) is a monotonic, non-increasing function that is essentially a reverse cumulative distribution, cumulating across time the proportion of observations that have yet to experience the event. For example, a plot of survival curves for the first-term attrition of high school graduates and non-graduates who enlisted for three years into MOS 11B (Infantryman) indicates that the survival rate of graduates is much higher than for non-graduates. It takes just under two years for the original sample of graduates to be reduced to 80 percent, whereas it takes just 10 months to reduce the sample of non-graduates to 80 percent of its original size.

Because the survivor function is monotonically non-increasing, its shape remains relatively unchanged, regardless of the rate at which events occur over time; however, the probability of leaving the military increases rapidly during the first three months of service and decreases to a relatively stable rate thereafter. The function describing the distribution of event occurrence across time is h(t), the hazard function. The dependent variable for the hazard function is the relative rate at which the event is occurring at any given point, or interval, in time.

The Proportional Hazards Model

Since its introduction by Cox in 1972, the proportional hazards model has become one of the most widely used event history models. Formally, the model is:

$$\ln [h(t)] = \ln [h_0(t)] + \beta X \qquad (1)$$

where ln is the natural logarithm, h(t) is the hazard rate, $h_0(t)$ is a baseline hazard rate that can take any form (i.e., it is non-parametric), β is a vector of regression coefficients, and $\mathbf{X}$ is a vector of predictor variables.

The model gets it name from the fact that for any two individuals with covariate vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, the ratio of their hazards is a constant value, k. If the hazards prove not to be proportional, a stratified analysis may be conducted where "group" is the stratifying variable. The baseline hazards are allowed to differ for each group, but the regression parameters are assumed to be the same across groups. This is somewhat analogous to estimating regression equations that allow group intercepts to differ while constraining the slopes to be equal across groups.

Whether the hazards are proportional for all groups or only within stratified groups, the effect of the independent variables on the hazard is to shift the baseline hazard up or down, depending on the values of the variables.

An Application to First-Term Attrition

The proportional hazards model was used to analyze attrition data in the Project A/Career Force Longitudinal Validation sample. The specific objective was to determine the extent to which information from the Project A/Career Force Experimental Battery could add to the prediction of attrition over the full course of the first tour of duty.

Subjects and Measures. The subjects were the approximately 49,000 first-term soldiers from the Project A/Career Force Longitudinal Validation (LV) sample (Campbell & Zook, 1994a). Rather than running analyses by MOS, job groups were formed by splitting the 21 MOS into two groups: Combat (MOS 11B, 12B, 13B, 16S, 19E, and 19K) and Non-Combat (all others). These two groups--C and NC--were further subdivided by enlistment terms (3 years or 4 years).

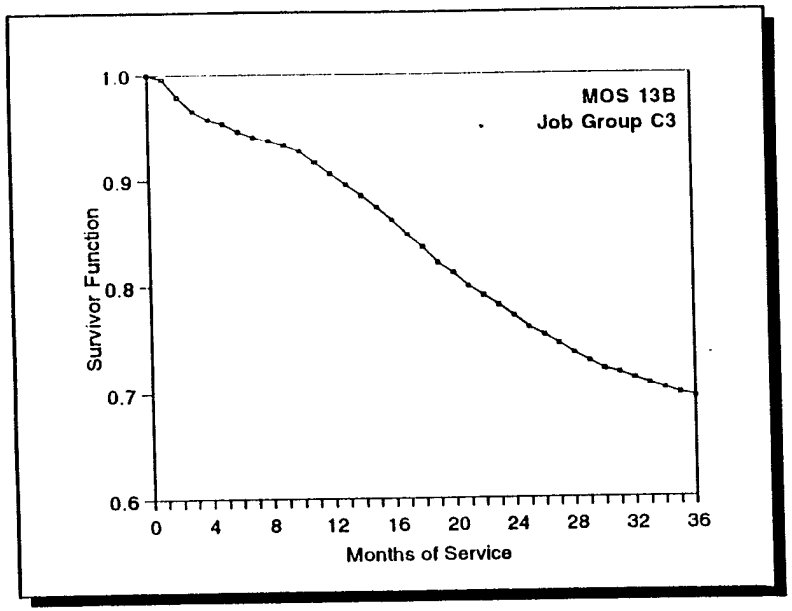
All measures were taken from the Project A/Career Force data base. The predictor measures were the same ones used in the LVI validation analyses.

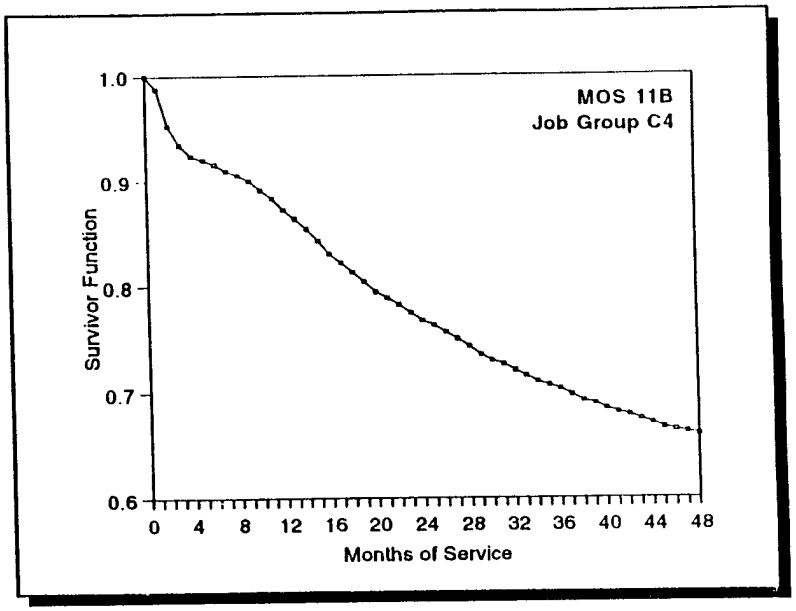
Attrition is defined as a premature separation from first-term service for reasons that are viewed negatively from the military perspective. As developed in another project sponsored by the Office of the Secretary of Defense, the Compensatory Screening Model (CSM) was used to group separation types into "pejorative" and "non-pejorative" categories (McBride, 1993). Separations that the CSM identified as "pejorative," and that correspond to Knapp's (1993) Army separation behavior categories four and five were used to define negative separations.

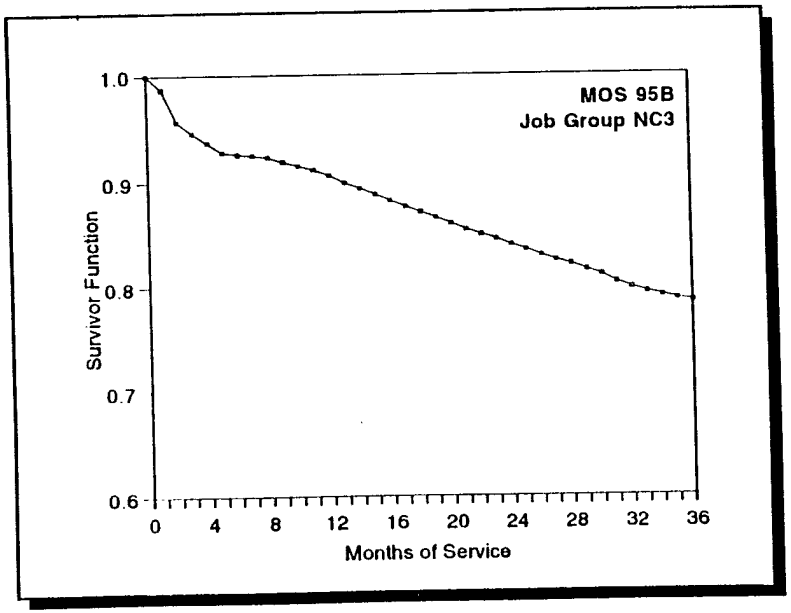
Analysis. Before estimating the proportional hazards regression equations, an empirical test of the proportionality assumption was conducted for each MOS within each of the four job groups by examining the significance of the parameter for a single MOS dummy variable (D_{MOS}) interacting with time. In all instances, the parameter was significant at $p = .001$, which rejects the proportionality assumption. The baseline survivor and hazard functions for one MOS from each of the four subgroups are given in Figures 1.14 and 1.15, respectively.

Because the hazards proved not to be proportional, a stratified analysis was conducted with MOS as the stratifying variable. Two types of proportional hazard analyses were performed for each of the four job groups.

The first entered predictor blocks hierarchically. The first block contained the four ASVAB composites and the dummy variable HSDG denoting high- school diploma graduate status. This block comprises the "base" variables -- the pre-enlistment variables currently available to the Army, against which incremental prediction was assessed, and they were included in all models. Blocks two through four contained the ABLE, AVOICE, and JOB scores, respectively. The incremental fit afforded by each block was







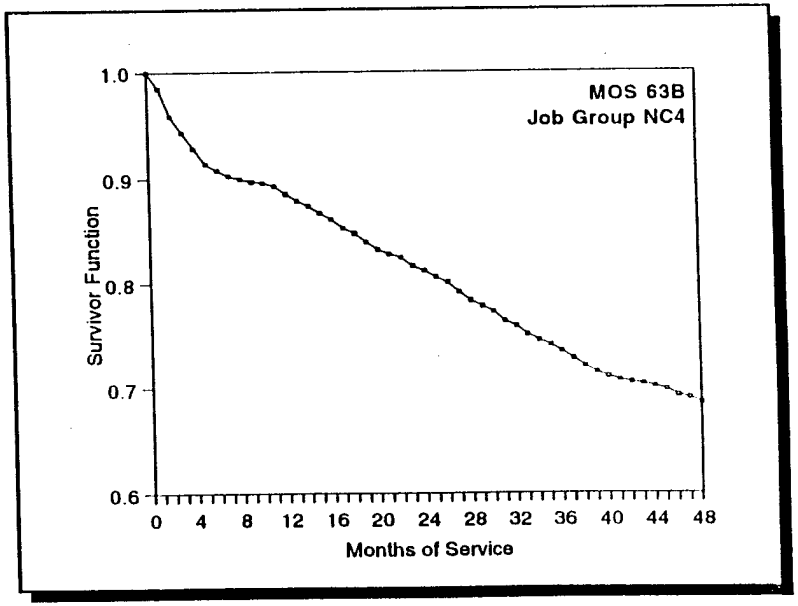
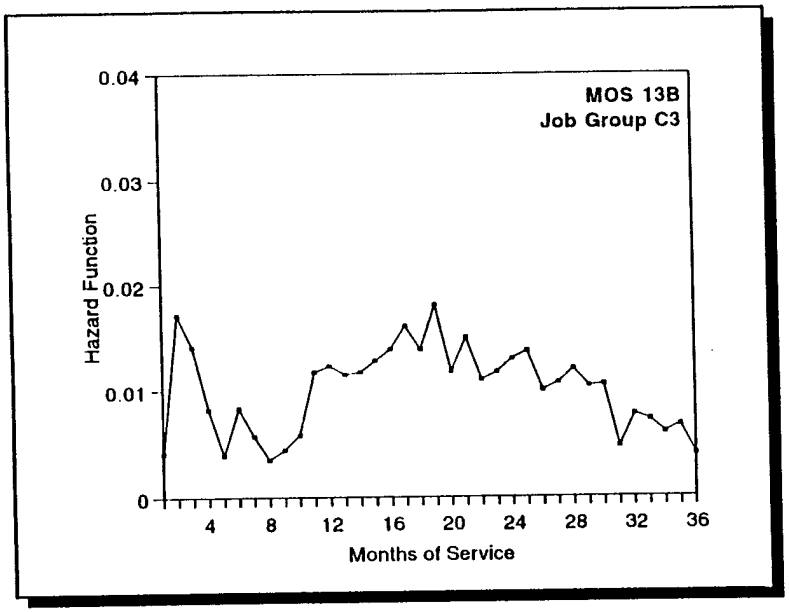
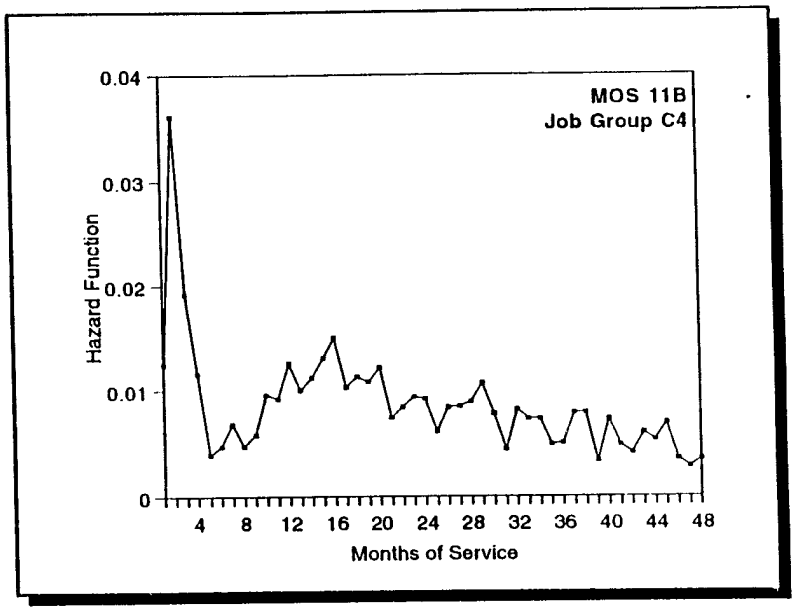
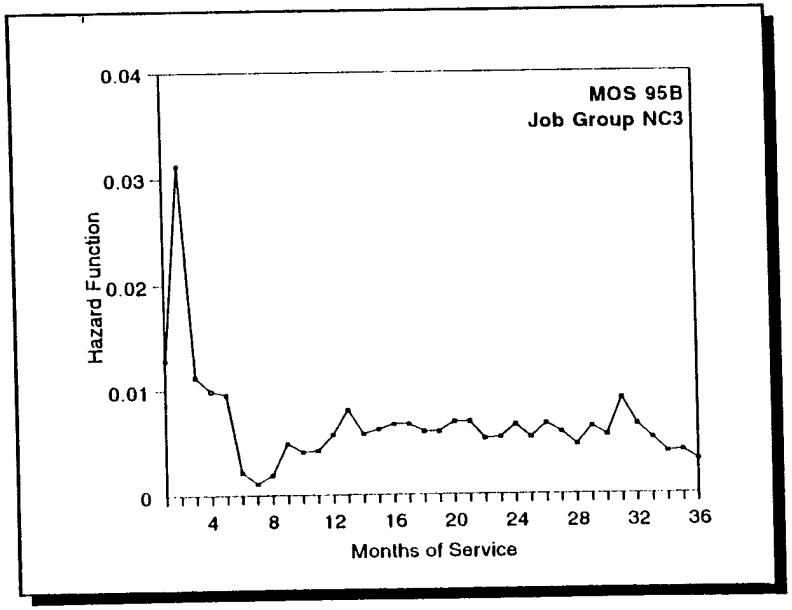


Figure 1.14. Baseline survivor functions for one MOS from each of the four job groups.







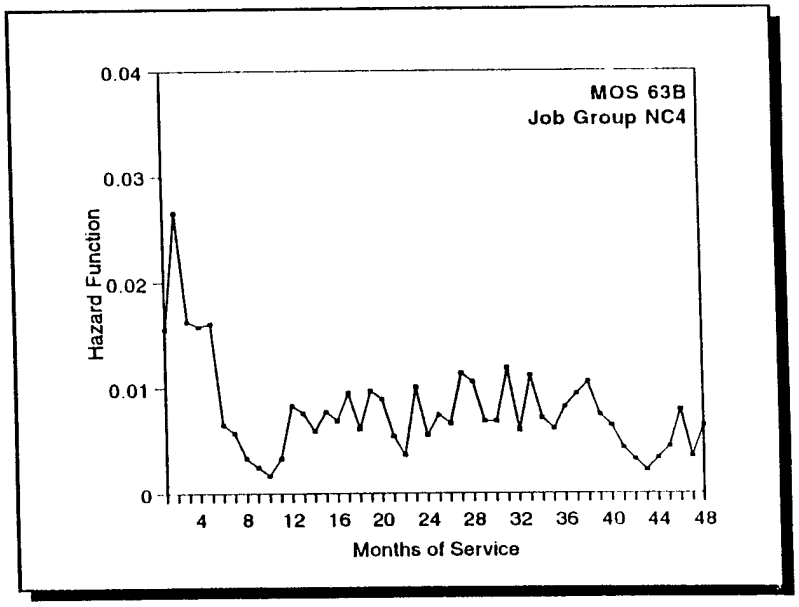


Figure 1.15. Baseline hazard functions for one MOS from each of the four job groups.

assessed relative to block one. Finally, a model was estimated containing all four blocks of predictors.

The second type of analysis used a best subset selection algorithm in the SAS procedure PHREG, in which the best j equations containing k specified predictors are identified. The model considered to be best for a given number of predictors is the one yielding the highest global score chi-squared statistic. Here, the best j = 3 equations containing k= 1, . . ., 26 predictors were obtained. The value for k is a function of the number of predictions that produce statistically significant increments in model fit. Likelihood ratio tests were calculated to evaluate the point at which additional predictors failed to increase the fit of the model significantly. Nested equations were selected from the best equations and re-evaluated in PHREG. Stringent p values were selected, due to the large samples and the dependence of X^2 on sample size. The effect of each variable, conditional on all other variables in the model, was also taken into account when deriving the "best" equation for each of the four groups.

Results. The results indicated the following:

- The base variables are significantly related to first-term attrition.
- The ABLE provides significant incremental fit to the base model for all four job groups.
- The AVOICE significantly increases the fit of the base model for 3-year enlistments but not 4-year enlistments. Although the sample sizes differ by a factor of nearly two to one (N3 = 20,252 and N4 = 10,780), this is probably not a power issue, given the absolute size of each group.
- Except for NC3, the JOB does not provide a statistically significant increase in fit over the base model.
- Addition of the AVOICE and JOB to the ABLE model increases the fit of the model to the data for job groups C3 and NC3 (Model 2 vs. Model 5); again, this is probably not a power issue.

Thus, the non-cognitive measures improve prediction of first-term attrition over and above current pre-enlistment information.

Best Subset Selection

The "variable traces" of the nested models for the best subset selection analyses end at the point of the last significant increase in model fit. All models began with k = 3 predictors.

The number of significant predictors ranged from seven to eleven. No AVOICE or JOB scales appear in any of the equations. Although the equations are slightly different across the job groups, six variables appear in each best model: high school

diploma graduate status, and five ABLE scales (Nondelinquency, Dominance, Physical Condition, Self-Esteem, and Social Desirability). In addition, the ASVAB Quantitative composite appears in all but one best equation for both Combat groups.

The results also indicate that the largest conditional effects are quite similar across the four job groups, with the largest effects provided by the ABLE scales Nondelinquency and Dominance. A higher Nondelinquency score translates into a decreased hazard for attrition, whereas the opposite is true for Dominance. For example, a one-unit (standard deviation) increase in the Nondelinquency score for Combat Soldiers with three-year enlistments decreases the hazard for attrition to 76.9 percent of its current level. Equivalently (and perhaps easier to picture), a one-unit (standard deviation) decrease in the Nondelinquency score yields an increase in the hazard of 30 percent.

One might ask what increase in survivability would be attained if the seven-variable attrition composite were used to screen recruits. The appropriate baseline for comparison would be the attrition behavior of the present sample first-term soldiers.

Figures 1.16 through 1.19 demonstrate the effects of the truncation on survivability. For all four plots, the truncation point of 10 percent translates to an increase in the expected survivability of three to four months, whereas the increase evidenced at 25 percent is seven to eight months. The maximal truncation of 33 percent affords an increase in survivability of between nine and ten months, a relatively small increase in survival time over the more modest 25 percent screen.

A second way of looking at the curves is in terms of the percentage of the groups remaining after six months. Taking a weighted average of the increase in retention, we would expect an average increase in cohort survival across the four job groups of 0.9 percent, 1.8 percent, and 2.1 percent after six months with truncations of 10 percent, 25 percent, and 33 percent, respectively. Similarly, at the two-year point, the expected gains are 1.6 percent, 4.3 percent, and 5.0 percent, respectively.

One central question is how much of a contribution the ABLE scales make to the prediction of attrition, over and above high school diploma graduate status. To examine this question, the five ABLE scales were standardized within job group, unit weighted (with Dominance and Social Desirability weighted negatively), and summed. This composite score was then standardized within job group. For each job group, high school diploma graduates and non-graduates were identified and split into two groups: those scoring above the mean and those scoring at or below the mean.

The effect of high and low ABLE scores on the survivability of graduates and non-graduates is most clearly demonstrated by the survivor functions for the four groups. High school diploma graduate status had a strong effect, with graduates having higher survival rates than non-graduates. For both groups of soldiers, those scoring high on the composite formed by the five ABLE scales that are components of the attrition composite have a better rate of survivability over those from the same educational group who score low on the composite.

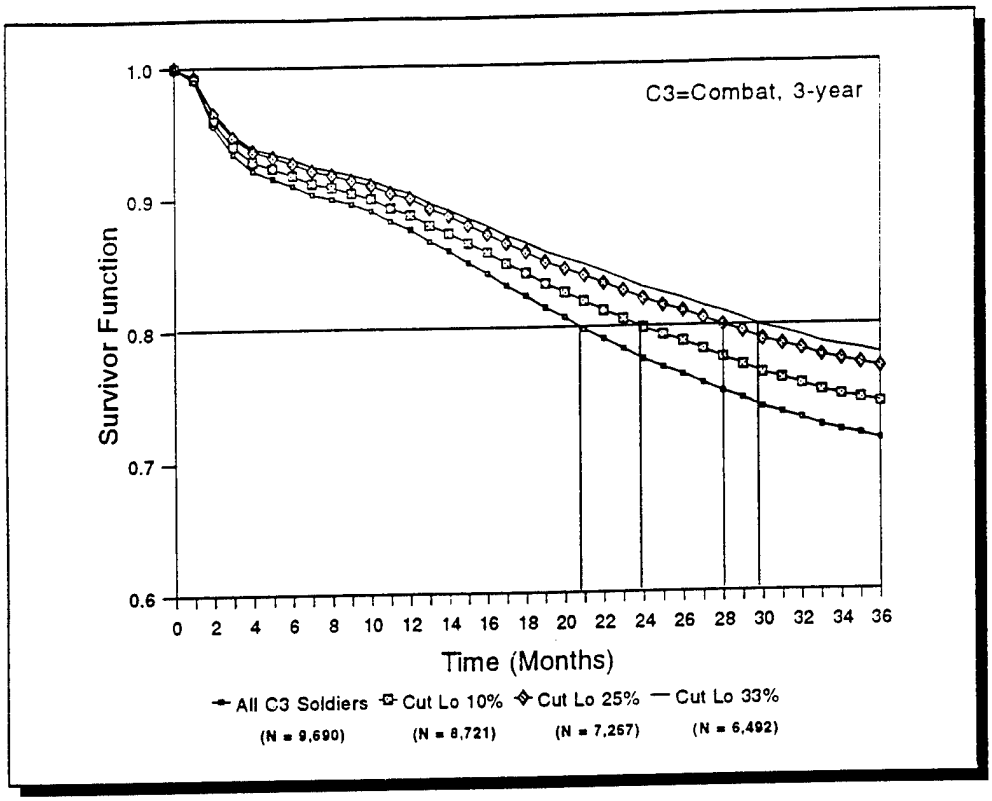


Figure 1.16. Survivor functions for C3 soldiers scoring above various truncation points on the attrition composite.

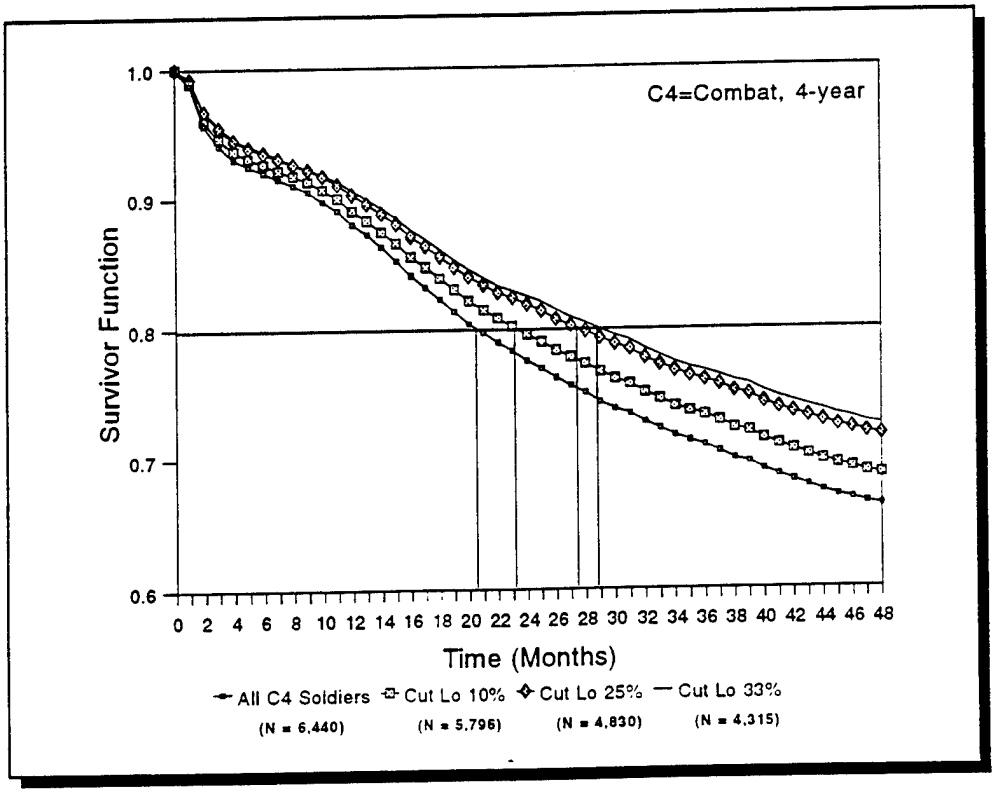


Figure 1.17. Survivor functions for C4 soldiers scoring above various truncation points on the attrition composite.

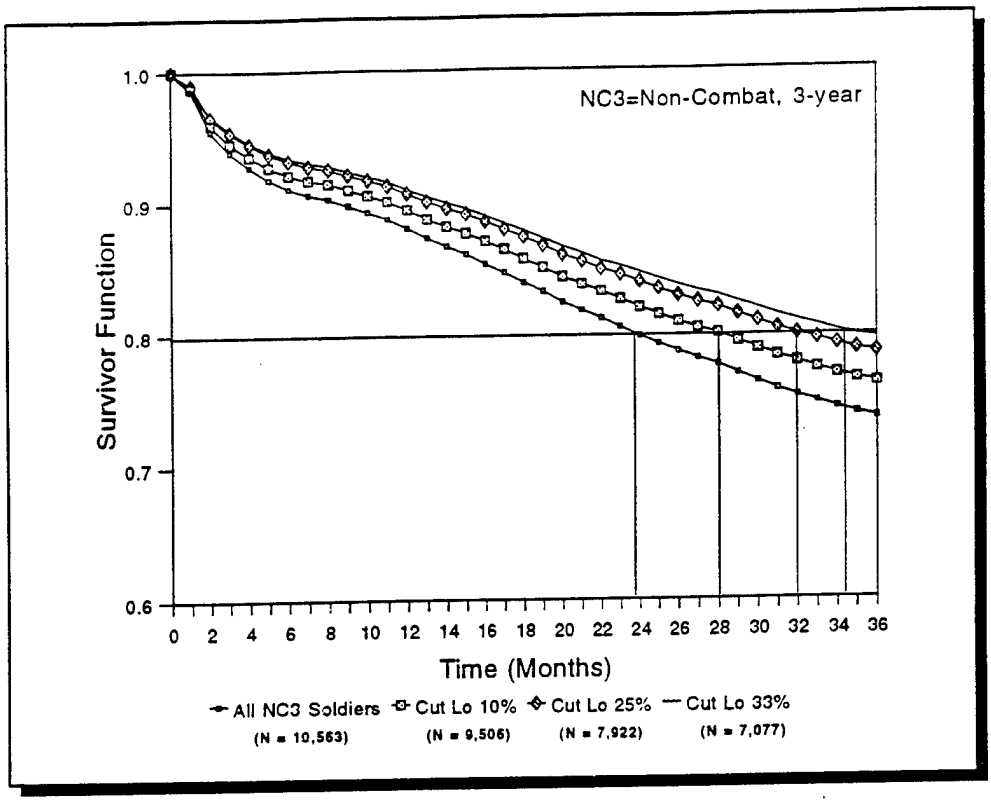


Figure 1.18. Survivor functions for NC3 soldiers scoring above various truncation points on the attrition composite.

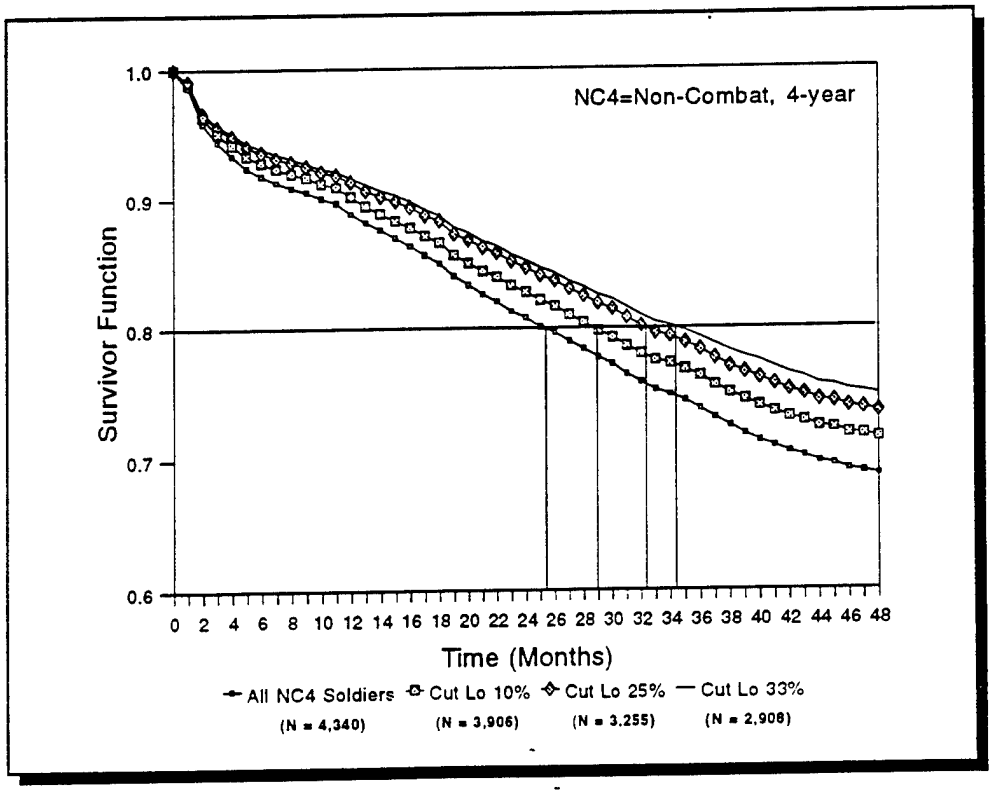


Figure 1.19. Survivor functions for NC4 soldiers scoring above various truncation points on the attrition composite.

Perhaps the most striking feature was that there is a greater increase in survivability for those high on ABLE for diploma graduates than for non- graduates. That is, although non-graduates who score high on the ABLE demonstrate increased survivability over non-graduates who score low on the ABLE, the difference between survival curves is much larger for graduates.

Summary

The results of the event history analysis of first-term attrition indicate that high school diploma graduate status remains a very powerful predictor, with the hazard rate for non-graduates being approximately twice that for graduates. Nevertheless, information provided by non-cognitive measures (specifically, scales from the ABLE) provides significant incremental prediction to pre-enlistment information currently available to the Army. Information provided by the AVOICE and the JOB contributed incremental prediction on occasion, but both were overshadowed when the ABLE was in the equation.

The use of regression typically raises the question of cross validity. Event history models do not produce a measure similar to the coefficient of determination, obviating the use of shrinkage formulae. Nevertheless, a procedure has been suggested for assessing the fit of a proportional hazards regression model in a second sample (McCloy, 1993). To the extent that the procedure is viable, questions of cross validity may be addressed.

The Role of Job Satisfaction in Performance, Attrition, and Reenlistment

The primary purpose of the job satisfaction measure was to serve as a predictor of turnover--both non-reenlistment and attrition. The measure was also intended to serve as an additional criterion measure that might be predicted by measures of temperament, interests, and cognitive ability (Knapp, Carter, McCloy, & DiZazio, in press).

It was hypothesized that there would be moderate correlations between job satisfaction and contextual performance indices, but negligible correlations between job satisfaction and scores on "can do" measures of the technical core of the job (i.e., job knowledge and hands-on test scores). In these analyses, the Effort and Leadership and Personal Discipline factor scores were considered to be indices of contextual performance.

Satisfaction and Turnover/Attrition

Very little research has examined the relationship between job satisfaction and military attrition (i.e., separation before the enlistment contract expires). This is understandable given that most attrition occurs early in the enlistment term and is hypothesized to be related to problems of adaptability rather than to job dissatisfaction per se. Although there is more research examining the relationship between job satisfaction and military reenlistment behavior, most studies of reenlistment focus on

factors other than job satisfaction (e.g., pay, reenlistment bonuses, spouse support, education).

It was hypothesized that both job satisfaction and reenlistment intention would be significantly correlated with reenlistment. It was further hypothesized that job satisfaction would provide incremental predictive value for predicting reenlistment over that provided by intention to reenlist. Parallel hypotheses were made for the prediction of attrition. That is, it was expected that job satisfaction would be significantly related to attrition even when the intention to attrit was included in the prediction equation.

Because of the paucity of directly related research, job performance indices were included in the reenlistment and attrition analyses on an exploratory basis. That is, no specific hypotheses related to job performance were proposed.

Development of the Army Job Satisfaction Questionnaire (AJSQ)

A pool of several hundred items was assembled from previously developed military instruments related to satisfaction. Five project researchers independently reviewed the items and identified several common facets of satisfaction that might be useful to incorporate in the new instrument. A draft of the AJSQ was developed by drawing items from previously-existing instruments to cover each of the categories of satisfaction selected for measurement. A second section of the draft AJSQ was designed to help explain differences in job satisfaction levels expressed in Section I. Items in Section II were related primarily to reasons for enlisting and reenlistment intentions, and most were drawn from other Army surveys (e.g., the Army's New Recruit Survey). Section I of the field test version comprised a total of 62 items covering six facets of satisfaction: satisfaction with supervision, co-workers, promotions, pay, work, and the Army as an organization. Section II included 30 items.

The field test version of the AJSQ and the short form of the Minnesota Satisfaction Questionnaire (MSQ; Weiss, Dawis, England, & Lofquist, 1967) were administered to 271 first- and second-tour soldiers at three Army installations. Factor analysis of the Section I satisfaction items resulted in the expected six-factor structure and the total number of items was reduced from 62 to 37. The pattern of correlations demonstrated reasonable levels of discriminant and convergent validity for the AJSQ subscales.

To conserve administration time, the AJSQ was shortened again before it was administered to soldiers in the LVI/CVII data collection. Section I of the LVI/CVII version contained 20 items related to the six facets of satisfaction, with three to four items defining each facet. Section II included 13 items.

Prior to administration of the questionnaire to soldiers in LVII, the AJSQ was revised once again. In Section I, two items were added to the work subscale and a single overall satisfaction item was added as well. Section II was substantially revised and expanded. Four items were dropped and 25 items were added for a total of 34 items. Again, most of the items came from existing Army surveys (e.g., Army Career

Satisfaction Survey, Army Families Research Questionnaire). They cover (a) reasons for enlisting, (b) reenlistment intentions, (c) unit morale, (d) perceived fairness of treatment, and (e) the impact of socio-political events (e.g., Operation Desert Storm, Army downsizing) which had occurred since administration of the LVI AJSQ measure.

Results

AJSQ data were collected from 11,140 soldiers in the first-tour Longitudinal Validation sample (LVI), 1,025 second-tour soldiers from the Concurrent Validation sample (CVII), and 1,574 second-tour soldiers from the Longitudinal Validation sample (LVII).

Scale intercorrelations are provided in Table 1.65. As expected, scales showed low to moderate correlations with each other. The analyses that were conducted and the results that were obtained are summarized in turn for (a) the relationship of satisfaction to performance and (b) the prediction of turnover.

In the FY93 report, the research questions associated with satisfaction, performance, and turnover were addressed using data from the LVI sample. Related analyses using the CVII and LVII data will be described at a later date. Note also that the analyses reported are restricted to the nine Batch A MOS because most required soldiers to have a complete array of job performance data.

Table 1.65
AJSQ Score Intercorrelations

Scale	Supervisors	Co-Workers	Promotions	Pay	Work	Army	Comp
Supervision	--	.19	.30	.21	.38	.33	.62
Co-Workers	.16	--	.25	.17	.31	.28	.52
Promotions	.16	.18	--	.33	.38	.39	.67
Pay	.13	.17	.25	--	.30	.48	.62
Work	.32	.31	.25	.18	--	.42	.74
Army	.22	.20	.30	.42	.30	--	.75
Composite	.55	.50	.59	.57	.70	.68	--
Overall Satisfaction	.32	.28	.28	.24	.63	.47	.64

Note. Upper diagonal is LVI (n = 11,140); lower diagonal is LVII (n = 1,574).

Satisfaction and Performance. The sample for this set of analyses comprised 6,352 LVI soldiers from the nine Batch A MOS. In addition to the six subscores and AJSQ composite score described above, a "Challenge" subscore was computed by summing two items tapping satisfaction with job challenge and satisfaction with the opportunity to use skills and abilities.

Table 1.66 shows sample-size weighted mean correlations between the eight job satisfaction scores and 19 job performance scores across the nine MOS. Correlations based on the measures of contextual performance are printed in bold type. Mean correlations across MOS are presented because analyses showed that, except for correlations based on satisfaction with promotions, most or all of the interoccupation variance in these correlations could be attributed to sampling error.

Several points can be noted. First, the correlations tend to be low. Second, the relative magnitude of the correlations based on different performance indices tends to support the a priori hypotheses. The only correlations that do not support the hypotheses are those based on the Awards and Certificates variable. Third, patterns of correlations vary widely by job satisfaction dimension.

A further comparison of the correlations based on supervisor and peer ratings revealed consistently higher correlations between job satisfaction and supervisor evaluations of performance than between job satisfaction and peer evaluations of performance, except for the correlations based on the pay and coworkers subscales. Correcting for criterion unreliability did not substantially influence the relative sizes of the correlations.

Prediction of Turnover. The base sample for the turnover analyses consisted of 5,721 LVI soldiers representing the nine Batch A MOS. Most analyses used the overall satisfaction composite score and an overall job performance score as independent variables. The overall performance score was computed by summing the five standardized factor scores.

An intention to reenlist variable was extracted from Section II of the AJSQ. Approximately 69 percent of the soldiers in this sample indicated that they would leave the Army, 15.5 percent indicated that they would reenlist, and 15.5 percent indicated that they were undecided. An intention to attrit variable was obtained from a survey designed by the Army Family Research Program which was administered in conjunction with the LVI data collection (Research Triangle Institute, 1988). Three-year enlistees had a 10 percent attrition rate during the period between testing and the end of their enlistment term, compared to a 20.5 percent for 4-year enlistees. Base rate differences across enlistment terms were controlled by conducting turnover analyses for soldiers in each enlistment term separately.

To assess turnover, reenlistees were compared to soldiers who completed their initial term of enlistment and who were eligible to reenlist. Soldiers who left the Army before completing their enlistment terms or who were otherwise ineligible to reenlist were excluded. However, analyses that used job performance as the only type of

Table 1.66
Mean Correlations Between Job Satisfaction and Performance Weighted by Sample Size

Performance Measure	Army Job Satisfaction Questionnaire (AJSQ) Scales							
	Supervisor	Co-Workers	Promotions	Pay	Work	Army	Overall Composite	Challenge
Newly Developed Contextual Performance Variables								
Contextual Army-Wide Rating Composite	**.184**	**.090**	**.149**	**.104**	**.152**	**.144**	**.212**	**.124**
Contextual Combat Scales Composite	**.162**	**.083**	**.135**	**.092**	**.117**	**.113**	**.178**	.091
Army-Wide Rating Composites								
Technical Skill and Effort	.155	.088	.129	.070	.132	.115	.177	.107
Personal Discipline	**.185**	**.089**	**.156**	**.106**	**.148**	**.129**	**.209**	**.124**
Physical Fitness/Bearing	.112	.099	.143	.063	.122	.110	.165	.107
Overall Army-Wide Ratings Composite	.183	.105	.161	.093	.159	.140	.216	.131
Administrative Variables								
Awards and Certificates	.024	.030	.042	.016	.025	.035	.043	.022
Disciplinary Actions	-.083	-.048	-.117	-.090	-.107	-.098	-.140	-.099
M16 Qualification	-.013	.022	-.038	-.052	-.038	-.034	-.042	-.028
Other Measures								
Overall MOS - Specific Rating Composite	.129	.089	.108	.057	.122	.086	.152	.102
Overall Effectiveness Rating	.164	.096	.137	.073	.138	.124	.188	.115
NCO Potential Rating	.168	.103	.157	.077	.137	.127	.196	.111
Hands-on Score	.068	.023	.030	.021	.023	.041	.052	.021
Knowledge Test Score	.109	-.002	.035	.080	.018	.069	.077	.006
Performance Model Factors								
Effort and Leadership	**.148**	**.096**	**.131**	**.067**	**.126**	**.118**	**.175**	.104
Personal Discipline	**.174**	**.088**	**.233**	**.124**	**.159**	**.146**	**.237**	**.134**
Core Technical Proficiency	.096	.020	.039	.049	.043	.058	.077	.036
General Soldiering Proficiency	.088	-.010	.020	.066	-.005	.050	.051	-.012
Physical Fitness/Military Bearing	.071	.093	.104	.025	.085	.077	.114	.070

Note: Sample sizes range from 5,222 to 6,345. Performance rating scores are based on combined peer and supervisor rating data. Correlations based on measures of contextual performance are printed in bold type.

predictor did include soldiers who were ineligible for reenlistment. The reenlistment rate across MOS and enlistment terms, excluding those soldiers who were ineligible, was 38 percent. This is substantially higher than the rate for reenlistment intention, which was 15 percent. For the attrition analyses, soldiers who separated prematurely for avoidable reasons were compared to those who did not. In this sample, the percentage of soldiers who exited prematurely is small (13%) because attrition is most likely to occur early in the first term.

The bivariate correlations among the variables described above are presented in Table 1.67.

Table 1.67
Intercorrelations Among Major Turnover Analysis Variables

	JS	JP	Reenlist Intent	Attrit Intent	Reenlist	Attrit
Job Satisfaction	1.000					
Job Performance	.189	1.000				
Reenlist Intention	.268	.073	1.000			
Attrit Intention[a]	.196	.192	.099	1.000		
Reenlistment[b]	.158	.132	.369	.104	1.000	
Attrition[c]	-.089	-.263	-.060	-.264	-.263	1.000
Proportion Term Completed	-.081	.126	*-.024*	.067	*.026*	-.196

Note. n = 4,098 for all correlations involving attrition intention and 5,721 for all others. All correlations except those in italics are significant at p = .0001.

[a] Given the 96% intention-not-to-attrit base rate, the maximum possible r_{pb} is approximately .49; attrition coded 1 if soldier plans to complete term and 0 if not.

[b] Given the 32% reenlistment base rate, the maximum possible r_{pb} is about .78; reenlistment coded 1 if reenlisted and 0 if not.

[c] Given the 13% attrition base rate, the maximum possible value r_{pb} is approximately .60; attrition coded 1 if attrited and 0 if not.

The relative contribution of the three types of predictors (i.e., intention, job satisfaction, and job performance) was assessed by evaluating the improvement of model fit with the addition of each predictor to the prediction equation. Intention was the first predictor entered into the model because it is generally believed to be the most immediate precursor to turnover. The overall job satisfaction composite was entered next. Overall job performance was entered last because it was not expected to be related as strongly to reenlistment as satisfaction would be. Separate reenlistment analyses were conducted for soldiers in each enlistment term (2, 3, and 4 years).

The effect of adding each variable to the model was evaluated by conducting likelihood ratio tests. In addition, an index of the predictive strength of each model was constructed by calculating the point-biserial correlation between the predicted probability of reenlistment, as determined by each model, and actual reenlistment behavior.

For all three enlistment terms, all three models are statistically significant. However, the incremental contribution of job satisfaction and job performance over reenlistment intention was minimal (Table 1.68). Clearly, the most powerful and consistent predictor of reenlistment behavior in this sample is reenlistment intention.

Table 1.68
Point-Biserial Correlations Between Reenlistment and Predicted Probability of Reenlistment for Three Models

	Enlistment Term		
Alternative Models	2-Year	3-Year	4-Year
Intention	.548	.399 } b	.293
+ Satisfaction	.553 (.546)[a]	.402 (.400) } b	.293 (.287) } b
+ Performance	.553 (.544)	.402 (.399)	.301 (.293) } b
Reenlistment Rate[c]	14%	45%	39%
Maximum r_{pb}[d]	.60	.81	.80

Note. All models have chi-squares significant at $p < .05$. 2-year n = 492; 3-year n = 2,753; 4-year n = 1,322.

[a] Correlations in parentheses are adjusted for shrinkage using formula from Stein (1960).
[b] Differences in model chi-squares are significant at $p < .05$.
[c] Excluding soldiers ineligible for reenlistment.
[d] Estimated from Figure 4-5 in Nunnally (1967, p. 133).

Prediction of Attrition. Logistic regression was used to assess the incremental contribution of four variables to the prediction of avoidable, late-term attrition. These variables were (a) the proportion of term completed when measures were administered, (b) intention to attrit, (c) overall job satisfaction, and (d) overall job performance.

Event history analysis was also used to determine whether job satisfaction contributed significantly to model fit over the seven best predictors identified by the analyses summarized in the previous section (i.e., high school diploma status, ASVAB

Quantitative composite, and the following scores from the ABLE: Nondelinquency, Dominance, Physical Condition, Self-Esteem, and Social Desirability).

In both sets of analyses, the dependent variable was avoidable attrition (both voluntary and involuntary) as described in Knapp (1993). All other turnover-related outcomes (i.e., reenlistment, completion of first term, unavoidable attrition) were treated as censored observations in the event history analyses. In the logistic regression analyses, unavoidable attritions were dropped from the analysis sample.

Analyses were performed separately for soldiers with 3-year and 4-year enlistment terms. For the event history analyses, analyses were also conducted separately for combat and non-combat MOS and were stratified by MOS.

Analyses compared the overall fit of the following four nested models: (a) proportion of term completed when measures were administered; (b) proportion of term completed and intention to attrit; (c) proportion of term completed, intention, and satisfaction; and (d) proportion of term completed, intention, satisfaction, and performance. The proportion of term completed was considered a statistical control variable and was always entered first.

The four models are compared in Table 1.69. All models were statistically significant ($p<.05$) for both 3- and 4-year enlistees.

Results of the event history analyses showed that in all cases (3-year term, combat MOS; 4-year term, combat MOS; 3-year term, noncombat MOS; and 4-year term non-combat MOS), satisfaction significantly improved model fit.

However, unlike the previous analyses using an unrestricted sample, most of the pre-enlistment predictors were not significantly related to attrition. Whereas all seven of these predictors were related to attrition in earlier analyses, the ABLE Nondelinquency composite and job satisfaction were the only predictors consistently shown to be significantly related to attrition in the LVI sample where attrition could only be counted if it occurred after the data collection point for the LVI performance measures.

These findings support the notion that late-term attrition and the more common early-term attrition are not equivalent criteria. They also show that, despite this apparent nonequivalence, the Nondelinquency measure retains substantial predictive power throughout the first term of enlistment.

Summary

Satisfaction is related to contextual measures of performance, though not as strongly as anticipated. Both satisfaction and performance are related to reenlistment, but much of their predictive power is eroded when reenlistment intention is added into

Table 1.69
Point-Biserial Correlations Between Attrition and Predicted Probability of Attrition for Four Models[a]

Alternative Models	Enlistment Term: 3-Year			Enlistment Term: 4-Year		
Proportion of term completed	-.115		b	-.144		b
+ Intention	-.300	(-.297)		-.258	(-.250)	
+ Satisfaction	-.302	(-.298)	b	-.281	(-.271)	b
+ Performance	-.370	(-.366)		-.372	(-.363)	
Avoidable attrition rate	10%			20.5%		
Maximum r_{pb}[c]	.55			.70		

Note. All models have chi-squares significant at $p < .05$. 3-year n = 2,540; 4-year n = 1,197.

[a] Correlations in parentheses are adjusted for shrinkage using formula from Stein (1960).
[b] Differences in model chi-squares are significant at $p < .05$.
[c] Estimated from Figure 4-5 in Nunnally (1967, p. 133).

the prediction equation. The findings regarding attrition are similar to those associated with reenlistment, although performance retains appreciable predictive power even when intention to attrit is included. Satisfaction does predict attrition very well compared to pre-enlistment cognitive ability and temperament measures.

A final observation is that the length of enlistment term and the point in that term in which attitudes and performance are measured have a substantial impact on the ability to predict turnover-related behavior.

ORGANIZATION OF THE CURRENT REPORT

This is the final technical report that will be produced under the auspices of the Retaining the Career Force Project. It will attempt to portray the final stages of the data analyses and model building in a manner that integrates the project's work into a meaningful whole and speaks directly to the original objectives.

Chapter 2 reports the results of developing reliability estimates for all the Project A/Career Force performance factor criterion scores. These estimates will be used to correct the major validity coefficients produced by the project for attenuation. Chapter 2 includes a reexamination of the correlation of current performance with future performance after corrections for unreliability.

Chapter 3 reports an extensive effort to develop and evaluate empirical keys for the AVOICE. The criterion of principal interest was Core Technical Proficiency.

Chapter 4 describes the procedure, analyses, and results of the Project efforts to identify the most appropriate set of optimal prediction equations for predicting each major performance factor in each MOS. That is, when all the available information is used in the optimal fashion, what is the nature of the resulting prediction equations?

Chapter 5 is composed of a fairness analysis of the optimal equations using the Project's adaptation of the regression, or Cleary, model. Black/White and Male/Female differences are considered.

Chapter 6 describes the results of the full "roll-up" model analysis, which was first conceptualized in the original design for Project A. The data from (a) ASVAB, (b) the Experimental Battery, (c) end-of-training performance, and (d) first-tour performance will be used to predict second-tour (LVII) performance within a hierarchical model. That is, the incremental validity obtained from each additional source of information, as it becomes available, will be evaluated.

Chapter 7 presents a complete and detailed analysis of the classification, or differential assignment, problem as it pertains to the Army context. It sets the stage for Chapter 8.

Chapter 8 reports on the development and evaluation of a new index for assessing gains from classification. Several alternative assignment methods and maximization functions are compared using empirical samples and are then used to evaluate the classification efficiency of the full Project A/Career Force prediction equations.

Chapter 9 attempts a succinct summary of all the Project A/Career Force Project findings organized around the original goals for the two projects. It also seeks to express the appreciation of the consortium in particular, and applied psychology in general, for the support of the Army Research Institute and the Army.

We hope this final report paints an appropriate picture of this very extraordinary research effort.

Chapter 2

ESTIMATING THE RELIABILITY OF THE SCORES ON PROJECT A/CAREER FORCE PERFORMANCE CRITERION FACTORS

Doug Reynolds, Anthony Bayless, and John P. Campbell

A major overall goal of personnel selection and classification research is to estimate population parameters, in the usual statistical sense of estimating a population value from sample data. For example, a population parameter of particular interest is the validity coefficient that would be obtained if the prediction equation developed on the sample was used to select all future applicants from the same population of applicants. The sample value is of interest only in terms of its properties as an efficient and unbiased estimate of a population value.

In personnel research, there are at least two major potential sources of bias in sample estimates of the population validity. First, restriction of range in the research sample, as compared to the decision (applicant) sample, acts to bias the validity estimate downward. Second, if the sample data are used to develop differential weights for multiple predictors (e.g., multiple regression) and the population estimate (e.g., of the multiple correlation coefficient) is computed on the same data, then the sample estimate is biased upward because of fortuitous fitting of error as valid variance. For all Project A/Career Force analyses done to date, the sample values have been corrected for these two sources of bias. That is, the bias in the sample estimator was reduced by using the multivariate correction for range restriction and the Rozeboom correction for "shrinkage."

It also seems a reasonable goal to estimate the validity of the predictor battery for predicting true scores on performance. That is, the criterion of real interest is a performance measure that is not attenuated by unreliability. If different methods are used to measure the same performance factor, the estimate of validity would differ across methods simply because of differences in reliability. To account for these artifactual differences in validity estimates and to provide an estimate of a battery's validity for predicting true scores on a performance dimension, the sample-based estimates can be corrected for attenuation.

ESTIMATING CRITERION RELIABILITIES

Corrections for attenuation are an accepted procedure for removing the downward bias in the population estimates that is caused by criterion unreliability. Consequently, one of the project's final analysis tasks was to develop reliability estimates for each of the performance factors used as criterion measures in each of the principal research samples. Once the reliability estimates were available, the final corrections to the major validity estimates were made.

Research Samples

The research samples from the following four data collections were used for the reliability estimation analyses: the Concurrent Validation (CVI), the Longitudinal

Validation End-of-Training (LVT), the Longitudinal Performance Measurement (LVI), and the Longitudinal Validation Second-Tour (LVII). The data collection procedures for each sample have been described in detail in previous reports (e.g., see Campbell & Zook, 1990). The results reported in this chapter are limited to the Batch A MOS.

Each research sample of individuals had been selected from the two representative sets of MOS--Batch A and Batch Z. A detailed description of the MOS selection process is contained in previous reports (e.g., Campbell, 1987b). The list of MOS by batch was shown in Figure 1.1. Depending on the data collection, Batch A soldiers received the hands-on, job knowledge, and/or the school knowledge tests, all rating scales, and the administrative measures. On the other hand, the Batch Z soldiers received only the school knowledge test, the Army-wide rating scales, and the administrative measures, regardless of data collection. See Table 1.4 for detail about which criterion measures were administered to which MOS, for which samples.

Reliability Computation

Reliability estimates were calculated for the CVI, LV-EOT, LVI, and LVII criterion composites to correct criterion-related validities for attenuation. All composites in these samples have the same general form: Each is a linear composite of several standardized criterion measures. Thus, the reliability estimates all were derived using a modification of a formula for composite reliability for use with weighted standardized variables (cf. Nunnally, 1967).[1] This formulation of composite reliability requires the intercorrelation among the variables in a composite, the weight applied to each variable, and the reliability of each variable. All intercorrelatons were computed by MOS, within each of the data collection samples, among the variables that constitute each composite. Computation of the weights and reliabilities is described in more detail below.

Weights for each variable are a function of the manner by which the variables in each composite were combined. When standardized variables are simply added together to form a composite score, all components receive equal weight. When components are combined into subcomposites before being added together to create a larger composite, components are differentially weighted. For example, if three standardized variables are added to form a composite, each would receive equal weight (e.g., .33). If two of the three variables were added together, restandardized, and added to the third variable, the sum of the weights of the first two variables would equal the weight of the third

[1] Composite reliability was derived using the following equation:

$$r_{yy} = 1 - \frac{\sum w_j^2 - \sum w_j^2 r_{xx}}{\sum w_j^2 + 2 \sum \sum w_j w_k r_{jk}}$$

where w_j = the weight for component j, w_k = the weight for component k, r_{xx} = the reliability estimates for each component, and r_{jk} = the intercorrelation between the total scores on the components.

(e.g., .25, .25, .50). The weights for each criterion variable were applied in accordance with the manner by which the variables were combined in each composite under consideration.

Generally, reliabilities were computed at the MOS level for MOS-specific measures and across MOS for Army-wide measures. The Army-wide rating scales were the only exception; rating-scale based reliabilities were adjusted to account for differences in the average number of raters in each MOS. The methods for estimating the reliability of the various project criterion measures are described in the following sections.

School Knowledge, Job Knowledge, and Hands-on Tests

The reliability analyses for the school knowledge, job knowledge, and hands-on tests were conducted separately for each criterion composite and each MOS. Split-half reliability estimates were obtained for the job knowledge and school knowledge criterion measures, using an odd-even split method. Estimates of the reliability of equivalent forms were derived for the hands-on criterion measure by having a subject matter expert (who was familiar with the various MOS and their tasks) separate "equivalent" tasks into two groups. Scores for each criterion measure were derived by summing the constituent items or tasks of each half or equivalent form and correlating them within criterion measure to produce split-half reliability estimates. These estimates were corrected using the Spearman-Brown prophecy formula.

For those MOS that required tracked tests (i.e., different forms of the tests were necessary for some MOS because different types of equipment were used within the MOS), corrected split-half reliability estimates were derived for each of the tracks. To obtain a single reliability estimate for the tracked MOS, the weighted average of the corrected reliability estimates across tracks was computed.

In all samples, the school knowledge, job knowledge, and/or hands-on tests were used to form the Core Technical Proficiency and General Soldiering Proficiency criterion composites.

Rating Scales

Three types of rating scales were used across the four samples examined here: Army-wide scales, combat prediction scales, and MOS-specific scales. K-rater reliability estimates for each MOS are available in earlier project reports for all MOS-specific rating measures; those estimates were used here.

Reliability estimates for the Army-wide ratings were computed differently depending on how the rating scale scores were constructed in each sample. In the LV-EOT sample, only peer ratings were used in computing the composite scores. Similarly, in the LVII sample, only supervisor ratings were used. In these samples, the single-rater intraclass reliability estimate computed across all cases and MOS in the sample was adjusted by the average number of peer (LV-EOT) or supervisor (LVII) raters within each MOS.

In the CVI and LVI samples, an average of the peer and supervisor ratings was used to develop the criterion composites. For these samples, rating reliability was estimated by using an intraclass correlation analogous to Cronbach's α, where the reliability of the combined ratings is a function of the variance of the individual components (i.e., the peer or the supervisor ratings) and the variance of the combined ratings.

All of the required variances were computed for each rating scale basic score within MOS to produce an estimate of the reliability. The reliability estimates for the combat-prediction scales used in the CVI sample were also developed using this procedure. Rating-scale basic scores were used in composites involving effort, leadership, personal discipline, and physical fitness in each of the data collection samples.

Role Plays

The LVII data collection included three role plays. The scores resulting from these measures were added together to form a role-play basic score. The reliability of the role-play basic score was estimated as an unweighted linear composite (Nunnally, 1967; equation 7-11). This formulation requires the variances of the role-play scores, the variance of the total score, and the reliability of each measure. The required variances were computed, across MOS, with LVII data. However, because of the very small number of soldiers who were shadow-scored in the LVII data collection, the reliabilities for the individual role plays could not be computed for LVII data. Instead, single-rater reliability estimates computed for the CVII role plays were used; these reliabilities were reported in an earlier project report (Campbell & Zook, 1990). The resulting reliability estimate was .791 for the LVII role-play scores. These scores were used as one component of the LVII Leadership composite.

Administrative Measures

A number of administrative measures were used in the four data collections examined here. These included the Total Awards and Letters score, the Articles 15/Flag actions score, the Promotion Rate Deviation score, and the Physical Readiness score. Each of these scores was estimated to be .90 across all MOS. This "arm-chair" estimate was based on the small but probable error that may result from soldiers incorrectly remembering, or purposefully distorting, their self-reported administrative data. Administrative measures were used in composites involving effort, leadership, personal discipline, and physical fitness in each of the data collection samples.

Situational Judgment Test

Prior project research has estimated the internal-consistency reliability of the SJT at .81 across all MOS (Campbell & Zook, 1994a). This estimate was used as one component of the LVII Leadership composite.

Results

Criterion composite reliabilities for Batch A MOS are reported in Tables 2.1 through 2.4 for each data collection. Median reliabilities across MOS are shown in Table 2.5. In general, the reliabilities of the factor composites are quite high. There are several reasons for this result. First, the individual components had gone through a lengthy development process that attempted to maximize their relevant variance. Second, the data collections themselves had been carried out as carefully as possible. Third, each criterion score is a composite of a number of components. Finally, when ratings served as component measures, multiple raters were used. In fact, the reliabilities of the factors that are based largely on ratings measures are as high as, or higher than, the factor scores based on the hands-on and/or knowledge tests.

The reliabilities of the "will-do" factors for the training performance factors tend to be somewhat lower than for the "can-do" factors because the number of scales in each composite is smaller.

Table 2.1
Reliabilities for CVI Batch A MOS Criterion Composites

	Composite Reliability				
MOS	**CTP**	**GSP**	**ELS**	**MPD**	**PFB**
11B	.905	N/A	.857	.752	.824
13B	.905	.871	.846	.755	.824
19E	.822	.869	.827	.762	.811
31C	.904	.872	.836	.816	.853
63B	.898	.770	.870	.801	.842
64C[a]	.834	.872	.849	.803	.802
71L	.912	.798	.879	.817	.858
91A	.896	.778	.872	.820	.855
95B	.704	.857	.872	.786	.860

Note: CTP = Core Technical Proficiency
GSP = General Soldiering Proficiency
ELS = Effort and Leadership
MPD = Maintaining Personal Discipline
PFB = Physical Fitness/Military Bearing
N/A = Not Applicable, CTP = GSP for 11B.

[a] Subsequently MOS 88M.

TRUE SCORE CORRELATIONS OF PAST PERFORMANCE WITH FUTURE PERFORMANCE

The first application of the correction for attenuation was to the correlations of performance with performance. The annual report for the fourth year of the Career Force Project (Campbell & Zook, 1994c) reported the correlations between training performance and subsequent first-tour performance, between first-tour performance and

Table 2.2
Reliabilities for LV-EOT Batch A MOS Criterion Composites

	Composite Reliability					
MOS	Tech	Basic	ETS	MPD	PFB	Lead
11B	.870	N/A	.649	.678	.678	.628
13B	.905	.687	.667	.696	.696	.647
19E	.821	.840	.662	.691	.691	.642
19K	.888	.878	.667	.696	.696	.647
31C	.925	.796	.660	.688	.688	.639
63B	.919	.794	.522	.555	.555	.499
71L	.829	.634	.630	.660	.660	.608
88M	.901	N/A	.535	.567	.567	.512
91A	.901	.389	.665	.694	.694	.645
95B	.819	.831	.711	.734	.734	.692

Note: Tech = Technical Knowledge Score
Basic = Basic Knowledge Score
ETS = Effort and Technical Skill
MPD = Maintaining Personal Discipline
PFB = Physical Fitness/Military Bearing
Lead = Leadership Potential
N/A = Not Applicable, CTP = GSP for 11B, 88M

Table 2.3
Reliabilities for LVI Batch A MOS Criterion Composites

	Composite Reliability				
MOS	CTP	GSP	ELS	MPD	PFB
11B	.846	N/A	.864	.811	.822
13B	.827	.833	.867	.816	.832
19E	.720	.779	.831	.823	.830
19K	.600	.763	.843	.792	.820
31C	.878	.801	.824	.751	.785
63B	.776	.683	.838	.808	.784
71L	.823	.709	.789	.692	.825
88M	.751	.829	.851	.796	.814
91A	.852	.774	.867	.816	.843
95B	.508	.770	.881	.840	.869

Note: CTP = Core Technical Proficiency
GSP = General Soldiering Proficiency
ELS = Effort and Leadership
MPD = Maintaining Personal Discipline
PFB = Physical Fitness/Military Bearing
N/A = Not Applicable, CTP = GSP for 11B

Table 2.4
Reliabilities for LVII Batch A MOS Criterion Composites

	Composite Reliability					
MOS	CTP	GSP	AE	PD	PFB	LDR
11B	.764	N/A	.867	.800	.829	.875
13B	.880	.706	.860	.810	.825	.869
19K	.678	.721	.859	.786	.797	.854
63B	.575	.605	.853	.798	.838	.857
71L	.765	.731	.856	.796	.832	.843
88M	.460	.809	.834	.780	.827	.842
91A	.721	.886	.856	.776	.828	.856
95B	.690	.782	.858	.798	.832	.872

Note: CTP = Core Technical Proficiency
GSP = General Soldiering Proficiency
AE = Achievement and Effort
N/A = Not Applicable, CTP = GSP for 11B
PD = Personal Discipline
PFB = Physical Fitness/Military Bearing
LDR = Leadership

Table 2.5
Median Reliabilities (Across Batch A MOS) for the LVT (EOT), LVI, and LVII Performance Factor Scores

LVT (EOT)		LVI		LVII	
Factor	r_{xx}	Factor	r_{xx}	Factor	r_{xx}
Tech	.895	CTP	.800	CTP	.706
Basic	.795	GSP	.774	GSP	.731
ETS	.661	ELS	.847	AE	.857
MPD	.685	MPD	.810	PD	.797
PFB	.670	PFB	.824	PFB	.829
LEAD	.641			LDR	.857

Note: Tech = Technical Knowledge Score
Basic = Basic Knowledge Score
ETS = Effort and Technical Skill
MPD = Maintaining Personal Discipline
PFB = Physical Fitness/Military Bearing
LEAD = Leadership Potential
CTP = Core Technical Proficiency
GSP = General Soldiering Proficiency
ELS = Effort and Leadership
AE = Achievement and Effort
PD = Personal Discipline
LDR = Leadership

second-tour performance, and between training performance and second-tour performance. That is, the performance factor scores obtained at one point in time were correlated with performance factor scores obtained at a later point in time in a true longitudinal design. The performance scores that were used were the factor scores produced by the performance modeling analysis for performance at the end of training (LVT), first-tour performance (LVI), and second-tour performance (LVII). The composition of each of the three sets of scores is shown in Figures 2.1, 2.2, and 2.3.

In general, there were substantial correlations of current performance at one stage with performance at a subsequent stage. Also, the pattern of the correlations showed considerable convergent and divergent validity across performance factors, even for those factors measured principally by ratings.

The previously reported LVT x LVI, LVI x LVII, and LVT x LVII intercorrelations after they have been corrected for attenuation are shown in Tables 2.6, 2.7, and 2.8. Three correlations are shown for each relationship. The top figure is the mean correlation across MOS corrected for restriction of range (using the training sample as the population) but not for attenuation. These values were first corrected for range restriction within MOS and then averaged (weighted across MOS). The first value in the parentheses is this same correlation after correction for unreliability in the measure of "future" performance, or the criterion variable when the context is the prediction of future performance from past performance. The second value within the parentheses is the value of the mean intercorrelation after correction for unreliability in both the measure of "current" performance and the measure of future performance. It is an estimate of the correlation between the two true scores.

The reliability estimates used to correct the upper value were the median values (shown in Table 2.5) of the individual MOS reliabilities. The mean values across MOS were slightly lower and thus less conservative than the median.

In general, training performance is a strong predictor of performance during the first tour of duty. For example, the single-scale peer rating of leadership potential obtained at the end of training has a correlation of .46 with the single-scale rating of NCO potential obtained during the first tour, when the NCO potential rating is corrected for attentuation. The correlation between the true scores is .58. Correlations of first-tour performance with second-tour performance are even higher, and provide strong evidence for using measures of first-tour performance as a basis for promotion, or for the reenlistment decision. The true score correlation between the first-tour single-scale rating of NCO potential and the second-tour single-scale rating of overall effectiveness is .68.

In subsequent chapters of this report the criterion reliability estimates will be used to examine the "corrected" coefficients for additional relationships of special interest. For example, what happens when all available predictor information is used in an optimal fashion to predict subsequent performance and the sample estimate is fully corrected for both restriction of range and criterion unreliability?

SCORES BASED ON TRAINING ACHIEVEMENT TESTS

1) Basic Knowledge Score

- Items measuring knowledge requirements common to all MOS.

2) Technical Knowledge Score

- Items measuring technical knowledge requirements specific to each MOS.

SCORES BASED ON RATING SCALES (PEER RATINGS)

3) Effort and Technical Skill (ETS) Score

- Degree of effective acquisition of technical knowledge/skill*
- Degree to which individual demonstrates extra effort

4) Leadership Potential (LEAD) Score

- Degree of expected leadership effectiveness

5) Maintaining Personal Discipline (MPD) Score

- Degree to which individual adheres to regulations and orders
- Degree to which individual practices effective self-control

6) Physical Fitness and Military Bearing (PFB) Score

- Degree to which individual maintains proper military appearance
- Degree to which individual maintains military standards of physical fitness

* Each of these describes a behavioral summary rating scale that was constructed to parallel the analogous rating scale developed for the assessment of first-tour performance (i.e., parallel to CVI and LVI).

Figure 2.1. Six EOT performance factor scores based on measures of training performance obtained at the end of basic and technical training.

1) Core Technical Proficiency (CTP)
   - Hands-on Test Score - MOS-Specific Tasks
   - Job Knowledge Test Score - MOS-Specific Tasks

2) General Soldiering Proficiency (GSP)
   - Hands-on Test Score - Common Tasks
   - Job Knowledge Test Score - Common Tasks

3) Effort and Leadership (ELS)
   - Administrative Index - Number of Awards and Certificates
   - Army-Wide BARS Overall Effectiveness Rating Scale
   - Army-Wide BARS Effort/Leadership Ratings Factor
   - Average of MOS BARS Ratings Scales

4) Maintaining Personal Discipline (MPD)
   - Administrative Index - Number of Articles 15 and Flag Actions
   - Administrative Index - Promotion Grade Deviation Score
   - Army-Wide BARS Personal Discipline Ratings Factor

5) Physical Fitness and Military Bearing (PFB)
   - Administrative Index - Physical Readiness Score
   - Army-Wide BARS Fitness/Bearing Ratings Factor

Figure 2.2. Five LVI first-tour performance factor scores and the basic criterion scores that define them, as obtained from the first-tour performance measures.

1) Core Technical Proficiency (CTP)
   - Job-Specific Hands-On Test Score
   - Job-Specific Knowledge Test Score

2) General Soldiering Proficiency (GSP)
   - General Hands-On Test Score
   - General Job Knowledge Test Score

3) Effort and Achievement (EA)
   - Administrative Index: Number of Awards
   - Army-Wide BARS Technical Skill/Effort Ratings Factor
   - Overall Effectiveness Rating
   - Average of MOS BARS Ratings

4) Leadership (LEAD)
   - Administrative Index: Promotion Rate
   - Army-Wide BARS Leading/Supervisory Ratings Score
   - Discipline Role Play: Structure Score
   - Discipline Role Play: Communication Score
   - Discipline Role Play: Interpersonal Skill Score
   - Counseling Role Play: Diagnosis/Prescription Score
   - Counseling Role Play: Communication/Interpersonal Skills Score
   - Training Role Play: Structure Score
   - Training Role Play: Motivation Maintenance Score
   - Situational Judgment Test: Total Score

5) Maintaining Personal Discipline (MPD)
   - Administrative Index: Number of Disciplinary Actions (reversed score)
   - Army-Wide BARS Discipline Ratings Factor

6) Physical Fitness/Military Bearing (PFB)
   - Administrative Index: Physical Readiness Score
   - Army-Wide BARS Fitness/Bearing Ratings Factor

Figure 2.3. Six LVII performance factor scores for second-tour NCO performance and the basic criterion scores that define them, as obtained from the second-tour performance measures.

Table 2.6
Zero-Order Correlations of Training Performance (EOT) Variables With First-Tour Job Performance (LVI) Variables: Weighted Average Across MOS

	EOT:TECH	EOT:BASC	EOT:ETS	EOT:MPD	EOT:PFB	EOT:LEAD
LVI: Core Technical Proficiency (CTP)	**.48** (.54/.57)	.38 (.42/.45)	.22 (.25/.26)	.15 (.17/.18)	.05 (.06/.06)	.18 (.20/.21)
LVI: General Soldiering Proficiency (GSP)	.49 (.56/.62)	**.45** (.51/.57)	.23 (.26/.29)	.17 (.19/.22)	.04 (.05/.05)	.16 (.18/.20)
LVI: Effort and Leadership (ELS)	.21 (.23/.28)	.17 (.18/.23)	**.35** (.38/.47)	.25 (.27/.33)	.28 (.30/.37)	.35 (.38/.47)
LVI: Maintain Personal Discipline (MPD)	.17 (.19/.23)	.14 (.16/.19)	.31 (.34/.42)	**.36** (.40/.48)	.21 (.23/.28)	.27 (.30/.36)
LVI: Physical Fitness and Bearing (PFB)	-.01 (-.01/-.01)	-.02 (-.02/-.03)	.26 (.29/.34)	.13 (.14/.17)	**.44** (.48/.58)	.31 (.34/.41)
LVI: NCO Potential (NCOP)	.18 (.23/.25)	.16 (.21/.23)	.35 (.45/.56)	.26 (.34/.41)	.29 (.37/.45)	.36 (.46/.58)

Note. Total pairwise Ns range from 3,633 - 3,908. Corrected for range restriction. Correlations between matching variables are in bold. Leftmost coefficents in parentheses are corrected for attenuation in the future criterion. Rightmost coefficients in parentheses are corrected for attenuation in both criteria.

Labels:
EOT: Technical Total Score (TECH)
EOT: Basic Total Score (BASC)
EOT: Effort and Technical Skill (ETS)
EOT: Maintain Personal Discipline (MPD)
EOT: Physical Fitness and Bearing (PFB)
EOT: Leadership Potential (LEAD)

Table 2.7
Zero-Order Correlations of First-Tour Job Performance (LVI) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS

	LVI:CTP	LVI:GSP	LVI:ELS	LVI:MPD	LVI:PFB	LVI:NCOP
LVII: Core Technical Proficiency (CTP)	.44 (.52/.59)	.41 (.49/.55)	.25 (.30/.33)	.08 (.10/.11)	.02 (.02/.03)	.22 (.26/.29)
LVII: General Soldiering Proficiency (GSP)	.51 (.60/.68)	**.57** (.67/.76)	.22 (.26/.29)	.09 (.11/.12)	-.01 (-.01/-.01)	.19 (.22/.25)
LVII: Effort and Achievement (EA)	.10 (.11/.12)	.17 (.18/.20)	**.45** (.49/.53)	.28 (.30/.33)	.32 (.35/.38)	.43 (.46/.50)
LVII: Leadership (LEAD)	.36 (.39/.42)	.41 (.44/.47)	**.38** (.41/.45)	.27 (.29/.32)	.17 (.18/.20)	.41 (.44/.48)
LVII: Maintain Personal Discipline (MPD)	-.04 (-.04/-.05)	.04 (.04/.05)	.12 (.13/.15)	**.26** (.29/.32)	.17 (.19/.21)	.16 (.18/.20)
LVII: Physical Fitness and Bearing (PFB)	-.03 (-.03/-.04)	-.01 (-.01/-.01)	.22 (.24/.27)	.14 (.15/.17)	**.46** (.51/.56)	.30 (.33/.36)
LVII: Rating of Overall Effectiveness (EFFR)	.11 (.14/.16)	.15 (.19/.22)	.35 (.45/.49)	.25 (.32/.36)	.31 (.40/.44)	.41 (.53/.68)

Note. Total pairwise Ns range from 333 - 413. Corrected for range restriction. Correlations between matching variables are in bold. Leftmost coefficents in parentheses are corrected for attenuation in the future criterion. Rightmost coefficients in parentheses are corrected for attenuation in both criteria.

Labels:
LVI: Core Technical Proficiency (CTP)
LVI: General Soldiering Proficiency (GSP)
LVI: Effort and Leadership (ELS)
LVI: Maintain Personal Discipline (MPD)
LVI: Physical Fitness and Bearing (PFB)
LVI: NCO Potential (NCOP)

Table 2.8
Zero-Order Correlations of Training Performance (EOT) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS

	EOT:TECH	EOT:BASC	EOT:ETS	EOT:MPD	EOT:PFB	EOT:LEAD
LVII: Core Technical Proficiency (CTP)	**.48** (.57/.60)	.41 (.49/.52)	.22 (.26/.28)	.15 (.18/.19)	.08 (.10/.10)	.17 (.20/.21)
LVII: General Soldiering Proficiency (GSP)	.49 (.57/.64)	**.43** (.50/.56)	.19 (.22/.25)	.11 (.13/.14)	.06 (.07/.08)	.11 (.13/.14)
LVII: Effort and Achievement (EA)	.10 (.11/.13)	.15 (.16/.20)	**.25** (.27/.33)	.17 (.18/.23)	.19 (.21/.25)	.24 (.26/.32)
LVII: Leadership (LEAD)	.32 (.35/.43)	.39 (.42/.53)	.29 (.31/.39)	.19 (.21/.26)	.15 (.16/.20)	**.25** (.27/.34)
LVII: Maintain Personal Discipline (MPD)	.08 (.09/.11)	.09 (.10/.12)	.21 (.24/.28)	**.26** (.29/.35)	.16 (.18/.22)	.21 (.24/.28)
LVII: Physical Fitness and Bearing (PFB)	-.05 (-.05/-.07)	-.01 (-.01/-.01)	.12 (.13/.16)	.07 (.08/.09)	**.32** (.35/.42)	.21 (.23/.28)
LVII: Rating of Overall Effectiveness (EFFR)	.11 (.14/.15)	.16 (.21/.23)	.24 (.31/.38)	.18 (.23/.28)	.17 (.22/.26)	.21 (.27/.35)

Note. Total pairwise Ns range from 333 - 413. Corrected for range restriction. Correlations between matching variables are in bold. Leftmost coefficents in parentheses are corrected for attenuation in the future criterion. Rightmost coefficients in parentheses are corrected for attenuation in both criteria.

Labels:
EOT: Technical Total Score (TECH)
EOT: Basic Total Score (BASC)
EOT: Effort and Technical Skill (ETS)
EOT: Maintain Personal Discipline (MPD)
EOT: Physical Fitness and Bearing (PFB)
EOT: Leadership Potential (LEAD)

Chapter 3
DEVELOPMENT AND EVALUATION OF AVOICE EMPIRICAL KEYS, SCALES, AND COMPOSITES

Cheryl Paullin, Ken Bruskiewicz, Mary Ann Hanson, Kristi Logan, and Mark Fellows

This chapter describes several alternative procedures for scoring the Army Vocational Interest Career Examination (AVOICE), a vocational interest inventory developed as part of Project A and Career Force. In previous analyses, the AVOICE has been found to be a valid predictor of several important army criteria (Campbell & Zook, 1990). However, the procedures that have been used to score the AVOICE differ somewhat from those traditionally used to score interest inventories (e.g., Strong, 1943). The research described here was conducted to determine whether alternative scoring procedures can be developed that have potential for increasing the validity of the AVOICE or otherwise enhancing the usefulness of this inventory.

BACKGROUND

AVOICE Development and Scoring

The AVOICE is designed to measure a wide variety of interests relevant to jobs in the Army. It was developed primarily using a rational scale construction strategy with some use of internal (e.g., factor analytic) scale development techniques. The starting point was an interest inventory developed by the Air Force, called the Vocational Interest Career Examination (VOICE; Alley & Matthews, 1982). The VOICE initial item pool was developed rationally to cover interest constructs relevant for Air Force enlisted personnel. These items were grouped into scales based on similarity in content.

The initial VOICE was revised and refined based on internal (e.g., factor analysis) and rational considerations. Form B contains 300 items that describe a variety of occupational titles, work tasks, leisure time activities, and desired learning experiences. Respondents are asked whether they like, dislike, or feel indifferent toward each. The items are grouped into 18 homogeneous basic interest scales. The VOICE was included in the first large-scale Project A predictor data collection (the Preliminary Battery).

Based on the results of this administration, the VOICE was revised: Some items were dropped, new items were added, and the response format was changed from a 3-point to a 5-point scale. This initial version of the AVOICE (the Pilot Trial Battery version) was administered to the Project A field test sample. The inventory was revised on the basis of field test data, administered to the Concurrent Validation (CV) sample, revised further, and finally administered to the Longitudinal Validation (LV) sample. More detailed discussion of VOICE development can be found in Hough, Barge, and Kamp (1987); L. M. Hough, McCloy, Ashworth, and M. M. Hough (1987); and Hough, McGue, Houston, and Pulakos (1987).

The current AVOICE (the Experimental Battery version) contains 182 items. The items are grouped into 22 scales that were developed and refined based on factor analyses, item-total correlations, and rational considerations (Hough, McCloy, et al., 1987). The AVOICE was initially intended to measure all six vocational interest constructs in Holland's (1966) hexagonal model of interests, with emphasis on the interest constructs most relevant to the Army. However, the current version does not include any scales from Holland's Enterprising theme. It is heavily weighted toward the Realistic theme, reflecting the fact that much Army work is of a realistic nature. The current AVOICE scales are listed in Table 3.1, organized according to Holland's themes.

AVOICE respondents rate a variety of jobs, work tasks, spare time activities, and desired learning experiences, using a scale ranging from "like very much" to "dislike very much." These item-level scores were summed to create 22 basic interest scores which, in turn (based on principal components analysis), have been grouped into eight summary composites (Peterson et al., 1990). VOICE validity has been assessed in past Project A/Career Force research, with multiple correlations computed between the eight AVOICE composites and a variety of performance criteria (Campbell & Zook, 1990, 1991).

Empirical Scoring Procedures

A scale construction strategy is a systematic procedure for grouping and keying item responses to form composite scores. The empirical approach involves selecting and/or weighting items according to evidence that the items differentiate between persons who score at different levels on a criterion variable. Because this approach maximizes the relationship between the scoring procedure and criterion variable scores in the development sample, empirical scoring procedures must be cross-validated to assess how much the procedures capitalize on chance and to ensure that the validity results hold up in an independent sample.

One important advantage of empirical scoring procedures is that they allow non-obvious relationships between item responses and criterion data to emerge. Empirical keys do not depend on the accuracy of any particular theory or set of hypotheses. Ideally, items in an empirical key will correlate maximally with the criterion variable but minimally with each other. Thus, internal consistency is not expected to be high, and a common theme or factor is not necessarily expected among the items. This can make it difficult to interpret what empirical keys measure and impossible to provide an a priori explanation of why persons score in a certain manner.

Table 3.1
Content of the AVOICE

Holland Theme	Scale Name	Number of Items
Realistic	Mechanics	10
	Heavy Construction	13
	Electronics	12
	Electronic Communications	6
	Drafting	6
	Law Enforcement	8
	Audiographics	5
	Rugged Individualism	16
	Combat	10
	Firearms Enthusiast	7
	Vehicle Operator	6
	Fire Protection	6
Investigative	Medical Services	12
	Science/Chemical	6
	Mathematics	3
	Computers	4
Artistic	Aesthetics	5
Social	Leadership/Guidance	12
Conventional	Clerical/Administrative	14
	Food Service Professional	8
	Food Service Employee	6
	Warehousing/Shipping	7

Past Research on the Empirical Approach

Many researchers have developed empirical keys to predict organizationally relevant criteria. Much of this work has used life experiences (i.e., biodata), vocational interests, and/or personality traits to predict criterion variables such as training or job performance, sales, and turnover. Owens (1976) identified 72 studies in which empirically derived biodata keys were used to predict organizationally relevant criterion variables. He reports an average validity of .35 for predicting sales success, .48 for performance in clerical or office jobs, and .48 for high-level talent or creativity.

Hough and Paullin (in press) reviewed 21 studies that directly compare the criterion-related validity of biodata scales developed using different construction strategies. The varied criterion measures included correct classification into clinical or medical diagnostic categories, self- or peer-reported standing on personality traits, and various measures of job or training performance. No scaling strategy appeared clearly superior to any other scaling strategy -- at least in terms of criterion-related validity.

Research is also available on the validity of empirically developed interest inventory scores. Much of this research has used the Strong Vocational Interest Blank (SVIB: currently called the Strong Interest Inventory). It has an impressive history of demonstrating both concurrent and predictive validity for a wide variety of criteria (Barge & Hough, 1988).

Strong obtained "good hits" in predicting occupational membership for nearly 80 percent of the 420 subjects in 5- and 10-year followup studies and "clean misses" for less

than 20 percent (Strong, 1943). Similar results have been reported by other researchers (e.g., see Bartling & Hood, 1980; Campbell, 1966). Interest scales have been shown to predict job satisfaction in a variety of jobs as well. Alley, Wilbourn, and Berberich (1976), using a predictive design for a sample of Air Force enlisted personnel, found significant multiple correlations between VOICE interest subscales and job satisfaction, ranging from a low of .20 (for administrative personnel) to a high of .57 (for mechanics). Researchers are also exploring the relationships between vocational interests and job performance. Gellatly, Paunonen, Meyer, Jackson, and Goffin (1991) found that vocational interests were significantly related to job performance in their sample of managers; the interest measures predicted variance in job performance that was not accounted for by cognitive ability.

In general, previous research shows that empirical scoring procedures can be valid predictors of important organizational criteria. There is reason to expect that empirical scoring procedures for the AVOICE may have appreciable validity, both to differentiate between soldiers in different occupations (i.e., MOS) and to predict organizationally relevant criteria such as performance or satisfaction.

Issues in Developing Empirical Scoring Procedures

Types of Scores

Numerous approaches can be taken to develop empirical scoring procedures. One basic distinction is the level of data used to develop the scoring procedures: response option-level, item-level, or scale-level data.

The result of empirical scoring procedures based on response option-level data will be referred to as *keys*. Weights are typically assigned to *each* response option within each item (although some options may be weighted zero). For example, a "middle" response option for a multiple-choice item could receive the largest weight.

The result of empirical scoring procedures based on item-level data will be referred to as *scales*. The raw data are item scores from the response options selected. The response scale metric typically increases monotonically and assumes a linear relationship between the item-level response and some underlying construct or criterion. For example, AVOICE scores range from 1 (like very much) to 5 (dislike very much).

Finally, empirically defined scores based on scale-level data will be referred to as *composites*. Selecting and weighting rational scales based on empirical relationships with a criterion variable is not typically thought of as an empirical scale development approach. However, the development of empirical composites is analogous to the development of empirical keys and scales. The scales within an empirical composite may be either unit-weighted or optimally weighted (e.g., using regression weights).

Empirically derived keys, scales, and composites vary in their potential to capitalize on error variance. Option-level empirical keys make use of the maximum amount of information available; this also means that empirical keying approaches have the greatest

potential to capitalize on error variance. Consequently, all other things being equal, empirical keying requires the largest sample size for development.

However, empirical keys may be more resistant to faking than scales or composites because keys are more likely to include weighting of non-obvious responses. Kluger, Reilly, and Russell (1991) found that a response-option level empirical key was more resistant to attempts to fake than an item-level key, perhaps because respondents instructed to fake tended to choose more extreme options. A faking strategy would affect item-level and scale-level scoring procedures much more than option-level keys.

Key/Scale/Composite Length

For internally constructed scales and most rationally constructed scales, longer scales are expected to have higher internal consistency reliabilities and consequently greater potential for validity. For empirical scoring procedures, the relationship between scale length and reliability is not as straightforward because they are not designed to be internally consistent. The length of empirical keys, scales, and composites also represents a potential confound when comparing the validity of empirically and rationally derived procedures. If empirical scoring procedures produce keys, scales, or composites that are consistently longer (or shorter) than scales or composites produced by rationally derived procedures, obtained differences in validity could be due, in part, to systematic differences in reliability.

Various rules of thumb can be used to determine how much data should be included in an empirical key, scale, or composite and how the data should be weighted. Mumford and Owens (1987) recommend excluding items failing to yield a difference in t-test, correlational, or chi-square analyses significant at the $p < .05$ or $p < .01$ level. However, if the scoring procedure is being developed in a large sample, this rule of thumb may lead to including information having only a weak relationship with the criterion variable. In these cases, Mumford and Owens recommend including only items that have a correlation of at least .10 or .15 with the criterion variable of interest. In general, the length of empirical keys, scales, and composites depends on the decision rules adopted during development.

Statistical Approaches to Empirical Scoring

Following is a brief description of four of the most common approaches to developing empirical keys.

(1) The vertical percent method (Stead & Shartle, 1940) involves three steps. First, contrasting groups are formed on the basis of some criterion variable. Often a natural dichotomy occurs (e.g., persons who attrite versus persons who stay in a job). When the variable is continuous, contrasting groups must be created; the middle of the criterion score distribution is often excluded (e.g., by contrasting the top 30% with the bottom 30%). Second, the percentage of persons choosing each response option in each criterion group is tabulated. Third, the difference between the percentages for the two

groups is calculated. The weight assigned to each response option depends on the size of this difference; the larger the difference, the greater the weight.

(2) The horizontal percent method (e.g., see Guion, 1965) also involves three steps. First, criterion groups are formed (just as for the vertical percent method). Second, the number of persons from the "more desirable" criterion group who chose each option is tabulated. Third, this number is divided by the total number of persons in both criterion groups. Option weights are created by dividing the resulting value by 10. All weights are positive and every option is included in the key.

(3) In the phi coefficient method (Lecznar, 1951), two criterion groups again are formed on the basis of some criterion variable. The correlation between the dichotomous response option variable and the dichotomous criterion group variable is the phi coefficient. Generally, a unit weight is assigned to response options that exhibit a statistically significant phi coefficient. Thus, key length depends both on the strength of the relationships and on the sample size. The direction of the weights corresponds to the direction of the phi coefficient. The point-biserial correlation can be substituted for the phi coefficient if the criterion variable is continuous.

(4) Unlike the preceding option-level approaches, the mean criterion method (Devlin, Abrahams, & Edwards, 1992) does not require creation of contrasting criterion groups. In this method, the weight assigned to each response option is the mean criterion score earned by all persons choosing the response option. All weights are positive and every option is included in the key.

Devlin et al. compared nine different weighting methods within the option-level empirical keying approach, including four variations of the vertical percent method, the horizontal percent method, the phi coefficient method, and the mean criterion method. They found that several variations of the vertical percent method yielded marginally, but consistently, higher validities than other option-weighting methods.

Available Project A/Career Force AVOICE Data

For two of the Project A/Career Force cohorts -- CV and LV -- data appropriate for the development of alternative, external (i.e., empirical) AVOICE scoring procedures are available. While soldiers of both cohorts were told that their AVOICE responses would be used for research only, the response set for these two groups may have differed. Because the CV soldiers had been in the Army for several years, they had no reason to suspect that their AVOICE responses could affect their Army careers. On the other hand, the LV respondents, who were in their reception processing, may have been somewhat concerned about potential consequences for their Army careers. Thus, the LV cohort may have been more motivated to respond in a socially desirable way.

In addition, AVOICE responses may systematically differ between these two cohorts because the LV cohort had virtually no Army experience whereas the CV cohort had been in the Army for several years. This Army experience may have changed some respondents' interests, and it certainly provided them with more information on which to

base their AVOICE responses (e.g., some items deal with operating armored vehicles). In contrast. AVOICE responses from the LV soldiers most closely approximate an applicant set.

The five first-tour performance composites are available as criterion data for both CV and LV cohorts: Core Technical Proficiency (CTP), General Soldiering Proficiency (GSP), Effort and Leadership (ELS), Personal Discipline (MPD), and Physical Fitness and Military Bearing (PFB). Also, for the LV cohort, the Army Job Satisfaction Questionnaire (AJSQ) was administered during their first tour of duty. The AJSQ provides several indices of job satisfaction as well as self-report of intent to reenlist.

A total of 9,349 soldiers from the CV cohort have both AVOICE data and first-tour performance data (the CVI sample), and about 10 percent of these soldiers are female. A total of 7,832 soldiers from the LV cohort have both AVOICE data and first-tour performance data (the LVI sample), and about 11 percent of these soldiers are female. Soldiers in each sample come from a variety of different MOS. The numbers of soldiers from each are shown for the CVI and LVI samples in Table 3.2. The numbers of male and female soldiers in each sample are also shown.

For many soldiers in the LV cohort (about 1,500), second-tour performance data were also collected. The measures used were similar to those used in the first tour, but were revised to reflect the somewhat higher performance requirements and the additional supervisory duties in second-tour jobs. Five performance composites identified for these second-tour soldiers are very similar to those in the model of first-tour performance; in addition, a sixth Leadership composite score was identified. Finally, attrition data are available for virtually the entire LV cohort (about 50,000 soldiers).

Sex Bias/Fairness Issues

The issue of sex bias is especially problematic in the area of interest measurement because researchers have consistently found systematic differences in interests expressed by males and females that appear at a very early age and continue throughout adulthood (Barge & Hough, 1988; Hansen, Collins, Swanson, & Fouad, 1993).

As expected, systematic differences have been found in the mean scores of males and females on the AVOICE rational scales and composites. Table 3.3 shows the effect sizes of the mean differences between male and female scores in the total LVI sample: effect size is the difference between male and female scores, divided by the pooled standard deviation (i.e., the standardized mean differences).

Table 3.2
Number of Soldiers in the Concurrent and Longitudinal Validation Cohorts With AVOICE and First-Tour Performance Data by MOS and by Sex

MOS	Concurrent			Longitudinal		
	Males	Females	Total	Males	Females	Total
Infantryman (11B)	694	0	694	810	0	810
Combat Engineer (12B)	703	0	703	592	0	592
Cannon Crewman (13B)	653	0	653	785	0	785
MANPADS Crewman (16S)	470	0	470	329	0	329
M1/M60 Armor Crewman (19E/K)	501	0	501	678	0	678
TOW/Dragon Repairman (27E)	141	6	147	40	3	43
Electronics Repairer (29E)	-	-	-	61	4	65
Single Channel Radio Operator (31C)	306	51	357	211	32	243
Carpentry/Masonry Specialist (51B)	107	1	108	94	0	94
Chemical Operations Specialist (54E)	423	10	433	319	27	346
Ammunition Specialist (55B)	271	18	289	129	17	146
Light Wheel Vehicle Mechanic (63B)	591	42	633	507	55	562
Motor Transport Operator (64C)[a]	617	61	678	215	75	290
Utility Helicopter Repairer (67N)	270	6	276	82	1	83
Administrative Specialist (71L)	227	278	505	84	268	352
Petroleum Supply Specialist (76W)	452	38	490	-	-	-
Unit Supply Specialist (76Y)	528	96	624	534	105	639
Medical Specialist (91A)	366	126	492	562	104	666
Food Specialist (94B)	497	110	607	609	87	696
Military Police (95B)	637	52	689	258	60	318
Intelligence Analyst (96B)	-	-	-	83	12	95
Total	8,454	895	9,349	6,982	850	7,832

[a] MOS 64C became MOS 88M in the Longitudinal Validation.

The difference between male and female mean scores is significant ($p < .01$) for all but four of the rational scales and all but one of the rational composites. Further, the direction of the effect sizes is consistent with stereotypical conceptions of sex differences. For example, females score higher than males on the Clerical/ Administrative, Aesthetic, Medical Services, and Fire Protection scales, and on the Administrative, Social, Audiovisual Arts, and Food Service composites. Males score higher than females on the Firearms Enthusiast, Combat, Heavy Construction, and Rugged Individualism scales, and on the Rugged/Outdoors, Structural/Machines, and Protective Services composites. In addition to highlighting the potential for subgroup

Table 3.3
AVOICE Rational Scales and Composites: Male/Female Effect Sizes for the Total LVI Sample[a]

	Effect Size[b]		Effect Size[b]
Scales		**Composites**	
Firearms Enthusiast	1.13	Rugged/Outdoors	1.15
Combat	0.91	Structural/Machines	0.86
Heavy Construction	0.87	Administrative	-0.47
Rugged Individualism	0.83	Social	-0.43
Clerical/Administrative	-0.78	Protective Services	0.34
Mechanics	0.69	Audiovisual Arts	-0.28
Aesthetics	-0.66	Food Service	-0.13
Electronics	0.63	Skilled Technical	-0.04
Medical Services	-0.47		
Fire Protection	-0.39		
Vehicle Operator	0.38		
Science/Chemical	0.23		
Leadership/Guidance	-0.21		
Drafting	0.21		
Food Service Professional	-0.20		
Law Enforcement	0.19		
Mathematics	-0.17		
Audiographics	-0.15		
Computers	-0.10		
Electronic Communications	-0.04		
Food Service Employee	-0.03		
Warehousing/Shipping	-0.03		

[a] Sample sizes for females range from 761-811; samples sizes for males range from 6,611 to 6,805.
[b] Effect sizes in this table are relative to the male subgroup (i.e., a positive effect size indicates that males score higher than females). Effect sizes larger than about .10 are significant at the $p < .01$ level.

differences, these large sex differences suggest that scoring procedures developed using a combined-sex or male sample may not be equally valid for both sexes.

There is much disagreement among researchers about how best to approach the issue of sex bias in measuring interests. Some have argued that separate scales should be developed for males and females because same-sex scales are better able to differentiate between occupational groups and general reference samples (e.g., Campbell & Hansen, 1981). Some have developed combined-sex scales but with separate norms for males and females (Kuder & Diamond, 1979); however, others have argued that separate sex norms result in lowered predictive accuracy (Gottfredson & Holland, 1975). Still another approach is to encourage both men and women to consider occupations that historically

approach is to encourage both men and women to consider occupations that historically have been dominated by the other sex. Primarily the controversy revolves around whether scaling or norming procedures should focus on increasing predictive efficiency or decreasing sex differences (Barge & Hough, 1988).

THE GENERAL PROCEDURE AND ANALYSIS DESIGN

As mentioned, the goal in the present research was to determine whether empirical scoring procedures have potential for improving AVOICE validity. Thus, efforts were made to assess a wide variety of different empirical scoring procedures and to focus on the criterion variables, scale construction techniques, and soldier samples which appeared to show the most promise for improving validity or utility.

Criterion Variables

Two aspects of job performance were included as criteria: the first-tour Core Technical Proficiency (CTP) factor and the second-tour Leadership (LDR) factor. In past research, the rationally derived AVOICE scales and composites have shown some promise for predicting CTP (Campbell & Zook, 1991; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990), and we wanted to determine whether an empirical approach to scale construction could enhance this validity. Leadership was included because it is an important aspect of job performance as soldiers move into their second tour of duty. In addition, performance is arguably more similar across MOS in the leadership and supervisory aspects of the job than in the technical aspects, so the entire LV cohort second-tour (LVII) sample could be used in developing and evaluating empirical scales to predict Leadership.

The first-tour Core Technical Proficiency criterion composite consists of scores on job-specific hands-on tasks and job knowledge test items covering tasks specific to each soldier's MOS (in contrast to general or common tasks). Overall score for the hands-on tasks is the average "percent-go" score across all tasks. Overall score for the job knowledge test is number correct. The CTP score is the unit-weighted sum of the two standardized overall scores.

The second-tour Leadership criterion composite is the unit-weighted sum of standardized scores from four performance measures targeted at the supervisory aspects of second-tour soldier jobs: the Situational Judgment Test (SJT), the Supervisory Simulation Exercises, the supervisory factor from the Army-wide performance rating scales, and promotion rate. In the SJT, soldiers are presented with supervisory situations and asked to identify the most and least effective response alternatives. The simulation exercises require those being tested to counsel and train several role-playing evaluators who then rate their performance. Behavior-based performance ratings were collected from the supervisors of LVII soldiers, and a composite was based on the ratings on the supervisory and leadership dimensions. Finally, self-reported promotion rate is included in the criterion.

Empirical scoring procedures were also developed to predict attrition, because vocational interests have at least a theoretical relationship with attrition. Attrition was defined using Knapp's (1993) turnover categorization scheme, which identifies soldiers in the LV cohort who attrited for what the military considers avoidable reasons (e.g., failure to meet minimal performance or behavioral criteria). Examination of the proportion of soldiers who attrited for avoidable reasons during each of the first 24 months of duty (after 24 months some soldiers have completed their tour of duty) showed that the percentage of soldiers attriting in each of the first 12 months is much larger than the percentages during subsequent months. Therefore, 12-month attrition was used as the criterion variable for developing empirical scoring procedures to predict attrition.

Finally, the empirical procedures were used to predict MOS membership, approximating procedures used in Strong's research on vocational interests (1943). Traditionally, job incumbents are used as the development sample when forming empirical scoring procedures to predict occupational membership. Therefore, the CV sample was used to represent job incumbents with several years of job experience.

Samples

Several factors entered into choosing MOS to include in our analyses. First, we wanted at least two MOS that are very similar and at least two that are very dissimilar; this would help to evaluate whether empirical scoring procedures developed in a similar MOS are more valid for a particular MOS than those developed in a dissimilar MOS. We also wanted at least one combat and at least one noncombat MOS. Finally, we wanted at least one MOS with large enough numbers of females to develop separate empirical scoring procedures based on females only.

Five MOS were chosen: Infantryman (11B), Cannon Crewman (13B), Light Wheel Vehicle Mechanic (63B), Administrative Specialist (71L) and Medical Specialist (91A). 11B and 13B are combat MOS; thus, they are arguably similar, and they avoid issues related to sex differences because they contain only males. The remaining three MOS are very different from one another and from the two combat MOS. Also, 91A and 71L each have a relatively large percentage of female soldiers.

A power analysis suggested that each sample should consist of at least 75 persons to have about a .60 chance of detecting a validity coefficient of about .25 (see Cohen & Cohen, 1975). The number of soldiers in some single-sex samples for the CTP and occupational membership criteria (e.g., MOS 91A females) was too small to split into development and cross-validation samples, so the entire sample was used as a development sample. Thus, we could not cross-validate these scales but we did transport them to other samples (e.g., the other cohort). Further, the 63B female sample was too small to create either a development or a cross-validation sample in either cohort for any criterion, so no empirical scoring procedures were developed.

Samples for Empirical CTP, Leadership, and Attrition Scoring Procedures

Both AVOICE data and first-tour CTP criterion data are available for large numbers of soldiers in both CV and LV and empirical scoring procedures were developed to predict CTP in both cohorts. Empirical scoring procedures for predicting CTP were developed within each of the five selected MOS, and for males and females separately where feasible.

The Leadership performance criterion composite is available for LV cohort members who were tested on performance during their second tour (i.e., the LVII sample), and these soldiers have AVOICE data as well. Empirical scoring procedures to predict Leadership were developed using the entire LVII sample (i.e., across all MOS).

Finally, AVOICE and attrition data are available for almost the entire LV cohort. Because interests were expected to show differential relationships across MOS, these analyses were conducted within MOS. Empirical scoring procedures to predict attrition were developed in the MOS 13B and 91A male samples only.

For each targeted criterion and sample, a development sample was created by randomly selecting about half of the relevant sample, with the remainder used as a cross-validation sample. Adequate cross-validation samples could not be formed for MOS 71L males in the LV cohort or for 91A females in the LV or CV cohort for predicting CTP.

In each sample, scoring procedures were developed using the development sample data and cross-validated by applying each procedure in the relevant cross-validation sample. This process provides a conservative estimate of the validity of the empirical scores. Table 3.4 shows development and cross-validation sample sizes for developing empirical scoring procedures designed to predict CTP, Leadership, and attrition.

Samples for Occupational Scoring Procedures

To develop scales to predict occupational (i.e., MOS) membership, we first randomly divided all male soldiers in the entire CVI sample into development and cross-validation samples. The female CVI soldiers were similarly divided. Most of the occupational scales were developed in the developmental half of these CV samples, but for two MOS sample size would not allow developing and cross-validating within the CV sample. For these MOS (91A females and 71L males), occupational scores were developed using the entire CV sample and then cross-validated only in the LV sample.

Soldiers were included in the target MOS samples only if they were in the top 90 percent of the performance distribution for Core Technical Proficiency, to ensure that soldiers being tested were performing their jobs with at least some degree of proficiency. Each target MOS has a corresponding general population, consisting of all soldiers remaining in the same-sex CV development sample after the target MOS soldiers have been removed (e.g., for the 91A male target MOS, the general population consisted of all

Table 3.4
Sample Sizes for Developing and Cross-Validating Empirical Scoring Procedures to Predict Core Technical Proficiency, Leadership, and Attrition

MOS	CV Cohort		LV Cohort	
	Development Sample	Cross-Validation Sample	Development Sample	Cross-Validation Sample
First-Tour CTP				
11B	248	243	393	365
13B	237	227	347	353
63B Males	229	213	243	240
71L Males	98	92	74	-- [a]
71L Females	123	114	122	113
91A Males	137	139	276	269
91A Females	116	-- [a]	99	-- [a]
Second-Tour Leadership				
All Batch A MOS	-- [b]	-- [b]	551	531
Attrition				
13B	-- [c]	-- [c]	2,361	2,406
91A Males	-- [c]	-- [c]	1,618	1,592

[a] Sample is too small to create a cross-validation sample.
[b] Leadership composite not available for this cohort.
[c] Attrition data not available for this cohort.

male soldiers in the CV sample except the 91As). Table 3.5 shows development and cross-validation sample sizes for setting up empirical occupational scoring procedures.

Combined-Sex Samples

A combined-sex development sample was generated for each MOS and for each criterion by pooling the two single-sex development samples already created for that MOS. Similarly, combined-sex cross-validation samples were formed for each MOS. Where the single-sex sample was too small to create a cross-validation sample, we randomly divided that single-sex sample in half and included half in the combined-sex development sample and the other half in the combined-sex cross-validation sample.

Table 3.5
Sample Sizes for Developing and Cross-Validating Occupational Keys, Scales, and Composites to Predict MOS Membership

	CV Cohort				LV Cohort (Cross-Validation Only)	
	Development Sample		Cross-Validation Sample			
MOS	Target MOS	General[a] Population	Target MOS	General[a] Population	Target MOS	General[a] Population
11B Males	336	3,880	358	3,880	791	5,996
13B Males	334	3,882	319	3,919	765	6,022
63B Males	289	3,927	302	3,936	495	6,292
71L Males	227	8,227	-- [b]	-- [b]	82	6,705
71L Females	148	313	130	304	255	554
91A Males	199	4,017	167	4,071	553	6,234
91A Females	126	769	-- [b]	-- [b]	102	707

[a] General population for each comparison consists of members of all other Batch A and Batch Z MOS in that sample.
[b] Sample is too small to create a cross-validation sample.

Screening

Screening of the AVOICE item-level data mirrored what has been done in previous analyses (Hough, McCloy, et al., 1987; Peterson et al., 1990). In both cohorts, if a soldier was missing more than 10 percent of the item-level AVOICE data or appeared to have responded randomly, that soldier's AVOICE data was set to missing. In each cohort, the missing data screen removed about 2 percent of the sample.

For the CV cohort, the Non-Random Response scale from the Assessment of Background and Life Experiences (ABLE; the biodata/personality inventory administered with the AVOICE) was used to screen the sample, on the assumption that soldiers who responded randomly on the ABLE probably also responded randomly on the AVOICE. Low scores on the scale indicate random responding, so data for a soldier who scored lower than six on the scale were dropped. About 8 percent of the sample was screened out for this reason (Peterson et al., 1990).

For the LV cohort, an Unlikely Responses scale made up of 12 AVOICE response options that are seldom endorsed was developed for the AVOICE. Soldiers who endorsed more than 5 of the 12 were assumed to be responding randomly. About 3 percent of the LV cohort data was eliminated based on these scale scores.

Data Analysis Procedures

Development of Empirical and Occupational Keys, Scales, and Composites

The initial tryout of the full range of empirical scoring procedures described previously was conducted for two criteria: (a) first-tour Core Technical Proficiency for soldiers in the 13B MOS, and (b) membership in the 91A MOS. These results were used to reduce the number of alternative procedures to a smaller set for the main comparisons using the remaining MOS and criteria.

Empirical and Occupational Keys. To develop the empirical keys to predict CTP, two criterion groups were identified within each MOS 13B development sample. The "high performance" group included soldiers whose CTP score fell in the top 30 percent of the 13B CTP distribution; the "low performance" group was made up of the bottom 30 percent.

We created vertical percent keys by computing the difference between the percentages of the two groups endorsing each response option and assigning weights to each option, using Strong's table of Net Weights. Two vertical percent keys were developed for each sample, one including only those items that showed at least a 5-point difference in response rates for high- and low-performance criterion groups and one including only items that showed at least a 10-point difference. Response options were assigned a positive weight if a larger percentage of the high-performance group endorsed the option and a negative weight if a larger percentage of the low-performance group endorsed the option.

We created the point-biserial empirical key by correlating each dichotomous response option score (did endorse/did not endorse) with the continuously scored CTP criterion measure. Response options with a point-biserial correlation at the $p < .05$ level were assigned a weight of +1 or -1 depending on the direction of the correlation. The remaining response options were assigned a weight of 0.

To develop the occupational keys for predicting MOS membership, we used a procedure very similar to Strong's (1943). We computed the percentage of respondents endorsing each response alternative in the target MOS (initially 91A males) and in the corresponding general population, and then computed the difference between the percentages. These differences were converted to empirical weights, using Strong's table of Net Weights, for each response alternative for all AVOICE items. The resulting scores will be referred to as the occupational key scores.

For both the empirical CTP and the occupational option-level keys, if more than 10 percent of the response options were missing for a respondent, the key score was not calculated and was treated as missing. If respondents were missing one or more response options but the percent missing was less than or equal to 10 percent, their key score was simply the mean of the non-missing responses.

Empirical and Occupational Scales. Empirical scales were developed by selecting items on the basis of the correlation between each item and a criterion variable of interest. Empirical CTP scales were developed by correlating each AVOICE item (scored 1, 2, 3, 4, or 5) with CTP scores within the target MOS development sample. We then rank-ordered the items in terms of their correlation with CTP and selected (a) the top 6 items, (b) the top 12 items, (c) all items whose correlation with CTP reached or exceeded the $p < .05$ significance level, and (d) 10 items beyond $p < .05$. All items in the item-level keys were unit weighted. Items negatively correlated with CTP were reverse-scored so that all items in the key could be summed to create a scale score that would be positively related to CTP. Similar procedures were followed to develop empirical scales to predict Leadership and attrition.

Occupational scales to predict MOS membership were developed using a similar method. Differences between the mean AVOICE item-level scores by members of the target MOS and by members of the corresponding general population were computed for each AVOICE item. Only items for which the mean difference was significantly different than zero ($p < .01$) were included. Items for which the general population had a higher mean score than the target MOS received a negative weight, and all items were summed to create occupational scales scores.

For all of the empirical and occupational scales, respondents missing more than 10 percent of relevant AVOICE item-level data did not receive a scale score. For respondents with 10 percent or less missing item-level data, the score associated with the mid-point of the item-level response options (i.e., the "indifferent" response which received 3 points) was used in place of each missing response. This procedure parallels the rational scoring system for the AVOICE. The midpoint was chosen because the effect on the overall mean for the entire sample was less than if the average of the non-missing items in the scale were used (Peterson et al., 1990).

Empirical and Occupational Composites. For predicting CTP, three different types of empirical composites were developed. First, we created a hand-picked scale-level key by combining all the AVOICE scales which correlated with CTP at the $p < .05$ level. These scales were unit weighted. Scales correlating positively with CTP were assigned a positive weight, those correlating negatively a negative weight.

Initially, to develop regression-picked composites, we tried both forward and stepwise regression, using a conservative criterion of $p < .01$ for entry into (and exit from, in the case of stepwise regression) the equation. In cases where no scales entered the regression equation, we relaxed the criterion to $p < .05$. Forward and stepwise regression produced identical solutions in most samples, so we present results for only the stepwise approach. Using stepwise regression, we developed two regression-picked composites. The scales within one composite were unit weighted; the scales within the other were optimally weighted, using regression weights from the development sample. The sign of the unit weights mirrored the sign of the regression weights.

Several hand-picked occupational composites were developed by computing differences between the AVOICE rational scale means from the target MOS and the

general population. We used significance values and meaningfulness as criteria for choosing varying lengths. Composite scores were simply the unit-weighted sum of the positively or negatively weighted scale scores.

When empirical and occupational composites were being created, persons with 10 percent or more missing item-level data were excluded from the analyses. Further, if any scale score was missing, the composite score was set to missing.

Evaluation of Keys, Scales, and Composites

Empirical keys, scales, and composites were developed in the present research to serve two conceptually different purposes: prediction of organizationally relevant criteria and prediction of occupational membership. Procedures appropriate for evaluating these keys, scales, and composites differed for these two categories.

For evaluating the keys, scales and composites developed to predict CTP, Leadership, and attrition, the statistic of greatest relevance is the correlation between each set of empirical scores and the criterion variable of interest (i.e., the validity) within the relevant MOS. In contrast, for evaluating occupational keys, scales, and composites developed to predict MOS membership, the statistic of greatest relevance is the difference between the mean scores earned by members of the target MOS and by the corresponding general population sample. The higher the mean score obtained by members of the target MOS (relative to members of the general population), the better the scale "works." These differences can be expressed as effect sizes -- the standardized mean difference between two groups' scores. Because these effect sizes are standardized, occupational scoring procedures developed or applied in different samples can be directly compared.

Transportability Analyses. The empirical CTP and occupational scales were evaluated in a variety of additional samples to determine the extent to which certain systematic differences between these samples affect the validity (i.e., transportability) of these scales. First, the empirical and occupational scoring procedures developed within a particular sample in one cohort were transported to the cross-validation sample of the same sex and the same MOS in the opposite cohort. In other words, cohort was the only characteristic that varied systematically between the development and "transportability" samples.

Second, we conducted cross-MOS transportability analyses for the empirical scales and composites developed to predict CTP. Those developed within a particular MOS were transported to the same-sex, cross-validation samples for each of the other MOS within the same cohort. For these analyses, MOS was the only characteristic that varied systematically between samples.

Finally, we investigated cross-sex transportability for scoring procedures designed to predict both CTP and occupational membership. Keys, scales, and/or composites developed within one sex were transported to the opposite-sex cross-validation sample

within the same MOS and within the same cohort. In other words, sex was the only characteristic that varied systematically between samples.

The impact of sex differences was also assessed by comparing empirical scoring procedures developed within each sex and in combined-sex samples. In analyses intended to illuminate male-female differences, we computed mean scores for males and females separately on the empirical CTP and occupational scoring procedures developed within male, female, or combined-sex samples. We then calculated the effect size of the difference between the male and female scores on each key, scale, or composite.

Comparison of Rationally Derived and Empirically Derived Scoring Procedures. Another goal of the present effort was to determine whether empirically derived scoring procedures can be more valid or useful than the rationally derived system, so we computed the validity of the rationally derived system for predicting CTP in each cross-validation sample. As discussed previously, eight rational composites were derived for the LV sample (six for CV). To be consistent with previous Career Force analyses, the validity of the rational composites for a particular criterion was estimated by computing the multiple regression of all eight AVOICE composites on that criterion variable. The resulting multiple correlation was statistically adjusted for shrinkage expected as a result of capitalizing on sample characteristics (using Rozeboom, 1978). The square root of the adjusted multiple correlation was then compared with the cross-validities of the empirical scales and composites.

The validities of the AVOICE rational composites and the CTP empirical scales for predicting first-tour CTP were compared with and without corrections for range restriction. Procedures described by McHenry et al. (1990) were used in correcting validities for range restriction. Multiple regression was used to compute the incremental validity of the AVOICE rational composites and empirical scales over the four ASVAB composite scores for predicting first-tour CTP. The resulting validities were adjusted for shrinkage, using Rozeboom (1978).

Comparison of Empirical Scales and Occupational Scales. Finally, the effectiveness of empirical scoring procedures designed to predict CTP was compared with those developed to predict MOS membership. These analyses include the empirical scales (in contrast to keys or composites) developed in the CV cohort to predict first-tour CTP and MOS membership. Their effectiveness was compared in the LVI sample. (Empirical scales developed to predict second-tour leadership and attrition were not included in this analysis.) The comparisons involved computing the correlation between scores on the occupational scales and CTP in the relevant MOS, and comparing the mean CTP empirical scale score obtained by the target MOS with that obtained by the general population (i.e., computing an effect size).

RESULTS: COMPARISONS OF SCORING PROCEDURES

Comparative Results for Keys vs. Scales vs. Composites

The validity of various empirical keys, scales, and composites for predicting CTP in the MOS 13B development and cross-validation samples, for both the CV and LV cohorts, is shown in Table 3.6. In this and all subsequent tables, a positive validity coefficient indicates that soldiers expressing stronger interest in, or liking for, the content of the AVOICE items included on that scale tend to score higher on the criterion variable (in this case CTP).

As expected, shrinkage in validity from the development to the cross-validation sample is greatest for empirical keys, moderate for empirical scales, and smallest for empirical composites. Table 3.6 also shows that the three approaches tend to produce similar levels of validity in the cross-validation samples. The unit-weighted and optimally weighted empirical composites produce similar levels of validity.

Table 3.6
Comparison of Empirical Keys, Scales, and Composites Developed to Predict Core Technical Proficiencies in the MOS 13B Development and Cross-Validation Samples

	Correlation With First-Tour CTP in the CV Cohort		Correlation With First-Tour CTP in the LV Cohort	
	Development Sample	Cross-Validation Sample	Development Sample	Cross-Validation Sample
Sample Size	237	225-227	346-347	352-353
Empirical Keys				
Vertical Percent (5)	.53	.37	.52	.24
Vertical Percent (10)	.51	.36	.52	.22
Point-Biserial	.55	.29	.55	.21
Empirical Scales				
6-Item Scale	.31	.23	.32	.23
12-Item Scale	.33	.27	.30	.28
All-Significant Scale	.33	.31	.32	.23
All-Significant Plus Ten Scale	.33	.31	.31	.21
Empirical Composites				
Hand-Picked	.26	.29	.24	.22
Stepwise (Unit weights)	.23	.21	.23	.22
Stepwise (Optimal weights)	.23	.21	.23	.23

Note. All correlations are significantly different from zero at the $p < .001$ level.

Empirical keys take considerably more time and effort to develop than do empirical scales or composites, yet they do not appear to predict CTP in the cross-validation sample any better than do the scales and composites. Based on these results, we decided to focus only on empirical scales and composites for predicting CTP (and other organizationally relevant criterion variables) for the remaining MOS. In addition, because the optimally weighted empirical regression composites were no more valid than the unit-weighted regression composites and because they entail more risk of capitalizing on chance than the unit-weighted regression composites, we decided to compute only unit-weighted regression composites for the remaining analyses.

Descriptive statistics are presented in Table 3.7 for the occupational key, scale, and composite developed using the CV male sample to predict membership in MOS 91A and applied in the LVI 91A male and corresponding general population samples. The occupational key approach resulted in the greatest differentiation between the target MOS (91A males) and the general population. The effect size of 2.21 indicates that the 91A males had a mean occupational key score over two standard deviations higher than that of the general population. The occupational scale approach resulted in slightly less differentiation than the key approach, with a difference of 1.69 standard deviations between the 91A males and the general population. The composite approach resulted in the smallest differentiation, with a mean difference of only 1.30. Results for the occupational composites of various lengths were extremely similar, so results are presented on Table 3.7 for only one length, that including five rational scales.

Table 3.7
Comparison of Occupational Keys, Scales, and Composites Developed to Predict MOS 91A Membership for Males

	LV 91A Males			LV Male General Population			
	N	Mean	SD	N	Mean	SD	Effect[a] Size
Key	546	65.67	9.61	6,069	47.66	8.00	2.21
Scale	543	63.21	7.92	6,046	48.24	8.92	1.69
Composite	552	59.40	9.15	6,152	48.02	8.69	1.30

[a] Effect sizes in this table are relative to the target MOS (i.e., a positive effect size indicates that the target MOS scores higher than the general population). All effect sizes are significant at $p < .001$ level.

Not only did the occupational composite yield the poorest differentiation, but the content was less interesting, conceptually, than the content of the keys or scales, because many of the rational scales ended up being included in the various occupational composites. Consequently, all subsequent work for the remaining MOS was focused on developing only occupational keys and scales.

Comparative Results Related to Scale Length

Within the item-level approach (i.e., the development of empirical scales), effect of scale length was assessed by developing scales of four different lengths to predict first-tour CTP and to predict occupational membership for a variety of different MOS. These results were used to determine the most appropriate and interesting lengths, discussed below.

For CTP, empirical scales of varying lengths were formed by computing and rank ordering the correlations between each AVOICE item and CTP (i.e., item-level validities). The first two scales include the 6 most valid items and the 12 most valid items. The third scale includes all items for which the item-level validity was significantly different from zero at the $p < .05$ level, and the final scale adds the next 10 items beyond the $p < .05$ level. Each successive scale includes all of the items from the preceding shorter scales.

The length of the all-significant and the all-significant plus 10 item scales varies depending upon how many items reached statistical significance in each sample. In the CV cohort, the length of the all-significant scales to predict CTP ranges (across MOS) from 9 to 79 items, with a median length of 31. In the LV cohort, the length ranges from 10 to 74 items, with a median length of 17.

Cross-validities for scales of different lengths developed to predict CTP are shown in Table 3.8. It includes the five samples for which we were able to cross-validate empirical scales in both CV and LV cohorts. The short, 6-item scale is usually less valid and is never more valid than longer scales. The longest scale (i.e., all-significant plus 10 items) is sometimes more and sometimes less valid than the two medium-length scales; overall, it appears that adding more items to a scale beyond those significantly correlated with the criterion does not consistently produce a higher validity. Therefore, for the remainder of this report we present only the results for the 12-item and all-significant scales.

To predict occupational membership, scales of several lengths were developed for different MOS. The results for the MOS 91A males are illustrative of the results in general and are presented in Table 3.9. These scales were developed in the CV sample and applied in the LVI sample. Scale lengths of 6 and 12 items were examined because they are consistent with the lengths of the CTP scales already discussed, thus allowing a direct comparison. The all-significant item scale represents the point at which the mean AVOICE item-level difference between the CV 91A males and the corresponding CV general population is significantly different from zero at the $p < .01$ level. In other words, it contains all items that significantly differentiate between the target MOS and the general population. This more conservative significance level was chosen for the occupational scales because most of the AVOICE items would have been included in each scale if we had chosen a less conservative level (e.g., $p < .05$). The all-significant occupational scales range from 27 to 115 items, with a median length of 70. Finally, we examined the scale length between the 12-item scale and the all-significant item scale at

Table 3.8
Cross-Validation Sample Correlations Between Core Technical Proficiency and Empirical Scales of Various Lengths Developed to Predict CTP

MOS	6-Item Scale	12-Item Scale	All-Significant Scale	All-Significant Plus 10 Scale
CV Cohort				
11B	.21**	.28**	.31**	.32**
13B	.23**	.27**	.31**	.31**
63B Male	.41**	.43**	.46**	.47**
71L Female	.11	.22	.18	.24*
91A Male	.15	.15	.20	.23*
Median	.21	.27	.31	.31
LV Cohort				
11B	.17**	.26**	.25**	.25**
13B	.23**	.28**	.23**	.21**
63B Male	.26**	.25**	.23**	.21**
71L Female	.15	.23*	.20	.26*
91A Male	.15*	.17*	.17*	.15*
Median	.17	.25	.23	.21

* Correlation is significant at $p < .01$. ** Correlation is significant at $p < .001$.

which a relatively distinct "drop" in AVOICE item mean difference occurs between the target MOS and the general population (for the 91A males this was 31 items).

Table 3.9 shows the results when the CV 91A occupational scales are applied in the LV 91A male sample and the corresponding LV male general population. The 12-item scale showed the greatest differentiation between the target MOS and the general population, but the differences between the 6-, 12-, and 31-item scales are very small (effect sizes 1.68 to 1.69). The all-significant item scale works least well. This probably reflects the increasing diversity of item content as more items are added to these scales. In addition, as more items are included, the average item-level mean difference between the target MOS and the general population for included items decreases.

To maintain consistency with empirical scales developed to predict CTP, we report only the results for the 12-item and the all-significant occupational scales for the remainder of this report.

Table 3.9
Comparison of Occupational Scales of Various Lengths Developed to Predict MOS 91A Membership for Males

	LV 91A Males			LV Male General Population			
Scale Length	N	Mean	SD	N	Mean	SD	Effect[a] Size
6-Item Scale	549	63.29	7.69	6,115	48.42	8.93	1.68
12-Item Scale	543	63.21	7.92	6,046	48.24	8.92	1.69
31-Item Scale	548	62.31	8.68	6,100	47.84	8.59	1.68
All-Sig. Scale[b]	547	59.77	9.12	6,106	47.93	8.64	1.36

[a] Effect sizes in this table are relative to the target MOS (i.e., a positive effect size indicates that the target MOS scores higher than the general population). All effect sizes are significant at $p < .001$ level.
[b] Includes 69 items.

RESULTS: VALIDITY ESTIMATES

Results for Empirical Scales and Composites Developed to Predict Organizationally Relevant Criteria

This section describes results for the empirical scales and composites developed to predict three organizationally relevant criteria: (a) first-tour Core Technical Proficiency, (b) the second-tour Leadership criterion composite, and (c) attrition.

Cross-Validation Results for Core Technical Proficiency

The top half of Table 3.10 shows correlations in the CV cross-validation samples between CTP and scales/composites developed to predict CTP using the CV cohort. The bottom half of the table shows correlations in the LV cross-validation samples between CTP and scales/composites developed to predict CTP using the LV cohort. Two of the hand-picked empirical composites (CV MOS 71L Male and LV 71L Female) and six of the regression-picked composites (CV 13B Male, CV 63B Male, CV 71L Male, LV 63B Male, LV 71L Female, and LV 91A Male) consist of a single AVOICE rational scale, so they are not really composites. For these samples, the validity of the empirical "composite" is simply the validity of a single AVOICE rationally derived scale.

The extent to which these empirical scales and composites cross-validate varies a great deal. Cross-validities range from .07 (ns) to .46 ($p < .001$) in the CV cohort and from .01 (ns) to .28 ($p < .001$) in the LV cohort. The median level of cross-validity is very similar across cohorts and across scale construction methods (i.e., scales versus composites).

Table 3.10
Cross-Validities for Empirical Scales and Composites Developed to Predict First-Tour Core Technical Proficiency

	Scales		Composites	
MOS	12-Item Scale	All-Significant Scale	Hand-Picked	Regression-Picked
CV Cohort				
11B	.28**	.31**	.24**	.28**
13B	.27**	.31**	.29**	.21**
63B Male	.43**	.46**	.44**	.38**
71L Male	.14	.16	.25	.25
71L Female	.22	.18	.22	.07
91A Male	.15	.20	.24*	.24*
Median	.24	.25	.24	.24
LV Cohort				
11B Male	.26**	.25**	.23**	.24**
13B Male	.28**	.23**	.22**	.22**
63B Male	.25**	.23**	.22**	.23**
71L Female	.23*	.20	.16	.16
91A Male	.17*	.17*	.10	.01
Median	.25	.23	.22	.22

* Correlation is significant at $p < .01$. ** Correlation is significant at $p < .001$.

Empirical scales and composites are most valid for predicting CTP for CV MOS 63B soldiers (r = .38-.46, all significant at $p < .001$). For other MOS samples, including LV 63B, empirical scales and composites show low to moderate cross-validities. The lower cross-validities occur in the smaller samples (CV and LV 71Ls and LV 91A males). Thus, it is not surprising that several of the cross-validities in these samples are not significantly different from zero at the $p < .01$ level.

Cross-Validation Results for Leadership

Correlations between the second-tour Leadership criterion and empirical scales and composites developed to predict Leadership in the LV cohort are shown in Table 3.11, for both the development and the cross-validation samples. As expected, scales show greater shrinkage than composites. However, both scales and composites exhibit relatively low cross-validities for predicting Leadership.

Table 3.11
Development and Cross-Validation Sample Validities for Empirical Scales and Composites Developed to Predict the Second-Tour Leadership Criterion Composite[a]

	Development Sample	Cross-Validation Sample
Sample Size	550-551	529-531
Scales		
12-Item Scale	.28**	.09
All-Significant Scale	.29**	.13*
Composites		
Hand-Picked	.19**	.12*
Regression-Picked	.16**	.19**

* Correlation is significant at $p < .01$. ** Correlation is significant at $p < .001$.

[a] LV cohort only.

Cross-Validation Results for Attrition

Empirical scales and composites were developed to predict 12-month attrition within two MOS samples from the LV cohort: MOS 13B and 91A males. Ten percent of the 13B sample and 13 percent of the 91A male sample attrited within their first 12 months for reasons that were categorized as avoidable. The number of female attrites was too small to create an adequate cross-validation sample.) Because the two criterion groups (i.e., remained in the Army versus attrited) were markedly different in size, we explored whether logistic regression procedures would produce different results than linear regression procedures. Results were virtually identical so, to maintain consistency with the other analyses, attrition analysis results are presented for the linear regression procedures only.

The validity of empirical scales and composites for predicting 12-month attrition in the LV development and cross-validation samples is shown in Table 3.12. Positive validity coefficients indicate that soldiers expressing interest in, or liking for, AVOICE items included in the empirical scale or composite tended to stay in the Army for at least 12 months. As this table shows, empirical scales were not very valid for predicting 12-month attrition in these two MOS. Therefore, we did not develop empirical scales or composites to predict attrition for any additional MOS.

Table 3.12
Development and Cross-Validation Sample Validities for Empirical Scales and Composites Developed to Predict Attrition[a]

	Development Sample		Cross-Validation Sample	
	13B	91A Male	13B	91A Male
Sample Size	2,361	1,618	2,406	1,592
Scales				
12-Item Scale	.14**	.16**	-.02	.07*
All-Significant Scale	.14**	.15**	-.02	.08**
Composites				
Hand-Picked	-- [b]	.08**	-- [b]	.05
Regression-Picked	-- [b]	.07*	-- [b]	.05

* Correlation is significant at $p < .01$. ** Correlation is significant at $p < .001$.

[a] LV cohort only.
[b] No composite could be formed.

Content of the Empirical Scales and Composites

The empirical scale construction strategy provides no guidance for interpreting the content of items included in empirical scales or composites. Indeed, strict adherence to this strategy precludes content interpretation. Typically, however, scale developers do examine such content. For example, it is interesting to determine whether empirical scales and composites designed to predict CTP within a particular MOS tend to include items that seem as if they should predict CTP in that MOS.

The existing AVOICE rational scoring system greatly facilitates interpreting the content of the empirical scales and composites. The AVOICE was explicitly designed to measure vocational interests relevant for military occupations, and the grouping of AVOICE items into internally consistent rational scales, and of scales into internally consistent rational composites, provides a carefully considered framework for interpreting the content of our empirical scales and composites.

Appendix A provides a "map" of the content of each empirical AVOICE scale and composite developed to predict CTP. For empirical scales, these tables show how many items come from each of the rationally derived AVOICE scales and the direction of the relationship between item-level scores and CTP. For composites, the tables show which rationally derived scales are included in each empirical composite and the direction of the relationship between scale-level scores and CTP.

In general, empirical scales and composites developed to predict CTP draw items from a number of different rationally derived scales. In other words, the content of empirical scales and composites is often quite heterogeneous. For most samples the empirical scales draw some items from "obvious" rationally derived scales (e.g., items from the Mechanics scale for MOS 63B). However, for most samples the scales also draw several items from "non-obvious" (or at least less obvious) rationally defined scales (e.g., the LV 91A female scales include several negatively weighted items related to interest in computers). In many cases, it is not hard to think of *post hoc* explanations for certain items being included in empirical scales and composites. However, the rationalizations are not so obvious that we would have been able to predict which items and scales would be in the empirical scales or composites for any particular MOS.

It is also noteworthy that most empirical scales include at least some negatively weighted items. This is one advantage of the empirical approach to scoring. Unlike the rational scoring system, empirical scales can incorporate patterns of likes and dislikes, at the item level of response, to create scales for a particular criterion variable.

Appendix B "maps" the content of empirical scales and composites developed to predict the second-tour Leadership criterion composite. They incorporate a wide variety of AVOICE items, many of which are negatively weighted. For example, in the hand-picked composite only one of the seven rationally defined scales (Rugged Individualism) is positively related to Leadership. The remaining six are negatively weighted, and include activities often considered routine and/or menial (e.g,. clerical, food service, vehicle operation). Interestingly, the empirical scales include either none or only three of the 12 items in the rationally defined Leadership/Guidance scale. Similarly, that scale is not included in either empirical composite. A *post hoc* content interpretation suggests that soldiers who earn higher scores on the Leadership criterion composite tend to like rugged, outdoor activities and to dislike routine or menial activities.

Appendix C "maps" the content of empirical scales and composites developed to predict 12-month attrition. The content of these scales varies widely, and is different for the MOS 13B and 91A samples. The low development sample validities suggest that the content may primarily reflect random "noise." Therefore, no attempt was made to interpret the content of these empirical scales and composites.

Transportability Results for Core Technical Proficiency

Cross-Cohort Transportability for CTP

Table 3.13 shows the correlations with CTP for scales and composites applied in cross-validation samples within the development cohort, and when these same scales and composites are transported to the opposite cohort. Cross-validities for scales and composites developed to predict CTP in each cohort are about equal when applied within the same cohort (median validities range from .22 to .25).

However, scales and composites show much more shrinkage when transported to the LV cohort than when transported to the CV cohort. When those developed in the

Table 3.13
Cross-Cohort Transportability of Empirical Scales and Composites Developed to Predict Core Technical Proficiency

	Correlations Between CTP and Scales/Composites Developed in CV Cohort		Correlations Between CTP and Scales/Composites Developed in LV Cohort	
	Applied in CV Cohort (6 samples)	Transported to LV Cohort (6-7 samples)	Applied in LV Cohort (5 samples)	Transported to CV Cohort (7 samples)
Scales				
12-Item Scale				
Median	.24	.15	.25	.27
Range	.14-.43	.00-.31	.17-.28	.05-.43
All-Significant Scale				
Median	.25	.14	.23	.29
Range	.16-.46	.06-.26	.17-.25	.04-.44
Composites				
Hand-Picked				
Median	.24	.16	.22	.28
Range	.22-.44	-.04-.23	.10-.23	-.19-.43
Regression-Picked				
Median	.25	.13	.22	.25
Range	.07-.38	-.28-.23	.01-.24	-.06-.38

CV cohort are transported to the LV cohort, the validities shrink by about .10. In contrast, when scales and composites developed in the LV cohort are transported to the CV cohort, the validities do not shrink at all -- in fact, they are slightly higher. These results suggest that first-tour CTP is more predictable in the CV cohort than in the LV cohort.

We can speculate about reasons for this finding. First, self-reported interests may be less accurate for applicants (i.e., the LV cohort) than for incumbents (i.e., the CV cohort), because applicants are less familiar with some of the vocational interests measured by the AVOICE (e.g., interest in combat-related activities). This could lead to greater error variance in the AVOICE item-level data for the LV cohort than for the CV cohort which, in turn, could affect the correlations between AVOICE items and CTP.

Second, differences in the applicant versus incumbent response sets could affect the AVOICE data. If applicants "faked" their AVOICE responses more than incumbents did, the relationship between item-level responses and CTP may be impacted by an artifactual source of variance -- variability due to differences in "faking."

Cross-Cohort Content Comparison for CTP

To compare the content overlap of scales and composites developed to predict CTP for the same MOS in different cohorts, we used information from Appendix A to prepare Table 3.14. For empirical scales (12-item and all-significant), to calculate the number of items shared by CV cohort and LV cohort scales, we simply counted the number of items across the two empirical scales that came from the same rationally derived AVOICE scales. Items were counted as "shared" if they came from the same rationally derived scale, even if they were not exactly the same item. The number of exact content matches would clearly be smaller, but the AVOICE rational scales are very internally consistent so the number of items that come from the same rational scale seemed to be the most interesting comparison. Items are considered "unshared" if they come from different rational AVOICE scales or are included on the empirical scale developed for only one cohort.

For the empirical composites, to calculate the number of scales shared by the CV and LV cohorts, we simply counted the number of rationally defined scales that appeared in the empirical composites in both cohorts.

One way to interpret the data in Table 3.14 is to compare the number of shared items (or scales) to the total number of shared and unshared items (or scales). For empirical CTP scales, this ratio ranges from 2:22 (9%) to 21:39 (54%). For empirical CTP composites, the ratio ranges from 0:7 (0%) to 1:1 (100%). Caution must be exercised when interpreting these ratios for the all-significant scale and for the composites because they can vary in length from one cohort to the next. If the all-significant scale in one cohort contains more items than the analogous all-significant scale in the other cohort, the maximum possible overlap is less than 100 percent. The same is true for the empirical composites.

The greater the discrepancy in length between the two comparison scales or composites, the lower the maximum possible overlap. For example, the all-significant MOS 63B male scale in the CV cohort includes 59 items; the analogous LV cohort scale contains 30 items. If all 30 items in the shorter scale overlap with the longer 59-item scale, the maximum possible overlap is 30:59 (50%). The observed overlap between these two scales (Table 3.14) is 24:64 (38 percent). However, given that 50 percent is the maximum possible overlap for scales differing this much in length, the observed overlap actually represents 76 percent of the maximum possible overlap.

At the composite level, the hand-picked composite for MOS 63B males consists of eight scales in the CV cohort, but only four in the LV cohort. Thus, the maximum possible overlap if all four LV scales were also included in the CV cohort composite would be 50 percent (4:8). The observed overlap is four out of eight scales. Thus, the degree of overlap is as high as it can be, given the difference in the number of scales in the two composites. The percentage of possible overlap between the CV- and LV-developed hand-picked composites for 63B males is 100 percent.

Table 3.14
Content Overlap[a] Between Empirical Scales/Composites Developed to Predict Core Technical Proficiency in the CV and the LV Cohorts

	Scales				Composites			
	12-Item		All-Significant		Hand-Picked		Regression-Picked	
MOS	No. of Shared Items	Total Shared and Unshared Items	No. of Shared Items	Total Shared and Unshared Items	No. of Shared Scales	Shared and Unshared Scales	No. of Shared Scales	Shared and Unshared Scales
11B	8	16	33	83	6	12	0	7
13B	6	18	21	39	3	6	0	3
63B Male	10	14	24	64	4	8	1	1
71L Male	5	19	5	20	-- [b]	-- [b]	-- [b]	-- [b]
71L Female	2	22	4	28	0	3	0	3
91A Male	6	18	11	34	2	7	1	3
91A Female	6	18	5	18	1	3	1	3

[a] Items from the same rationally derived scale are counted as a match, even if the two items are not exactly the same.
[b] No composites could be formed in the LV 71L Male sample.

The largest discrepancies in scale or composite length occur for the MOS 63B male and 11B samples. As illustrated above, the degree of overlap for the CV and LV 63B male scales and composites is very high. Similarly, the overlap for the 11B sample is actually higher than the data in Table 3.14 suggest. For example, the all-significant scale achieves 66 percent of the maximum possible overlap across cohorts, given that the scale includes 42 items for the CV cohort and 74 items for the LV cohort. Further, the 11B hand-picked composite actually achieves 78 percent of the maximum possible overlap, given that it includes 7 scales for the CV cohort and 11 for the LV cohort. In general, the content overlap between scales and composites developed in the two cohorts is very high for 63B males, moderate for the 11B, 13B, and 91A male samples, and quite low for the 71L male and female samples.

Cross-MOS Transportability for CTP

Table 3.15 shows the validities for empirical scales and composites applied in their own MOS cross-validation sample and transported to other MOS cross-validation samples within the same cohort. As might be expected, the median validity is higher when empirical scales and composites are applied in the MOS for which they were developed than when they are transported to other MOS; further, the range of validities is larger in the cross-MOS samples than in the within-MOS samples.

Not surprisingly, the degree of similarity between occupations has an impact on the degree to which scales and composites developed to predict CTP in one MOS can be transported to a different MOS. Among the MOS included in our analyses, the two combat MOS (11B and 13B) are more similar to each other than they are to the other MOS (63B, 71L, and 91A). Empirical scales and composites developed in the 11B MOS sample show little shrinkage when transported to the 13B sample, a moderate amount when transported to the 63B male sample, and a great deal when transported to the 71L male and 91A male samples. Scales and composites developed in the 13B sample follow the same pattern.

The information in the appendices can be used to compare content overlap of each MOS sample to every other MOS sample.

Cross-Sex Transportability for CTP

Single-Sex Scales and Composites. The cross-sex transportability analyses were hindered by the fact that, in most MOS, there are either no females (e.g., 11B and 13B) or too few females to adequately develop and/or cross-validate empirical scales and composites (e.g., MOS 63B). Further, for one MOS in which the number of females was adequate (LV 71L), there were too few males. We were able to *develop* empirical scales and composites for CV and LV 63B males, CV and LV 71L males and females, and CV and LV 91A males and females. However, we were unable to *cross-validate* the LV 71L male, CV 91A female, or LV 91A female scales and composites. Consequently, there are only three within-sex cross-validation samples in the CV cohort and only two in the LV cohort.

Table 3.15
Transportability Across MOS of Empirical Scales/Composites Developed to Predict Core Technical Proficiency

	Correlation With CTP in Same MOS (Cross-Validity)		Correlation With CTP in Other MOS	
	CV Cohort (6 samples)	LV Cohort (5 samples)	CV Cohort (22 samples)	LV Cohort (22 samples)
Scales				
12-Item Scale				
Median	.24	.25	.16	.09
Range	.14-.43	.17-.28	-.07-.40	-.05-.26
All-Significant Scale				
Median	.25	.23	.17	.11
Range	.16-.46	.17-.25	-.03-.44	-.02-.27
Composites				
Hand-Picked				
Median	.24	.22	.22	.10
Range	.22-.44	.10-.23	-.05-.44	.01-.23
Regression-Picked				
Median	.24	.22	.10	.07
Range	.07-.38	.01-.24	-.09-.32	-.06-.19

The median and range of validities obtained when applying single-sex empirical scales and composites in same-sex cross-validation samples and when transporting these same scales and composites to opposite-sex cross-validation samples are shown in Appendix Table D.1. These figures all represent within-cohort analyses (i.e., scales and composites developed in the CV cohort are applied in CV cohort, and those developed in the LV cohort are applied in the LV cohort). The results vary by cohort. In the CV cohort, empirical scales developed for one sex are actually somewhat more valid when transported to opposite-sex samples, but empirical composites are equally valid when applied in same sex samples or transported to opposite sex samples. In the LV cohort, all scales and composites are slightly more valid when applied in same-sex samples than when transported to opposite sex samples. The data do not show any consistent pattern. In some cases, scales and composites developed in one sex are more valid when applied in a same-sex sample than when applied in an opposite-sex sample. In other cases, the reverse is true.

Combined-Sex Scales and Composites. The median and range of validities when combined-sex empirical scales and composites are applied in the combined-sex cross-validation samples within the same cohort and when they are transported to the opposite cohort are shown in Appendix D.2. The results can be compared to the results for single-sex scales and composites, shown within cohort in Table D.1. The first two columns of the table provide a direct comparison of the cross-validity of single-sex scales/composites and combined-sex scales/composites. In both cohorts, the combined-sex scales and composites show cross-validities that are about the same as those for single-sex scales and composites.

Male-Female Effect Sizes for Empirical CTP Scales

Male-female effect sizes in the total LVI sample for empirical scales developed to predict CTP are shown in Table 3.16. A positive effect size for an empirical scale indicates that males in the LVI sample tend to score higher, on average, than do females. All but two of the effect sizes shown on this table are positive, indicating that males in the LVI sample tend to score higher than females regardless of the sex of the sample used to develop the scales. The two exceptions occur when empirical scales are developed in combined-sex samples, and one of these is very small (-.06). The largest effect sizes occur for scales developed in the MOS 63B male and combined-sex samples; the smallest occur for scales developed in the 71L male and combined-sex samples. It is interesting to note that the effect size favoring males is actually larger for scales developed using a female sample than for scales developed using a male sample.

Comparison of Empirical CTP Scales/Composites With Rational Scoring Procedures

Table 3.17 shows, for several samples in each cohort, the validity of the rationally derived composites and the median cross-validity across four empirical scales and composites for predicting CTP, Leadership, and 12-month attrition. The validities for the rational composite have been adjusted for shrinkage; since cross-validation provides a more conservative estimate of shrinkage than the statistical correction, the cross-validities of empirical scale/composite are overcorrected, relative to the rational composite, to an unknown degree. No other corrections have been applied.

The validity of the rational composite for predicting first-tour CTP is consistently lower than the median cross-validity of empirical scales and composites developed to predict CTP, particularly in the LV cohort samples. The difference is often quite small, however. For the second-tour Leadership criterion, the validity of the rational composite is somewhat higher than the median validity of the empirical scales and composites (r = .21 versus .13, respectively). Finally, the AVOICE shows little or no validity for predicting attrition, whether scored rationally or empirically (range, -.02 to .07).

Table 3.16
Male/Female Effect Sizes on Scales Developed to Predict Core Technical Proficiency[a] in the LV Cohort[b]

MOS	Development Sample	Effect Size[c]
12-Item Scale		
63B	Males	0.89
	Combined	0.82
71L	Males	0.17
	Females	0.71
	Combined	-0.06
91A	Males	0.37
	Females	0.93
	Combined	0.36
All-Significant Scale		
63B	Males	1.24
	Combined	1.41
71L	Males	0.03
	Females	0.73
	Combined	-0.44
91A	Males	0.54
	Females	0.99
	Combined	0.77

[a] Scales and composites developed in the CV cohort.
[b] Sample sizes for females, 714-772; for males, 6,225-6,657.
[c] All effect sizes in this table are relative to the male subgroup (i.e., a positive effect size indicates a higher score by males). Effect sizes larger than about .10 are significant at the $p < .01$ level.

The validities of the rational composites and empirical scales were corrected for range restriction in five CV and four LV cohort samples.[1] The validities before and after correction are compared in Table 3.18. In most samples, validities become larger when they are corrected for range restriction -- sometimes much larger. The correction for range restriction tends to be greater for the rational composites than for the empirical scales. After correction for range restriction, the validities in some samples are quite high. For example. both rational composites and empirical scales achieve very

[1] The MOS 13B samples could not be included in this analysis because a change in database management procedures made it impossible to add variables needed to perform the range restriction correction to these samples.

Table 3.17
Comparison of Validity of Rational and Empirical Scales and Composites For Predicting First-Tour Core Technical Proficiency, Second-Tour Leadership, and Attrition

	CTP		Leadership[a]		Attrition	
	Rational Composite[b]	Empirical Scales/ Composites[c]	Rational Composite	Empirical Scales/ Composites	Rational Composite	Empirical Scales/ Composites
CV Cohort						
11B	.25	.28	--[d]	--[d]	--[d]	--[d]
13B	.24	.28	--[d]	--[d]	--[d]	--[d]
63B Male	.44	.44	--[d]	--[d]	--[d]	--[d]
71L Male	.00	.21	--[d]	--[d]	--[d]	--[d]
71L Female	.15	.20	--[d]	--[d]	--[d]	--[d]
91A Male	.00	.22	--[d]	--[d]	--[d]	--[d]
Median across MOS	.20	.25	--[d]	--[d]	--[d]	--[d]
Range across MOS	.00-.44	.20-.44	--[d]	--[d]	--[d]	--[d]
LV Cohort						
11B	.15	.25	--[d]	--[d]	--[d]	--[d]
13B	.16	.23	--[d]	--[d]	.04	-.02
63B Male	.08	.23	--[d]	--[d]	--[d]	--[d]
71L Female	.17	.18	--[d]	--[d]	--[d]	--[d]
91A Male	.00	.14	--[d]	--[d]	.07	--[d]
Median across MOS	.15	.23	.21	.13	.06	.02
Range across MOS	.00-.17	.14-.25	none	.09-.19	.04-.07	-.02-.06

[a] Pooled across MOS.
[b] Multiple regression of eight AVOICE rational composites; statistically adjusted for shrinkage using Rozeboom (1978).
[c] Median cross-validities across 12-item and all-significant scales and hand-picked and regression-picked composites.
[d] Not available

Table 3.18
Comparison of Validity of Rational Composites and Empirical Scales for Predicting First-Tour Core Technical Proficiency Before and After Correcting for Range Restriction

	Rational Composite[a]		Empirical Scales			
			12-Item		All-Significant	
	Validity[b]	Validity[c] Corrected for Range Restriction	Cross-Validity	Cross-Validity[d] Corrected for Range Restriction	Cross-Validity	Cross-Validity[c] Corrected for Range Restriction
CV Cohort						
11B	.25	.45	.28	.39	.31	.43
63B Male	.44	.58	.43	.52	.46	.57
71L Male	.00	.00	.14	.15	.16	.17
71L Female	.15	.47	.22	.41	.18	.42
91A Male	.00	.20	.15	.17	.20	.18
Median	.15	.45	.22	.39	.20	.42
Range	.00-.44	.00-.58	.14-.43	.15-.52	.16-.46	.17-.57
LV Cohort						
11B	.15	.36	.26	.40	.25	.40
63B Male	.08	.22	.25	.29	.23	.31
71L Female	.17	.17	.23	.32	.20	.24
91A Male	.00	.41	.17	.34	.17	.35
Median	.12	.29	.24	.33	.22	.33
Range	.00-.17	.17-.41	.17-.26	.29-.40	.17-.25	.24-.40

[a] Multiple regression of eight AVOICE rational composites.
[b] Statistically adjusted for shrinkage using Rozeboom (1978).
[c] Statistically adjusted for shrinkage using Rozeboom (1978) and corrected for range restriction.
[d] Statistically corrected for range restriction.

respectable corrected validities for MOS 11B soldiers in each cohort, 63B male and 71L female soldiers in the CV cohort, and 91A male soldiers in the LV cohort. On the other hand, even after correcting for range restriction, both rational composites and empirical scales show low validity for predicting CTP in 71L male and 91A male samples in the CV cohort.

Finally, the incremental validity of the rationally derived composites and the empirically derived scales for predicting CTP, when each is used in conjunction with the four ASVAB composites, was evaluated. Table 3.19 shows the validity of the ASVAB alone compared to the validity of the ASVAB when combined with the AVOICE in various forms. In general, the empirical scales add slightly more incremental validity, or detract less from the validity of the ASVAB alone, than the rationally derived composite.

Table 3.19
Incremental Validity[a] of AVOICE Rational Composites and Empirical Scales Over the ASVAB for Predicting First-Tour Core Technical Proficiency

	ASVAB Factors (A4) [4]	A4 + Rational AVOICE Composites [12]	A4 + Empirical 12-Item AVOICE Scale [5]	A4 + Empirical All-Significant AVOICE Scale [5]
CV Cohort				
11B	.71	.72	.72	.73
63B Male	.72	.74	.75	.75
71L Male	.73	.71	.72	.72
71L Female	.40	.28	.38	.38
91A Male	.62	.63	.62	.63
Median	.71	.71	.72	.72
Range	.40-.73	.28-.74	.38-.75	.38-.75
LV Cohort				
11B	.64	.63	.64	.64
63B Male	.56	.54	.57	.56
71L Female	.66	.66	.66	.66
91A Male	.69	.68	.70	.70
Median	.65	.65	.65	.65
Range	.56-.69	.54-.68	.57-.70	.56-.70

Note. Numbers in brackets are the numbers of predictor scores entering the prediction equations.

[a] Multiple regression statistically adjusted for shrinkage using Rozeboom (1978) and corrected for range restriction.

Comparison of Random and Empirical Keys

To make certain that the empirical keys are not simply capturing random characteristics of the data set, they were compared with randomly derived scales. The latter were created by randomly selecting items for inclusion in two different 12-item scales. These scales were then applied in the MOS 13B and 91A male cross-validation samples in the CV and LV cohorts. The random scales are much less valid for predicting CTP than are the empirically derived 12-item scales. For the 13B male sample, correlations for the empirical scale with CTP were .27 and .28 for the CV and LV cohorts respectively (both significant at $p < .001$); the random scale correlations with CTP were .04 and .01. For the 91A male sample, the correlations for the empirical scale with CTP were .15 and .17 for the CV and LV cohorts, respectively (both significant at $p < .01$); the random scale correlations with CTP were .12 and .02 respectively.

Overall Results for Occupational Keys and Scales

Descriptive statistics for the occupational keys and for 12-item occupational scales in several MOS are presented in Table 3.20. As described in the method section, these occupational keys and scales were developed using the CV sample only, and separately for the following CV MOS: 91A males, 91A females, 71L males, 71L females, 13Bs and 11Bs. These keys and scales were then applied in the LVI sample.

The occupational keys and scales both work best for the 91A males, as indicated by the difference in mean scores between target MOS and the general population. For example, the 91A males score 2.21 standard deviations higher than the corresponding general population on the 91A male occupational *key;* they score 1.69 standard deviations higher than the corresponding general population on the 91A male occupational *scale.* The occupational keys work least well for 13Bs and the occupational scales work least well for the 71L females.

The occupational keys work somewhat better than the scales for differentiating between the target MOS and the general population in two samples: 91A males and 11Bs. In contrast, the occupational scales work slightly better than the keys for the remaining samples. However, the differences between the keys and the scales are relatively small. On average, the target MOS score about one standard deviation higher than the corresponding general populations for both the occupational keys and the occupational scales. In sum, there were relatively minor differences in the ability of the occupational keys and scales to distinguish the target MOS from the general population.

In Table 3.21 the means for the 12-item occupational scales developed in the CV cohort are applied to the CV and LVI cross-validation samples. As expected, each of the MOS cross-validation samples scores higher than the corresponding general populations on the scale developed for their MOS. For example, in the LVI sample, 71L males score 1.50 standard deviations higher than the relevant general population on the 71L 12-item occupational scale. Effect sizes differentiating target MOS from the corresponding general population in the LVI sample ranged from a low of 0.40 for 71L

Table 3.20
Mean Occupational Key and Scale Scores in Cross-Validation (LVI) Samples

	Keys							Scales (12-item)						
	Target MOS			General Population[a]				Target MOS			General Population[a]			
MOS	N	Mean	SD	N	Mean	SD	Effect[b] Size	N	Mean	SD	N	Mean	SD	Effect[b] Size
91A Males	546	65.67	9.61	6,069	47.66	8.00	2.21	543	63.21	7.92	6,046	48.24	8.92	1.69
91A Females	99	61.06	8.01	663	47.69	10.23	1.34	100	65.24	7.20	667	51.03	10.88	1.36
71L Males	74	59.63	8.71	6,541	48.85	9.44	1.14	74	61.66	8.70	6,545	48.48	8.81	1.50
71L Females	236	64.11	14.99	526	58.57	14.43	.38	243	60.83	10.60	529	56.29	11.52	.40
13B	744	53.37	9.33	5,871	49.77	10.11	.36	746	53.88	9.33	5,850	49.51	9.98	.44
11B	759	55.96	9.75	5,856	50.26	9.49	.60	764	55.08	8.45	5,889	50.50	9.43	.49

[a] General population for each composite consists of members of all other Batch A and Batch Z MOS in that sample.

[b] Effect sizes in this table are relative to the target MOS (i.e., a positive effect size indicates that the target MOS score higher than the general population). All effect sizes are significant at the $p < .001$ level.

Table 3.21
Mean 12-Item Occupational Scale Scores in the Cross-Validation and Cross-Cohort Samples

	CV Cross-Validation							LV Cohort						
	Target MOS			General Population[a]				Target MOS			General Population[a]			
MOS	N	Mean	SD	N	Mean	SD	Effect[b] Size	N	Mean	SD	N	Mean	SD	Effect[b] Size
91A Males	165	62.15	8.22	3,965	48.98	9.47	1.40	543	63.21	7.92	6,046	48.24	8.92	1.69
91A Females	--	--	--	--	--	--	--	100	65.24	7.20	667	51.03	10.88	1.36
71L Males	--	--	--	--	--	--	--	74	61.66	8.70	6,545	48.48	8.81	1.50
71L Females	129	62.47	10.14	302	58.62	10.50	.37	243	60.83	10.60	529	56.29	11.52	.40
13B	308	52.82	10.30	3,812	49.75	9.98	.31	746	53.88	9.33	5,850	49.51	9.98	.44
11B	344	53.22	9.41	3,782	49.71	10.00	.35	764	55.08	8.45	5,889	50.50	9.43	.49
63B Males	290	58.93	7.84	3,843	50.31	9.39	.96	486	61.24	6.22	6,169	49.96	9.27	1.24

[a] General population for each composite consists of members of all other Batch A and Batch Z MOS in that sample.
[b] Effect sizes in this table are relative to the target MOS (i.e., a positive effect size indicates that the target MOS score higher than the general population). All effect sizes are significant at the $p < .001$ level.

females to a high of 1.69 for 91A males. In every case, the target MOS scored higher on their occupational scale than did any of the corresponding general populations.

Figure 3.1 represents the means obtained by males in each of the five target MOS on each occupational scale. The x-axis portrays the samples in which the occupational scales were developed, the y-axis portrays the scale scores, and the entries are the mean scale scores earned by each sample. There is marked differentiation between members of the target MOS and the remaining MOS using scales developed for the 63B, the 71L, and the 91A MOS. There is less differentiation for the 11Bs, but they still clearly score higher than members of the other MOS on the 11B scale. For the 13B scale, however, 71Ls actually score higher than do the 13Bs. The 13Bs have very similar mean scores across all five scales (i.e., they have a flat, but slightly elevated profile), although their highest score does occur on the 13B scale.

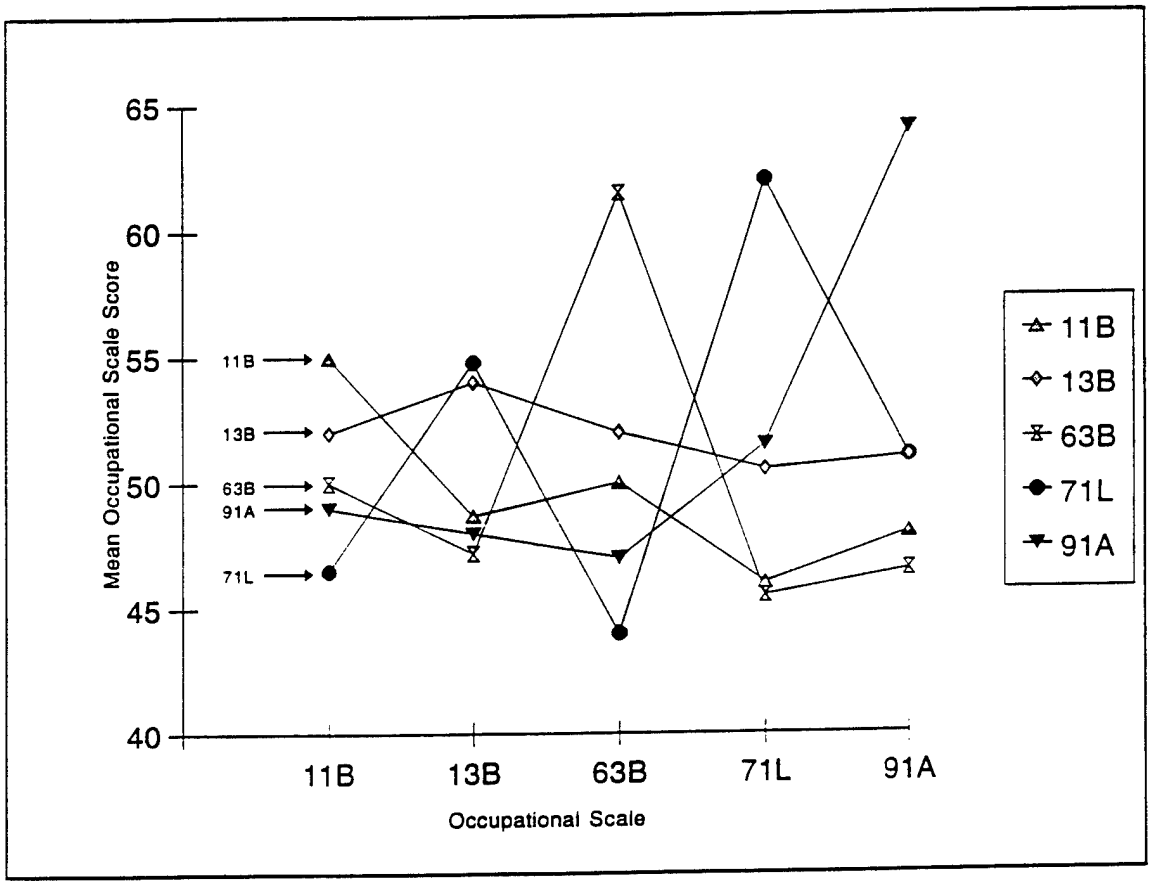


Figure 3.1. Mean occupational scale scores by MOS for males in the LV (cross-validation) sample.

The mean scores for the all-significant item occupational scale developed in the CV cohort are applied to the CV and LVI cross-validation samples in Table 3.22. For the most part, these results are very similar to those from the 12-item occupational scale. Each MOS scores higher on the scale developed for it than do any of the corresponding general populations. For example, in the LVI sample, the 91A females score 1.46 standard deviations higher on the 91A female all-significant occupational scale than their corresponding general population does. For the all-significant scale, effect sizes differentiating the target MOS from their corresponding general populations range from 0.37 to 1.46 (mean unweighted effect size = 0.90).

In the main, the all-significant scales work less well than the 12-item scales in differentiating target MOS from the general population. The target MOS means on their relevant scale tend to be smaller for the all-significant item scales than for the 12-item scales. The effect sizes of the general population compared to target MOS also tend to be smaller for the all-significant item scales than for the 12-item scales.

Content of the Occupational Scales

A "map" of the content of the 12-item and all-significant item scales designed to predict group membership is provided in Appendix E. For each occupational scale, the tables show how many items come from each of the rationally derived AVOICE scales and the direction in which the item is scored in the occupational scale.

Some interesting patterns emerge in examining which AVOICE rational scale items are included in the 12-item occupational scales. For example, both the MOS 91A male and the 91A combined-sex scales contain all 12 of the AVOICE Medical Services rational scale items; for the 91A female occupational scale, eight of the items come from the Medical Services scale. The 63B male 12-item occupational scale contains nine items from the rational Mechanics scale; the 63B combined-sex scale contains eight. For the 11B occupational scale, 8 of the 12 items come from either the Combat, Firearms Enthusiast, or Rugged Individualism rational scales and the remaining 4 come from the Law Enforcement rational scale. (Note: These rational scales all are designed to measure Holland's *Realistic* theme.)

The content of the MOS 13B 12-item occupational scale is perhaps the most heterogeneous. It consists of items from a variety of rational scales that were designed to measure Holland's *Realistic*, *Investigative*, and *Conventional* themes (all positively weighted). Finally, the 71L occupational scales contain the largest number of negatively weighted items. The 71L male 12-item occupational scale contains five items from the Clerical/Administrative rational scale and one item from the Computers rational scale, positively weighted. The remaining items come from a variety of other scales and all are negatively weighted. The 71L female scale contains only one item from the Clerical/ Administrative scale (positively weighted) and the remaining items come from a variety of other scales and receive negative weights.

Table 3.22
Mean All-Significant Occupational Scale Scores in the Cross-Validation and Cross-Cohort Samples

	CV Cross Validation							LV Cohort						
	Target MOS			General Population[a]				Target MOS			General Population[a]			
MOS	N	Mean	SD	N	Mean	SD	Effect[b] Size	N	Mean	SD	N	Mean	SD	Effect[b] Size
91A Males	164	58.97	9.48	3,984	48.44	9.09	1.16	547	59.77	9.12	6,106	47.93	8.64	1.36
91A Females	--	--	--	--	--	--	--	100	64.64	8.12	669	49.40	10.72	1.46
71L Males	--	--	--	--	--	--	--	74	59.40	9.77	6,545	48.52	8.78	1.24
71L Females	127	63.94	10.42	300	60.50	11.05	.32	241	62.56	11.21	529	58.31	11.69	.37
13B	314	53.17	10.00	3,864	49.74	9.96	.34	747	53.65	9.39	5,876	49.54	9.98	.41
11B	344	53.04	9.15	3,803	49.73	10.03	.33	757	55.88	8.66	5,851	50.57	9.17	.58
63B Males	291	58.46	9.79	3,823	50.63	9.12	.85	482	58.73	8.64	6,064	50.73	8.91	.90

[a] General population for each composite consists of members of all other Batch A and Batch Z MOS in that sample.
[b] Effect sizes in this table are relative to the target MOS (i.e., a positive effect size indicates that the target MOS score higher than the general populations). All effect sizes are significant at the $p < .001$ level.

Male-Female Comparisons on the Occupational Scales

Descriptive statistics are shown in Table 3.23 for male, female, and combined-sex samples on occupational scales developed using MOS 63B, 71L, and 91A males only, females only, or males and females combined. For the 71L male and 63B male samples, scales developed using a same-sex sample (i.e., a male sample) more clearly differentated the target MOS from the corresponding general population than did scales developed using a combined-sex or an opposite-sex sample. For the 91A male sample, however, a scale developed using the opposite-sex (i.e., female) sample provided better differentiation between the target MOS and the general population than a scale developed using a same-sex sample.

For 71L females, scales developed using an opposite-sex sample worked better to differentiate between that MOS and the general population than did scales developed using a same-sex or a combined-sex sample. For 91A females, a scale developed using a same-sex sample worked better than scales developed using an opposite-sex or combined-sex sample. For 63B females, there were minimal differences in effect sizes for scales developed using a same-sex versus a combined-sex sample. (The 63B female sample was too small to develop same-sex scales.)

Across MOS, scales developed in same-sex samples are slightly better at differentiating target MOS samples from the corresponding general populations (mean unweighted effect size = 1.27) than are scales developed in opposite-sex or combined-sex samples. Scales developed in opposite-sex and combined-sex samples are, on average, equally effective at differentiating between the target MOS and corresponding general population (mean unweighted effect size = 1.13 and 1.14, respectively). This latter finding may occur because the female samples are so small that they have very little impact on the content of the combined-sex scales.

Male-Female Effect Sizes for Occupational Scales

The effect sizes for mean score by sex on the occupational scales developed for MOS 71L, 91A, and 63B soldiers (both 12-item and all-significant scales) are shown in Table 3.24. Females score higher than males on the 71L 12-item (mean unweighted effect size = -1.28) and 91A 12-item scales (mean unweighted effect size = -0.45) regardless of whether the scales were developed using males only, females only, or a combined sample. Similar results were obtained for the all-significant item scales. Females scored higher on both the 71L and 91A all-significant item scales (mean effect size = -1.35, -0.88), regardless of which sample was used to construct the scales. For both the 63B 12-item and the all-significant item scales, males scored higher than females (mean effect size = 0.84, 1.40) regardless of whether the scales were constructed using males only or a combined sample.

Table 3.23
Same-Sex, Cross-Sex, and Combined-Sex Comparisons of 12-Item Occupational Scales Developed in Males Only, Females Only, and Males/Females Combined Samples

	MOS 63B			MOS 71L			MOS 91A		
	Male	Female	Combined	Male	Female	Combined	Male	Female	Combined
Scales Developed Using Males Only									
Target MOS Mean	61.24	53.80	60.54	61.66	67.74	66.29	63.21	65.17	63.52
General Population Mean[a]	49.96	42.48	49.18	48.48	59.32	49.29	48.24	52.89	48.70
Effect Size[b]	1.24	1.04	1.19	1.50	.82	1.81	1.69	1.21	1.64
Scales Developed Using Females Only									
Target MOS Mean	--[c]	--[c]	--[c]	55.90	60.83	59.68	64.43	65.24	64.56
General Population Mean[a]	--[c]	--[c]	--[c]	49.03	56.29	49.57	48.34	51.03	48.60
Effect Size[b]	--[c]	--[c]	--[c]	.73	.40	1.03	1.84	1.36	1.79
Scales Developed Using Males and Females Combined									
Target MOS Mean	61.00	53.72	60.33	58.99	65.62	64.06	63.21	65.17	63.52
General Population Mean[a]	49.99	42.36	49.20	48.47	60.69	49.38	48.24	52.89	48.70
Effect Size[b]	1.21	1.04	1.16	1.19	.48	1.54	1.69	1.21	1.63

[a] General population for each composite consists of members of all other Batch A and Batch Z MOS in that sample.
[b] Effect sizes in this table are relative to the target MOS subgroup (i.e., a positive effect size indicates that the target MOS score higher than the general population).
[c] Occupational scales were not developed for the female 63Bs because the sample size was not large enough.

Table 3.24
Occupational Scales[a]: Effect Sizes for Mean Score by Sex in the LVI Sample[b]

MOS	Development Sample	Effect Size[c]
12-Item Scale		
63B	Males	0.83
	Combined	0.85
71L	Males	-1.45
	Females	-0.89
	Combined	-1.50
91A	Males	-0.51
	Females	-0.32
	Combined	-0.51
All-Significant Scale		
63B	Males	1.38
	Combined	1.41
71L	Males	-1.43
	Females	-1.14
	Combined	-1.47
91A	Males	-1.11
	Females	-0.15
	Combined	-1.38

[a] Scales developed in the CV cohort.
[b] Sample sizes for females range from 756 to 772; for males, from 6,546 to 6,687.
[c] All effect sizes in this table are relative to the male subgroup (i.e., a positive effect size indicates a higher score by males).

Rational AVOICE Scale and Composite Descriptive Statistics by Sex and MOS

The standardized means scores for the AVOICE basic interest scales and composites within MOS and sex in the LVI sample are shown in Table 3.25. It is apparent from the mean scale scores that the AVOICE basic interest scales have some construct validity. The 11B and 13B MOS both score high on the Combat scale. The 63B males score highest on the Mechanics scale; however, 63B females score highest on the Clerical/Administrative scale (although the small number of female 63Bs makes this result difficult to interpret). The 71L males and females both score highest on the Clerical/Administrative scale and the 91A males and females both score highest on the Medical Services scale.

Table 3.25
AVOICE Rational Scale Means for the LVI Sample by MOS and by Sex

	Males					Females		
Scale	11B	13B	63B	71L	91A	63B	71L	91A
Sample size	775	756	490	78	548	52	249	100
Clerical/Administrative	46.19	51.53	47.06	58.62*	48.88	55.57	61.25*	51.03
Mechanics	49.48	50.07	60.37*	44.83	46.74	52.89	42.19	42.51
Heavy Construction	51.75	52.30	54.19	44.89	47.05	47.22	41.48	41.33
Electronics	49.15	52.18	54.39	48.87	48.50	48.68	44.35	42.50
Combat	54.10	53.70*	49.66	45.09	46.55	44.26	39.78	43.08
Medical Services	46.07	48.40	44.44	48.39	60.33*	50.98	50.63	62.34*
Rugged Individualism	54.42*	49.45	50.88	46.77	50.08	44.46	39.96	45.53
Leadership/Guidance	49.36	50.40	46.38	52.38	52.12	49.80	51.27	53.10
Law Enforcement	52.75	51.17	48.31	49.25	48.71	47.79	46.13	46.56
Food Service Professional	48.53	51.52	48.85	50.25	49.64	53.40	51.52	51.75
Firearms Enthusiast	54.00	51.06	51.82	46.13	49.19	43.21	39.48	39.24
Science/Chemical	49.61	51.45	48.08	50.36	51.96	47.36	46.60	50.29
Drafting	50.02	50.56	48.95	52.59	51.23	45.84	47.88	49.39
Audiographics	48.13	51.40	48.74	53.59	50.26	49.90	51.39	52.47
Aesthetics	48.52	49.48	46.43	52.04	51.40	52.75	55.77	57.92
Computers	47.66	52.44	48.38	54.15	49.32	51.53	54.08	47.07
Food Service Employee	49.75	52.09	49.22	46.89	48.84	51.18	50.28	47.38
Mathematics	48.15	51.13	48.26	52.67	50.51	49.72	54.03	49.54
Electronic Communications	49.40	51.98	49.10	51.00	48.57	49.30	51.10	48.76
Warehousing/Shipping	48.64	53.03	50.27	50.32	47.87	51.69	50.47	45.69
Fire Protection	50.90	51.73	48.74	48.63	50.94	48.38	44.34	46.99
Vehicle Operator	49.69	52.90	53.42	46.59	47.11	49.09	45.66	43.51
Range of SDs	(8.05-10.37)	(7.93-10.33)	(5.95-10.35)	(7.95-11.71)	(7.50-10.43)	(9.73-11.98)	(8.56-11.17)	(7.33-12.21)

Note. * Indicates "most relevant" scale for that MOS in the judgment of the researchers. In each column, the highest and lowest scores are underlined.

The pattern of mean scores for the AVOICE basic interest scales by MOS is generally in accordance with how soldiers score on the relevant empirically derived occupational scale. There is a slight tendency for the mean scores on occupational scales developed for each MOS to be higher than the mean scores on the AVOICE basic interest scale that conceptually seems most relevant for each MOS.[2] For example, LVI 91A females score 0.40 standard deviation higher on the 91A female 12-item occupational scale than they do on the AVOICE Medical Services basic interest scale. LVI 71L males attain scores 0.36 standard deviation higher on the 12-item 71L male occupational scale than they do on the AVOICE Clerical/Administrative rational scale. On average, target MOS score 0.20 standard deviation is higher on the 12-item occupational scales than on their "relevant" AVOICE rational scale.

AVOICE rational composite means are shown in Table 3.26 by MOS and sex for the LVI sample. The standardized means, with one exception, also provide evidence for construct validity of the AVOICE. For example, the 11B males score highest on the Rugged/Outdoors composite and lowest on the Administrative composite. The 63B males score highest on Structural/Machines and lowest on the Social composite, whereas the 63B females score highest on the Administrative composite and lowest on the Rugged/Outdoors. Male and female 71Ls score highest on the Administrative composite; however, male 71Ls score lowest on Structural/Machines and females score lowest on Rugged/Outdoors. Male and female 91As score highest on the Social composite and lowest on Structural/Machines. The 13B sample provides one piece of counterintuitive evidence; they score highest on the Administrative composite and lowest on the Social composite.

Comparison of CTP Empirical Scales and Occupational Scales

Mean scores on the 12-item occupational scales and 12-item CTP empirical scales are shown in Table 3.27 for soldiers in all relevant samples. Table 3.28 shows the same information for the all-significant scales. The results let us assess whether occupational scales are more effective than CTP empirical scales at predicting MOS membership. Effectiveness is defined as the degree to which scales differentiate between a target MOS and a relevant general population (as indicated by the effect size).

Except for MOS 11B and 63B male, the effect sizes are consistently much larger for occupational than for CTP empirical scales (for 11B and 63B male, the effect sizes are almost the same). These results indicate that occupational scales are more effective than empirical CTP scales at predicting MOS membership. This makes sense because occupational scales are specifically designed to achieve this purpose.

[2] The AVOICE rational scales considered "most relevant" according to researcher judgment are identified in Table 3.25.

Table 3.26
AVOICE Rational Composite Means for the LVI Sample by MOS and by Sex

	Males					Females		
Composite	11B	13B	63B	71L	91A	63B	71L	91A
Sample size	758	742-744	486	74	547	49	235-236	98-99
Administrative	47.05	52.52	48.51	55.41	48.21	54.26	56.91	48.24
Audiovisual Arts	48.53	50.59	47.39	53.68	51.26	49.64	52.44	54.26
Food Service	49.03	52.00	48.93	48.49	49.21	52.03	51.11	49.56
Structural/Machines	50.04	52.28	57.07	45.03	46.70	49.67	41.30	40.65
Protective Services	52.16	51.61	48.30	48.34	49.83	48.08	44.33	46.33
Rugged/Outdoors	54.81	51.52	50.94	45.04	48.45	42.61	37.89	41.88
Social	47.20	49.18	44.59	50.51	57.36	50.61	51.54	59.16
Skilled Technical	48.29	52.27	47.97	52.71	50.08	50.11	52.06	48.55
Range of SDs	(8.43-10.37)	(8.22-10.22)	(6.94-10.07)	(8.03-11.85)	(8.39-9.99)	(9.46-11.43)	(8.51-10.36)	(9.32-12.17)

Note. In each column, the highest and lowest scores are underlined.

Table 3.27
Comparison of Mean Scores on 12-Item Occupational Scales and 12-Item CTP Empirical Scales in LVI Sample

	12-Item Occupational Scales							12-Item Empirical Scales						
	Target MOS			General Population[a]				Target MOS			General Population[a]			
MOS	N	Mean	SD	N	Mean	SD	Effect[b] Size	N	Mean	SD	N	Mean	SD	Effect[b] Size
11B	756	55.15	8.43	3,081	51.30	9.37	0.42	756	55.47	9.88	3,081	50.97	8.95	0.49
13B	643	53.65	9.41	3,003	48.56	9.88	0.52	643	49.13	9.80	3,003	51.51	9.70	-0.24
63B Males	476	61.23	6.23	3,343	50.11	9.28	1.24	476	60.09	6.82	3,343	50.28	9.37	1.08
71L Males	74	61.66	8.70	3,766	47.66	8.48	1.65	74	49.91	10.37	3,766	50.22	9.87	-0.03
71L Females	236	60.82	10.62	297	54.91	11.33	0.54	236	41.34	9.94	297	45.96	10.42	-0.45
91A Males	540	63.25	7.92	3,288	48.11	8.72	1.76	540	58.29	9.47	3,288	50.44	9.29	0.84
91A Females	99	65.29	7.22	434	50.81	10.76	1.42	99	47.07	11.37	434	41.34	12.45	0.47

Note: Occupational and empirical scales were developed in the CVI sample and applied in the LVI sample.

[a] General population for each composite consists of members of all other Batch A and Batch Z MOS in that sample.
[b] Effect sizes in this table are relative to the target MOS (i.e., a positive effect size indicates that the target MOS score higher than the general population). All effect sizes are significant at the $p < .001$ level.

Table 3.28
Comparison of Mean Scores on All-Significant Occupational Scales and All-Significant CTP Empirical Scales in the LVI Sample

	All-Significant Occupational Scales							All-Significant Empirical Scales						
	Target MOS			General Population[a]				Target MOS			General Population[a]			
MOS	N	Mean	SD	N	Mean	SD	Effect[b] Size	N	Mean	SD	N	Mean	SD	Effect[b] Size
11B Males	756	55.89	8.66	3,081	51.60	8.58	0.50	756	55.50	9.45	3,081	51.27	8.65	0.48
13B Males	643	53.47	9.45	3,003	48.81	10.01	0.47	643	49.79	9.30	3,003	51.41	9.73	-0.17
63B Males	476	58.72	8.61	3,343	50.90	8.62	0.91	476	58.51	7.79	3,343	50.87	8.65	0.89
71L Males	74	59.40	9.77	3766	47.58	8.48	1.39	74	51.64	9.98	3,766	49.68	9.91	0.20
71L Females	236	62.61	11.21	297	56.61	11.56	0.53	236	41.32	10.48	297	45.70	10.55	-0.42
91A Males	540	59.86	9.13	3,288	46.87	8.53	1.51	540	56.15	9.39	3,288	51.18	8.94	0.55
91A Females	99	64.74	8.10	434	49.55	10.61	1.49	99	46.09	11.68	434	40.94	12.45	0.42

Note: Occupational and empirical scales were developed in the CVI sample and applied in the LVI sample.

[a] General population for each composite consists of members of all other Batch A and Batch Z MOS in that sample.
[b] Effect sizes in this table are relative to the target MOS (i.e., a positive effect size indicates that the target MOS score higher than the general population). All effect sizes are significant at the $p < .001$ level.

Table 3.29 shows correlations between scale scores and CTP for the 12-item occupational scales and the 12-item CTP empirical scales in several samples. Table 3.30 contains the same information for the all-significant scales. The results let us assess whether CTP empirical scales are more effective than occupational scales at predicting CTP. Effectiveness is defined as the size of the correlation between scale score and CTP (i.e., the validity).

The pattern of results is the same for the 12-item scales and the all-significant scales. Specifically, both occupational and CTP empirical scales show moderate levels of validity (significantly different from zero at the $p < .001$ level) for the 11B, 13B, and 63B male samples and low, non-significant levels for the 71L male, 71L female, and 91A female samples. For the 91A male sample, the CTP empirical scale validity is relatively low but significant ($p < .001$), and the occupational scale validity is low and non-significant. The CTP empirical scale validities are consistently higher than the occupational scale validities. These results suggest that scales developed specifically for the purpose of predicting first-tour CTP are at least somewhat more effective at predicting CTP than scales developed to predict MOS membership.

Next, the content overlap between occupational and CTP empirical scales was assessed. Table 3.31 shows the number of items they share. For 12-item scales, the ratio of overlap varies from 1:23 (4%) to 11:13 (85%). These ratios can be interpreted directly because all scales include the same number of items. The ratios for all-significant scales must be interpreted with more caution, since CTP empirical scales often include far fewer items than the occupational scales (due, at least in part, to differences in sample sizes used in development). Thus, the highest possible amount of overlap is often much lower than 100 percent.

When differences in scale length are taken into account, the amount of maximum possible overlap in the two types of scales ranges from 2.7 percent to 94 percent. Some items are weighted positively in one type of scale and negatively in the other -- sometimes for only one or two items (63B) and sometimes for many items (71L).

In general, the degree of overlap varies by MOS. It is high for the 11B and 63B samples, moderate for the 91A male sample, and relatively low for the 13B, 71L female, 71L male, and 91A female samples. The high content overlap for 11B and 63B explains why both occupational and CTP empirical scales tend to be fairly effective for both predictive purposes, that is, MOS membership and CTP. The two types of scales contain many of the same items.

In a final comparison, we computed the correlation between scores earned by each sample on each type of scale. Not surprisingly, the correlations correspond to the degree of content overlap. The average correlation between the occupational and CTP empirical scales (across the 12-item and all-significant scales) is .84 for the 63B male sample, .63 for 11B, .41 for 91A male, .21 for 91A female, and .09 for 13B. Interestingly, occupational and CTP empirical scales correlate negatively with each other in the 71L female and 71L male samples (average correlations are -.27 and -.42, respectively).

Table 3.29
Comparison of 12-Item Occupational and CTP Empirical Scale Correlations With Core Technical Proficiency in the LVI Sample

	Occupational Scales		Empirical Scales	
MOS	N	r	N	r
11B	756	.18*	756	.23*
13B	643	-.06	643	.17*
63B Males	476	.25*	476	.28*
71L Males	74	-.02	74	.11
71L Females	236	.09	236	.06
91A Males	540	.10	540	.15*
91A Females	99	.03	99	.16

Note. Occupational and empirical scales were developed in the CVI sample and applied in the LVI sample.

* Correlation is significant at $p < .001$.

Table 3.30
Comparison of All-Significant Occupational and CTP Empirical Scale Correlations With Core Technical Proficiency in the LVI Sample

	Occupational Scales		Empirical Scales	
MOS	N	r	N	r
11B	756	.24*	756	.26*
13B	643	-.04	643	.17*
63B Males	476	.23*	476	.27*
71L Males	74	-.02	74	.12
71L Females	236	.08	236	.09
91A Males	540	.09	540	.18*
91A Females	99	.10	99	.14

Note. Occupational and empirical scales were developed in the CVI sample and applied in the LVI sample.

* Correlation is significant at $p < .001$.

Table 3.31
Content Overlap[a] Between Occupational Scales and Empirical Scales Developed to Predict Core Technical Proficiency

MOS	12-Item Scales		All-Significant Scales		Maximum Possible Overlap (%)
	No. of Shared Items	Total Shared and Unshared Items	No. of Shared Scales	Total Shared and Unshared Scales	
11B	5	19	33 [b]	65	78
13B	3 [b]	21	13 [b]	58	27
63B Male	9	15	56	117	92
63B Combined-Sex	11	13	73 [b]	119	94
71L Male	1	23	7 [b]	98	67
71L Female	2 [b]	22	10 [b]	73	54
71L Combined-Sex	4 [b]	20	17 [b]	119	78
91A Male	3	21	19 [b]	79	57
91A Female	3 [b]	21	5 [b]	35	29
91A Combined-Sex	4	20	17 [b]	91	36

Note. Occupational and empirical scales were developed in the CVI sample and applied in the LVI sample.

[a] Items from the same rationally derived scale are counted as a match, even if the two items are not exactly the same.

[b] At least one shared item is weighted in one direction in the empirical scale developed to predict CTP and in the opposite direction in the occupational scale.

DISCUSSION

Psychometric Issues

Results of the analyses suggest that in samples as large as those included in the present research, most empirical keys, scales, and composites work equally well for predicting organizationally relevant criteria such as Core Technical Proficiency. The results for the occupational keys, scales, and composites indicate that while all three levels of

scale construction are useful for differentiating members of target MOS from their respective general populations, occupational keys and scales work much better than do the composites. It is interesting to note that while empirical composites work as well as empirical scales for predicting organizationally relevant criteria, the composite-level approach appears to be somewhat less useful for predicting MOS membership.

The findings for scale length vary somewhat depending on the purpose for which the scales are developed, but the 6-item scales are relatively ineffective for either purpose. This may be due to the fact that the AVOICE contains many items, and therefore is likely to contain more than six items that are useful for any given purpose.

For scales developed to predict CTP, the 12-item scales are *less* effective than the all-significant scales. Some of the all-significant CTP empirical scales include 25-75 items. These scale lengths tend to occur in the larger MOS samples (11B, 13B, 63B males, 91A males) and scales of this length are consistently more valid than 12-item CTP empirical scales. Other all-significant scales include fewer than 25 items. These scale lengths tend to occur in the smaller MOS samples (71L males, 71L females, 91A females) and to be more valid than the 12-item scales, but not consistently so. Finally, the longest CTP empirical scales (i.e., the all-significant plus 10-item scales) are no more valid than the all-significant scales. For scales developed to predict CTP, then, there appears to be a point of diminishing returns, in terms of validity, when adding items no longer significantly improves the prediction of the criterion variable of interest.

For scales to predict MOS membership, the 12-item scale is consistently more effective than the all-significant scale (which is always longer than 12 items) for differentiating between the target MOS and the relevant general population. The all-significant occupational scales often include many items (usually more than 50 and often more than 100). As more items are added, the content becomes much more heterogeneous, and the average item-level mean difference between the target MOS and the general population across the included items decreases. Thus, adding more items to the scales actually decreases ability to differentiate between the target MOS and the general population. Thus, for the occupational scales the point of diminishing returns appears to occur somewhere between the 12-item and the all-significant item scales.

Empirical Scales and Composites Developed to Predict Organizationally Relevant Criteria

Effectiveness in Prediction

Table 3.32 summarizes information concerning the effectiveness of empirical scales and composites for predicting organizationally relevant criteria. This table shows the cross-validities for scales and composites developed to predict three criterion variables: first-tour Core Technical Proficiency, attrition, and second-tour Leadership. In general, empirical scales and composites were not found to be very useful for predicting attrition or second-tour Leadership.

Table 3.32
Summary of Results for Empirical Scales and Composites Developed to Predict First-Tour Core Technical Proficiency, Attrition, and Second-Tour Leadership

	Core Technical Proficiency (CTP)				Attrition		Leadership	
	CV Cohort		LV Cohort		LV Cohort		LV Cohort	
	No. of Samples	Median r	No. of Samples	Median r	No. of Samples	Median r	No. of Samples	Median r
Scales								
12-Item	6	.24	5	.25	2	.03	1	.09
All-Significant	6	.25	5	.23	2	.03	1	.13*
Composites								
Hand-picked	6	.24	5	.22	1	.05	1	.12*
Regression-picked	6	.24	5	.22	1	.05	1	.19*

* Correlation is significant at $p < .01$.

Empirical scales and composites do, on average, predict first-tour CTP reasonably well. The median cross-validity ranges from .23 to .25 across both cohorts. When these values are corrected for range restriction, as shown in Table 3.18, the cross-validities become quite respectable for several of the samples. For 12-item scales, the median of the corrected cross-validities is .39 in the CV cohort and .33 in the LV cohort. For all-significant scales, the median is .42 in the CV cohort and .33 in the LV cohort.[3]

The median values reported in Tables 3.18 and 3.32 disguise the fact that there is considerably more variability, across sample, in the CV cohort cross-validities than in the LV cohort cross-validities, as shown in Table 3.10. For example, the highest cross-validity in the CV cohort is about .43 ($p < .001$) for the 63B male sample and the lowest is about .18 (ns) for the 71L female sample. In the LV cohort, the variability of cross-validities is smaller, ranging from about .25 ($p < .01$) for the 11B, 13B, and 63B male samples to about .12 (ns) for the 91A male sample.

To summarize, CTP is well-predicted by empirical scales and composites for some MOS (i.e., the 63B MOS, and to a lesser extent, the 11B and 13B MOS), but poorly predicted in other MOS (i.e., the 71L and 91A MOS) in the CV cohort. The pattern of results is the same in the LV cohort but there is less variation across MOS in the degree to which CTP is predicted. In the LV cohort, CTP is predicted only moderately well in the 63B, 13B, and 11B MOS and poorly in the 71L and 91A MOS.

[3] Table 3.32 includes scales developed to predict CTP in the 13B samples, but Table 3.18 does not, because we could not correct the MOS 13B cross-validities for range restriction.

The content of empirical scales and composites developed to predict organizationally relevant criterion variables is, in general, heterogeneous. Each scale or composite includes items from a variety of AVOICE rational scales. Some of the content would be expected on a rational basis (e.g., positive items on mechanical activities on a scale predicting CTP for MOS 63B). Some of the content, while not necessarily surprising, would be difficult to predict solely on a theoretical basis (e.g., negative items on fire protection tasks on a scale predicting CTP for 13B). A few of the scales contain counterintuitive content (e.g., the scale predicting CTP for 91A females contains more items related to Rugged Individualism, positively weighted, than items related to medical services, also positively weighted).

Transportability Analyses for CTP Empirical Scales

The two sample characteristics that clearly impact transportability of empirical scales and composites developed to predict CTP are cohort and MOS. The results are much less clear as to the impact of sample sex on transportability.

Validities of scales and composites developed in the LV cohort exhibit little or no shrinkage, on average, when transported to the CV cohort. However, the reverse is not true; validities of scales and composites developed in the CV cohort shrink substantially, on average by about .10, when transported to the LV cohort. These findings indicate that cohort or response set differences can have an important impact on the transportability of empirical scales and composites. The degree of content overlap between scales and composites developed in the two cohorts is high for the MOS 63B male samples, moderate for the 11B, 13B, and 91A male samples, and low for the 71L male, 71L female, and 91A female samples.

Many theories of vocational interests would predict that occupation (i.e., MOS) will have a strong impact on the transportability of empirical scales and composites designed to predict job performance. Given that persons in different jobs typically express different patterns of vocational interests, it is not unreasonable to expect different vocational interest items to predict job performance in different jobs. One would also expect this effect to be moderated by the degree of similarity between the job for which the scale was developed and the job to which the scale is transported. In general, this is what the results show. Empirical scale and composite validities exhibit little shrinkage when transported to a similar MOS (e.g., from 11B to 13B), but exhibit a great deal of shrinkage when transported to a very different MOS (e.g., 13B to 71L).

The results for cross-sex transportability analyses do not lead to any clear conclusions. Scales and composites developed to predict CTP for one sex sometimes do, and sometimes do not, work better for predicting CTP within the same-sex sample than when transported to an opposite-sex sample. These inconsistent findings may be partly due to the fact that several of the single-sex samples are quite small, and we would expect the validities of empirical scales and composites developed and applied in small samples to be relatively unstable. For the same reason, differences in the content of the empirical scales and composites developed for each sex may be as much a function of

chance as of true sex differences in the relationship between vocational interests and job performance.

Occupational Keys to Predict MOS Membership

Effectiveness in Prediction

Overall, the occupational scale results are very encouraging. Occupational keys and scales are very effective in differentiating members of a target MOS from the general population for the 11B, 63B, 71L, and 91A samples. They are less effective for the 13B sample, but still provide reasonable differentiation between 13Bs and the general population. Soldiers in all five MOS obtain higher mean scores than the corresponding general populations on the occupational scale developed for that MOS. In addition, they obtain higher mean scores on their own occupational scale than on any of the occupational scales developed for the other MOS. Finally, except for the 13B MOS, members of all the MOS samples obtain higher mean scores on their respective scales than do members of any other MOS samples.

Using the occupational scales, it was also possible to correctly classify (into their actual MOS) approximately 54 percent of the LV male soldiers from the five MOS for whom we developed occupational scales. This result is very similar to results obtained by other researchers. For example, Hansen and Tan (1992) reported direct matches between declared or intended college major and scores on the occupational scales of the Strong Interest Inventory for 64 percent of the males in their concurrent validation study.

The content of each of the 12-item occupational scales makes a great deal of sense. The 11B, 63B, and 91A MOS scales all contain items from rational scales that are conceptually relevant for that MOS and all are positively weighted. The 71L 12-item scale contains fewer items that seem directly relevant to their MOS, but also contains several negatively weighted items on which 71Ls could be expected to score very low (e.g., items from the Firearms Enthusiast rational scale).

The content of the 13B 12-item occupational scale is by far the most heterogeneous and difficult to interpret. It contains items that come from a wide variety of rational interest scales and represent several of Holland's themes. This diversity of item content is clearly reflected in the occupational scale scores for the 13Bs. In Figure 3.1, the 13Bs had a generally flat, but slightly elevated profile across the five occupational scales developed for their own and for the other four MOS. In other words, the 13B sample scored slightly higher than the general population on all five occupational scales, but did not score particularly high on any one occupational scale.

Transportability Analyses for Occupational Scales

Parallel to the findings for scales and composites to predict CTP, results show that cohort and MOS clearly impact the transportability of occupational keys and scales, but are much less clear on the impact of sample sex on transportability.

Occupational keys and scales consistently worked better when applied in the LVI sample than when applied in the CVI sample. This finding is surprising because application in the LVI sample is both cross-validation and cross-cohort. It is also surprising because the difference in response set between the CVI and LVI samples, if it has any impact at all, should work to weaken the effectiveness of the occupational scales when transported to the LVI samples. This result is also the opposite of that obtained for the empirical scales and keys designed to predict CTP, where the scales developed in the CV sample showed a fair amount of shrinkage when applied to the LV cohort. The fact that occupational keys and scales work better in the cross-cohort sample might mean that (a) the response set difference had little effect on AVOICE responses or (b) occupational membership is more "predictable" for the LVI sample in spite of greater potential for distortion of AVOICE responses in that sample.

As expected, most occupational keys and scales work better for predicting membership in the MOS for which they were developed than for other MOS. The only exception occurs for scales and keys developed to predict membership in the 13B MOS. The 71L male sample actually earns a slightly higher mean score than the 13B sample on the occupational keys and scales developed specifically for the 13B MOS.

In general, occupational keys and scales differentiate slightly better between the target MOS and the general population when they are applied in the same-sex validation sample than in an opposite-sex cross-validation sample. However, this advantage is very small and there are several exceptions.

Comparison of CTP Empirical and Occupational Scales

Comparisons of the CTP empirical and the occupational scales show that the empirical scales developed to predict CTP tend to do so better than do the occupational scales designed to predict MOS membership. Conversely, the occupational scales developed to predict MOS membership do so much better than do the CTP empirical scales. In those instances where both sets of scales tend to predict MOS membership and CTP fairly well (e.g., 11B, 63B males, and 91A males), there is considerable content overlap between the occupational and CTP empirical scales. Not surprisingly, empirical and occupational scales work best for the purpose they were designed to achieve.

Comparison of Rational and Empirical Scales

The validity of empirical scales and composites developed specifically to predict first-tour CTP is consistently slightly higher than the validity of the rational composites for predicting CTP. This is especially evident in the CV cohort. (Recall that other analyses suggest that CTP is better predicted in the CV cohort than in the LV cohort.) When this trend is combined with the fact that the cross-validation procedure used to evaluate the empirical scales provides a more conservative estimate of shrinkage than does the statistical procedure used to adjust for shrinkage of the rational composites, it appears that empirical scales and composites are somewhat better predictors of CTP than are the rational composites. When both scoring procedures are corrected for range

restriction (as shown in Table 3.18), the trend becomes stronger for the LV cohort, but less distinct for the CV cohort.

Finally, empirical scales developed to predict CTP add slightly more incremental validity when used in conjunction with the ASVAB to predict CTP than when rational composites are used in conjunction with the ASVAB to predict CTP, although the AVOICE does not add much incremental validity in either case.

Occupational keys and scales are also uniformly more effective than the rational AVOICE scales in differentiating between a target MOS and the general population, although the relevant rational scales also work quite well for this purpose.

CONCLUSIONS

The results described above suggest that empirical scoring procedures can improve the validity and/or usefulness of the AVOICE, at least in some instances. The degree of improvement appears to be larger for predicting occupational membership than for predicting organizationally relevant criterion variables. We believe several additional analyses could fully explore the usefulness of empirical scoring procedures for the AVOICE. We recommend the Army consider conducting the following analyses:

(1) Applying the occupational scales in the LVII sample. This would permit testing the hypothesis that the longer soldiers have been "on the job," the better occupational scales would differentiate between the target MOS and the general population. (This involves an assumption that soldiers who like their jobs better re-enlist.)

(2) Comparing the accuracy of classification based on the occupational scales with the accuracy of classification that could be achieved using the AVOICE rational scales. A one-way discriminant analysis could be used to predict MOS membership on the basis of the AVOICE rational scales.

(3) Developing and cross-validating occupational scales for the remaining MOS with adequate sample size. This could include some Batch Z MOS.

(4) Developing and cross-validating empirical scales to predict CTP in the remaining MOS with available CTP data (and adequate sample size).

(5) Developing and cross-validating empirical scales to predict one or two aspects of job satisfaction (e.g., satisfaction with the work itself) in one or two MOS. If initial analyses appear promising, develop empirical scales in several MOS.

(6) Assessing the relationship between occupational scale scores and one or two aspects of job satisfaction, as measured by the Army Job Satisfaction Questionnaire. We do not necessarily expect occupational scales to predict

overall job satisfaction, but hypothesize that occupational scales may predict satisfaction with the work itself.

(7) Developing and cross-validating empirical scales to predict the Effort and Leadership (ELS) criterion composite in the CV and LV cohorts. The initial analyses could be conducted in a pooled-MOS sample because the content of this performance factor is very similar across all MOS. If the initial analyses look promising, develop and cross-validate empirical scales and/or composites to predict ELS within one or two MOS to see if the level of prediction varies significantly by MOS.

(8) Exploring how race differences impact the effectiveness of empirical scoring procedures by comparing scales developed within Caucasian samples to scales developed within African-American samples (the only two races for which there is adequate data to develop and cross-validate empirical scales). These analyses could be focused on prediction of CTP or on prediction of MOS membership.

Chapter 4

MAXIMIZING SELECTION VALIDITY FOR PREDICTING FIRST-TOUR PERFORMANCE

Scott H. Oppler, John P. Campbell, and Norman G. Peterson

Up to this point, the analyses of predictor battery validities in Project A/Career Force have followed a standard format that has been referred to as the "basic validation analyses." Previous reports have presented the basic validation analyses for the prediction of training performance, the prediction of first-tour performance, and the prediction of second-tour performance. A basic validation analysis has consisted of two principal parts.

First, the predictor scores from each predictor domain have been correlated with each performance criterion factor score. The predictor "domains" are (a) subscores from the ASVAB, (b) the scores from the paper-and-pencil spatial tests, (c) the scores from the computerized measures of perceptual and psychomotor ability, (d) the subscores from the ABLE, and (e) the subscores from the AVOICE. In most, but not all, analyses the ASVAB has been represented by its four factor scores. Research with the ABLE has produced three sets of subscores, one based on the previous literature/theory-driven development of subscores and two based on factor analysis (see Chapter 2. Campbell & Zook, 1994a). Within each domain, the individual scores could be either unit weighted or regression weighted.

Second, the incremental validities of each predictor domain over the ASVAB when predicting each performance criterion factor score have been estimated. That is, when the four ASVAB factor scores were combined with the subscores from a second domain in the same prediction equation, and all predictor scores in the combined set were regression weighted, did the combined equation yield a higher validity than the four ASVAB factor scores alone?

The basic validation analyses do not use all the information from ASVAB and the Experimental Battery in one equation so as to maximize the degree of predictive validity that can be obtained from the full predictor battery. So far, the project analyses have not tried to estimate the upper limit of predictive accuracy when all information is used. The analyses reported in this chapter and Chapter 6 attempt to do so.

The overall objectives for the analyses reported in this chapter were as follows:

(1) Consider the complete array of full prediction equations for each of the five performance criterion factors in each of the nine Batch A MOS (5 x 9 = 45) and determine the minimum number of equations that can be used without loss of predictive accuracy. For example, for any particular criterion factor, does each MOS require a unique equation (i.e., nine equations) or will fewer unique equations, perhaps only one, yield the same level of validity in each MOS?

(2) Once the minimum number of equations had been identified, estimate the maximum validity (selection efficiency) that could be obtained from a "reduced" prediction equation when the purpose for reducing the length of the test battery was either:

(a) To maximize selection efficiency. That is, would a shorter battery yield a higher estimate of the population validity than a longer battery because the longer battery includes scores that add more error variance than relevant variance?

or

(b) To maximize classification efficiency. That is, if the goal was to reduce the number of predictors so as to increase the differences in the equations across MOS, or across criterion factors within MOS, how did that goal affect the selection efficiency of the battery?

Chapter 6 will discuss the set of analyses that attempted to estimate the maximum validity that could be obtained when prior information about individual training performance and/or job performance is combined with the available ASVAB and Experimental Battery scores. This is the so-called "roll-up" analysis that can be applied as individuals progress through training, through their first tour of duty, and on into their second tour.

DIFFERENTIAL PREDICTION ACROSS CRITERION CONSTRUCTS AND ACROSS MOS

The purpose of the analyses reported here was to examine the extent to which the number of equations developed to predict first-tour performance could be reduced from 45 (9 jobs x 5 equations, one equation per job for each of the five LVI criterion constructs) while minimizing the loss of predictive accuracy. Specifically, two separate sets of analyses were conducted: one set to evaluate the reduction of equations across criterion constructs, within MOS, and the other to evaluate the reduction of equations across MOS, within criterion construct.

Sample

The sample consisted of first-tour soldiers in nine of the 10 LVI Batch A MOS. Soldiers in 19E were not included in the analyses for the same reasons that they were not included in the original LVI validation analyses: (a) There were not very many of them, and (b) the MOS was in the process of being phased out (Oppler, Peterson, & Russell, 1994). To be included in the analyses, soldiers were required to have complete LVI criterion data, complete ASVAB data, and complete data for all composites derived from the Experimental Battery. This resulted in a total sample of 3,086 soldiers (11B = 235; 13B = 551; 19K = 445; 31C = 172; 63B = 406; 71L = 251; 88M = 221; 91A = 535; and 95B = 270).

Measures

Predictors

For the present investigation, two sets of predictors were examined. The first set included the four ASVAB factor composites, plus the one unit-weighted spatial test composite and the eight composite scores obtained from the computerized test measures (for a total of 13 predictors). The second set of predictors included the four ASVAB factors, plus the seven ABLE and eight AVOICE subscale composite scores (for a total of 19 predictors).

Criteria

Eight criteria were used in the present investigation: the five criterion constructs corresponding to the five factors from the LVI performance model (Core Technical Proficiency [CTP], General Soldiering Proficiency [GSP], Effort and Leadership [ELS], Maintaining Personal Discipline [MPD], and Physical Fitness and Military Bearing [PFB], plus three higher order composites of these five constructs. The higher order composites (labeled Can-Do, Will-Do, and Total) were formed by standardizing and adding together CTP and GSP (Can-Do), standardizing and adding together ELS, MPD, and PFB (Will-Do), and standardizing and adding together Can-Do and Will-Do (Total).

Analysis Procedures

Reducing Equations Across Criterion Constructs

For the analyses examining the reduction in prediction equations across criterion constructs, the data were analyzed using a variant of the Mosier (1951) double cross-validation design as follows. Generally, soldiers in each job were randomly split into two groups of equal size (plus or minus one soldier). Prediction composites developed using each of the criterion measures (i.e., CTP, GSP, ELS, MPD, PFB, Can-Do, Will-Do, and Total) in one group were used to predict each of the criterion measures in the other group (and vice versa).

More specifically, after the soldiers were divided into groups, covariance matrices (comprising the predictors and criterion measures described above) were computed in each group and corrected for multivariate range restriction (Lord & Novick, 1968, p. 147). These corrections were made using the covariances among the ASVAB subtests in the 1980 Youth Population (Mitchell & Hanser, 1984). Next, prediction equations were developed for each of the eight criteria, using the corrected covariance matrix in each group. The prediction equations developed in each group were then applied to the corrected covariance matrix of the other group to estimate the cross-validated correlation between each predictor composite and each of the five LVI criterion constructs. Finally, the results were averaged across the two groups within each job, and then averaged across jobs (after the results within each job were weighted by sample size).

In addition to analyzing the data separately by job, we also conducted a set of analyses using data that had been pooled across jobs. Specifically, the two covariance matrices per job described above were used to form two covariance matrices that had been pooled across jobs (i.e., each pooled covariance matrix was formed by pooling the data from one of the matrices computed for each job). These pooled matrices were then analyzed, using the same procedures described in the preceding paragraph, except that at the end it was not necessary to average the results across jobs. Actually, because soldiers in MOS 11B do not have scores for GSP, two sets of pooled covariance matrices were created and analyzed in order to retain 11B data where possible: One set included data from 11B but did not contain GSP in the covariance matrices, and the other set did not include data from 11B, but did contain GSP in the covariance matrices.

Reducing Equations Across Jobs

For the analyses examining the reduction in prediction equations across jobs, two sets of procedures were used. For the first set, a general linear model analysis was used to determine whether the predictor weights varied significantly across jobs. For the second set, an index of discriminant validity was used to estimate the extent to which predictive accuracy was improved when each job was allowed to have its own equation when predicting performance for a given criterion construct.

The analyses were conducted using two different subsets of the criterion measures. For the first set of predictors (ASVAB, spatial, and computer composites), a subset of five criteria was predicted: CTP, GSP, ELS, Can-Do, and Total). For the second set of predictors (ASVAB, ABLE and AVOICE composites), a second subset of criteria was created by including MPD, PFB, and Will-Do in place of CTP, GSP, and Can-Do. The criteria for each set were chosen because they were considered the most likely to be predicted by the predictors in those sets. Note that ELS was included in both sets because the LVI basic validation analyses indicated that it was predicted by predictors in both sets.

General Linear Model. For the general linear model analyses, deviation scores were created within job for all predictors and criteria. This was done to eliminate intercept differences across jobs which may have been caused by differences in selection requirements across jobs. The data were then pooled across jobs and a series of full and reduced linear models were estimated (one set per criterion for each predictor set). For the full models, regression weights were allowed to vary across job in the prediction of a given criterion measure; for the reduced models, regression weights were constrained to be equal across MOS. Finally, the multiple correlations associated with the full and reduced models were compared.

Discriminant Validity. For the discriminant validity analyses, raw data were used to compute a single covariance matrix for each job. These matrices were then corrected for multivariate range restriction (Lord & Novick, 1968, p. 147), using the covariances among the ASVAB subtests in the 1980 Youth Population (Mitchell & Hanser, 1984). For each predictor set-criterion combination (e.g., the predictor set with ASVAB, spatial, and computer composites and CTP), these matrices were then used to develop prediction

equations separately for each job. The multiple correlations associated with these equations were adjusted for shrinkage and averaged across jobs. This average is referred to here as the mean absolute validity.

Next, the prediction equation developed in each job was correlated with performance in all of the other jobs. These across-job correlations were also averaged (but were not adjusted for shrinkage, because they did not capitalize on chance). The mean of these correlations is referred to here as the mean generalizability validity.

Finally, discriminant validity was computed as the difference between the mean absolute validity and the mean generalizability validity. This index of discriminant validity represents an assessment of the extent to which there is intra-individual variation in predicted performance across jobs.

Results

The pooled results of using equations developed for each of the five criterion factors to estimate the validity of each equation for predicting scores on the other four criterion factors are shown in Tables 4.1 and 4.2 for two different predictor sets. Table 4.1 used the ASVAB + Spatial + Computerized measures and Table 4.2 used the ASVAB + ABLE + AVOICE.

When reading down the columns, the cross-validated correlation of the weighted predictor composite with its own criterion factor should be higher than the correlations of composites using weights developed on different criterion factors. This was the general result although the differences were not very large in some cases.

Based on the results of the within-MOS analysis, it was decided, for purposes of future analyses, to maintain a unique equation for each of the criterion factors.

Table 4.1
Estimates of Differential Prediction of Criterion Construct Scores for the ASVAB, Spatial, and Computerized Tests Predictor Set[a]

	Criterion Construct Scores				
Predicted Score	CTP	GSP	ELS	MPD	PFB
Core Technical Proficiency	.608	--	.365	.139	-.029
General Soldiering Proficiency	--	--	--	--	--
Effort and Leadership	.575	--	.373	.159	.019
Maintaining Personal Discipline	.402	--	-.290	.159	.043
Physical Fitness and Military Bearing	-.053	--	.055	.052	.137

Note. Mosier double cross-validation estimates, using pooled covariance matrices with MOS 11B included; no GSP scores are given because 11B does not have GSP scores.

[a] A total of 13 predictors.

Table 4.2
Estimates of Differential Prediction for Criterion Construct Scores for the ASVAB, ABLE, and AVOICE Predictor Set[a]

	Criterion Construct Scores				
	CTP	GSP	ELS	MPD	PFB
Predicted Score					
Core Technical Proficiency	.589	--	.368	.158	-.024
General Soldiering Proficiency	--	--	--	--	--
Effort and Leadership	.540	--	.386	.194	.076
Maintaining Personal Discipline	.331	--	-.275	.259	.095
Physical Fitness and Military Bearing	-.042	--	.089	.073	.294
Can-Do	.593	--	.368	.161	-.023
Will-Do	.371	--	.328	.220	.045
Total	.573	--	.387	.194	.045
Foldback correlations:					
Sample 1	.577	--	.421	.284	.356
Sample 2	.631	--	.400	.294	.332
Combined	.604	--	.403	.283	.335

Note. Mosier double cross-validation estimates, using pooled covariance matrices. MOS 11B data are included; no GSP scores are given because 11B does not have GSP scores.

[a] A total of 19 predictors.

For the comparisons of criterion prediction equations across jobs, the interpretation of the general linear model results was ambiguous because of the considerable disparity in degrees of freedom between the analysis of the full equation and the analysis of the reduced equations (e.g., 117 vs. 13 for predicting CTP). The difference between the adjusted and unadjusted estimates of R was so large that it swamped all the effects. The estimate of discriminant validity was approximately .03 for CTP, which constitutes only weak evidence for retaining a unique CTP prediction equation for each MOS. A grouping of the Batch A MOS into "job families" might have produced higher discriminant validity and reduced the number of unique equations from nine to three or four. However, for purposes of the current analysis, it was decided not to cluster MOS at this point, as using a unique equation for CTP for each MOS would constitute a benchmark for later analysis.

VALIDITY ESTIMATES FOR FULL AND REDUCED EQUATIONS

Twelve unique equations were identified in the previous analysis--that is, one equation for CTP in each MOS and one equation for all MOS for ELS, MPD, and PFB).

The objectives for this part of the analysis were to estimate the predictive validity, in the LVI sample, (a) of the full ASVAB + Experimental Battery predictor set (k = 28) for the 12 unique equations, and (b) of the same set of 12 unique equations after they were reduced in length with the goal of preserving either maximum selection efficiency or maximum classification efficiency.

Analysis Procedures

The Full Prediction Equations

For each of the 12 unique prediction situations, the appropriate covariance matrices, corrected for range restriction, were used to compute full least squares estimates of the multiple correlation between the full predictor battery (ASVAB plus EB) and the relevant criterion score. This estimate was adjusted using Rozeboom's (1978) formula 8. The correlations of the weighted composite of all predictors with the criterion were also computed for equal weights for all predictors and with the zero order validities used as weights. For unit weights and validity weights, no negative weights were used except for AVOICE. For the AVOICE subscores, if a particular composite had a negative correlation with the criterion the score was weighted negatively.

The validities for predicting Core Technical Proficiency were averaged across MOS. All validity estimates were corrected for attenuation in the criterion measure using the reliability estimates reported in Chapter 2.

Obtaining the Reduced Equations

The reduced equations were obtained via expert judgment, using a panel of three subject matter experts (the three authors of the present chapter). The task for the SMEs was to identify independently what they considered to be the optimal equations, in each of the 12 unique situations, for maximizing selection validities and the optimal equations for maximizing classification efficiency. The judgment task was constrained by stipulating in each case that the prediction equations (i.e., for selection and for classification) could contain no more than ten predictors. The judges were free to use fewer variables if they thought that a smaller number would reduce error while preserving relevant variance, or would improve classification efficiency without significantly reducing the overall level of selection validity.

The information used by the SMEs for the judgment task consisted of all prior LVI and CVI validation analyses and a factor analysis of the predictor battery, using a pooled 26 x 26 correlation matrix, which stipulated that the full set of 26 factors (unrotated) should be extracted. The factor scores were then correlated with pooled ELS, MPD, and PFB scores and with the CTP score in each MOS.

Each judge then identified two sets of 12 reduced equations. After looking at the results from the other SMEs, each revised his specifications. Remaining differences were eliminated via a series of discussions aimed at reaching a consensus.

The zero order validities, regression weights, and predictor battery validities were then recomputed for the two sets (selection vs. classification) of reduced equations.

Expert judgment was used instead of hierarchical regression or an empirical evaluation of all possible combinations because the latter was computationally prohibitive and the former runs the risk of too much sample idiosyncracy. In fact, there is no one optimal procedure for identifying such predictor batteries. For any procedure there is a tradeoff between adjusting for the capitalization on sample-specific chance factors and being able to maximize an unbiased estimate of the population validity.

Results

Predicting Core Technical Proficiency, Tables 4.3 and 4.4 (for selection and classification, respectively) show the results for the SME reduced equations in comparison to the results for the full equations, using multiple regression weights. The body of each table contains the recomputed standardized regression weights. Reading down each MOS column provides the list of the predictor variables included in the reduced equation for that MOS. Also shown are the foldback and adjusted (via the Rozeboom correction formula) multiple Rs for the reduced equations and the validity estimates, using zero order validity weights and unit weights. For comparison purposes, the multiple correlations using the full regression-weighted ASVAB plus Experimental Battery predictor set are shown at the bottom of Table 4.3.

The selection results for the reduced equations for the Will-Do factors -- ELS, MPD, and PFB -- are shown in Table 4.5.

In summary, a reduction of the 45 equations to a set of 12 (9 for CTP and one each for ELS, MPD, and PFB) was judged to be a conservative interpretation of the data and to be consistent with the conceptual interpretation of these variables. Consequently, all subsequent analyses in this chapter are focused on this subset of 12 unique prediction equations.

In terms of comparisons across MOS, the following points seem relevant. They are based on the overall pattern and relative magnitudes of the recomputed regression weights in Tables 4.3, 4.4, and 4.5.

Relative to selection validity for CTP, it is the Quantitative and Technical Knowledge factors on the ASVAB that make the greatest contribution. Overall, among the ASVAB factor scores, they yield the largest and most frequent regression weights. However, Perceptual Speed and Verbal do seem to make a contribution to potential classification efficiency, as judged by the SMEs. They were selected as predictors and had substantial regression coefficients for only a few MOS.

The spatial composite from the Experimental Battery was a uniformly strong contributor to selection validity and provided relatively little potential classification efficiency, except for distinguishing 71L and 91A from all other MOS.

Table 4.3
SME Reduced (Optimal) Equations for Maximizing Selection Efficiency for Predicting Core Technical Proficiency in LVI

Predictor	Standardized Regression Weights by MOS								
	11B (235)	13B (551)	19K (445)	31C (172)	63B (406)	71L (251)	88M (221)	91A (535)	95B (270)
ASVAB									
Quantitative	.079	.220	.128	.138	.098	.326	.198	.176	--
Speed	.135	--	--	.073	.047	.101	--	.130	-.041
Technical	.317	.115	.124	.222	.332	--	.193	.208	--
Verbal	.061	.049	.129	.016	.050	.294	--	.100	.408
Spatial	.196	.275	.308	.242	.198	.155	.264	.186	.388
Computer									
Movement Time	.136	--	--	--	.044	.059	--	--	.059
Number Speed/Accuracy	.094	--	--	--	--	--	--	--	--
Percept. Accuracy	--	.040	--	--	--	.079	--	.059	--
Percept. Speed	--	--	--	--	--	--	--	--	--
Psychomotor	--	.034	--	--	--	--	.065	-.021	--
Short-Term Memory	--	--	--	.029	.106	.010	.138	.092	.125
Basic Speed	--	--	--	.066	--	--	--	--	--
Basic Accuracy	--	--	--	--	--	--	--	--	.086
ABLE									
Achievement	.002	--	.017	--	.014	-.058	--	--	.002
Adjustment	-.000	-.046	-.036	--	--	--	--	--	--
Physical Condition	--	--	--	--	--	--	--	--	--
Internal Control	--	.046	--	--	--	--	--	--	.059
Cooperativeness	--	--	--	--	--	--	.069	--	--
Dependability	--	.106	.079	.133	.073	--	.070	.040	-.070
Leadership	--	--	--	--	--	--	--	--	--
AVOICE									
Administrative	--	--	--	--	--	--	--	--	--
Audiovisual Arts	--	--	--	--	--	.102	--	--	--
Food Service	--	--	--	-.156	--	--	-.179	--	--
Struct./Machine	--	--	--	--	.007	--	.262	--	--
Protective Services	--	--	--	--	--	--	--	--	--
Rugged/Outdoors	.010	-.008	.056	.044	--	--	-.081	.077	.041
Social	--	--	--	--	--	--	--	--	--
Skill/Technical	--	--	--	--	--	.031	--	--	--
Reduced Equations									
Foldback R	.789	.642	.637	.712	.710	.845	.719	.763	.836
Adjusted R	.765	.624	.617	.663	.691	.829	.683	.751	.820
Validity Weights	.773	.625	.626	.693	.691	.827	.687	.756	.798
Unit Weights	.737	.580	.556	.691	.651	.757	.632	.741	.665
Full Equations									
Foldback R	.810	.663	.659	.768	.722	.861	.745	.777	.853
Adjusted R	.747	.614	.596	.650	.670	.821	.649	.748	.813

Note. Dashes indicate that this predictor variable was not included in the reduced equation for this MOS.

Table 4.4
SME Reduced (Optimal) Equations for Maximizing Classification Efficiency for Predicting Core Technical Proficiency in LVI

Predictor	Standarized Regression Weights by MOS								
	11B (235)	13B (551)	19K (445)	31C (172)	63B (400)	71L (251)	88M (221)	91A (535)	95B (270)
ASVAB									
Quantitative	--	.279	--	--	.136	.480	--	.395	--
Speed	.154	.024	--	--	--	--	--	--	--
Technical	.353	--	--.228	.248	.354	--	.234	.274	.213
Verbal	--	--	--	--	--	.360	--	--	.298
Spatial	.214	.327	.406	.285	.209	--	.331	--	.297
Computer									
Movement Time	--.121	--	--	--	.046	.089	--	--	.049
Number Speed/Accuracy	.123	--	--	.153	--	-.046	--	--	--
Percept. Accuracy	--	.039	--	--	--	.104	.070	.093	--
Percept. Speed	--	--	.032	--	--	--	.035	--	--
Psychomotor	--	.048	--	--	--	--	.024	--	--
Short-Term Memory	--	--	--	.029	.116	--	.147	.150	.129
Basic Speed	--	--	--	.067	--	--	--	--	.026
Basic Accuracy	--	---	--	--	--	--	--	--	.083
ABLE									
Achievement	--	--	.014	--	--	--	--	--.031	-.033
Adjustment	--	--	--	--	--	--	--	--	--
Physical Condition	--	--	--	--	--	--	--	--	--
Internal Control	.041	.039	--	--	--	--	--	--	.047
Cooperativeness	--	--	--	--	--	--	.110	--	--
Dependability	--	.086	.079	.143	.074	--	--	.050	--
Leadership	--	--	--	--	--	--	--	--	--
AVOICE									
Administrative	-.101	--	--	--	--	--	--	--	-.006
Audiovisual Arts	--	--	--	--	--	.127	--	--	--
Food Service	.030	--	--	-.154	--	--	-.159	--	-.005
Struct./Machine	--	.038	--	--	--	--	.185	--	--
Protective Services	--	--	--	--	--	--	--	--	--
Rugged/Outdoors	--	--	.021	.030	--	--	--	.094	--
Social	--	--	--	--	--	-.037	--	--	--
Skill/Technical	.098	--	--	--	.041	--	--	-.103	--
Reduced Equations									
Foldback R	.792	.636	.622	.710	.710	.835	.707	.750	.842
Adjusted R	.770	.621	.606	.670	.695	.823	.672	.740	.826
Validity Weights	.776	.614	.605	.696	.702	.811	.674	.737	.825
Unit Weights	.744	.568	.521	.675	.652	.750	.613	.585	.749

Note. Dashes indicate that this predictor variable was not included in the reduced equation for this MOS.

Table 4.5
SME Reduced (Optimal) Equations for Maximizing Selection Efficiency for Predicting Will-Do Criterion Factors

Predictor	Regression Weights by Criterion Factors		
	ELS (3,086)	MPD (3,086)	PFB (3,086)
ASVAB			
Quantitative	.059	.073	-.010
Speed	.075	--	--
Technical	.189	.088	--
Verbal	--	--	--
Spatial	--	--	--
Computer			
Movement Time	.050	--	.052
Number Speed/Accuracy	--	--	--
Percept. Accuracy	--	--	--
Percept. Speed	--	--	--
Psychomotor	--	--	--
Short-Term Memory	.056	.040	--
Basic Speed	--	--	--
Basic Accuracy	--	---	--
ABLE			
Achievement	--	--	.064
Adjustment	--	--	--
Physical Condition	--	--	.245
Internal Control	--	-.086	--
Cooperativeness	--	--	--
Dependability	-.088	.207	.049
Leadership	.047	--	--
AVOICE			
Administrative	--	--	--
Audiovisual Arts	--	--	--
Food Service	--	--	--
Struct./Machine	--	--	--
Protective Services	--	--	--
Rugged/Outdoors	--	--	--
Social	--	--	--
Skill/Technical	--	--	--
Reduced Equations			
Foldback R	.354	.232	.310
Adjusted R	.347	.224	.304
Validity Weights	.346	.202	.297
Unit Weights	.341	.187	.252
Full Equations			
Foldback R	.372	.277	.346
Adjusted R	.349	.243	.321

Note. Dashes indicate that this predictor variable was not included in the reduced variable.

Among the computerized measures, Movement Time and Short-Term Memory were judged to make the most consistent contribution to selection validity (they were selected for the greatest number of MOS in Table 4.3) while Perceptual Speed, Psychomotor Ability, and the accuracy scores seem to make the greatest contribution (although small) to classification efficiency, as indicated by the specificity with which they were assigned to MOS in Table 4.4.

For the ABLE, it is the Dependability scale that is judged to make the greatest contribution to selection validity as indicated in Table 4.3. The ABLE was seen as contributing very little to potential classification efficiency (Table 4.4).

For the AVOICE, the Rugged Outdoor scale was selected as making the most consistent contribution to selection validity (Table 4.3), although the recomputed regression weights are not large. The AVOICE seems to have considerable potential for making a contribution to classification efficiency; the pattern of weights in Table 4.4 is very distinctive and seems to be consistent with the task content of the respective MOS.

The reduced equations for ELS, MPD, and PFB in Table 4.5 (which were obtained only for the goal of maximizing selection validity) show what are perhaps the expected differences in the equations. ASVAB is much more important for predicting ELS than for predicting MPD and PFB. The pattern of ABLE weights is consistent with expectations and the AVOICE is judged to contribute virtually nothing to the prediction of these three factors.

The mean results across MOS for predicting CTP are shown is Table 4.6. In general, differential predictor weights do provide some incremental validity over unit weights. However, zero-order validity coefficients as weights are virtually as good as regression weights and the reduced equations yield about the same level of predictive accuracy as the full equations. In fact, the reduced equations do slightly better.

Perhaps the most striking feature in Table 4.6 is the overall level of the correlations. The validities are very high. The best available estimate of the validity of the Project A/Career Force predictor battery for predicting Core Technical Proficiency is contained in the last column of the table which is the adjusted multiple R corrected for unreliability in the criterion (i.e., CTP in LVI). The estimated validities (averaged over MOS) in this column are .78 ± .01. The reduced equations produce this level of accuracy, which does break the so-called validity ceiling, just as readily as the full equation, with perhaps more potential for producing classification efficiency. The analyses that attempt to estimate the actual gains in classification efficiency are reported in Chapter 8.

Table 4.6
Estimates of Maximizing Selection Efficiency Aggregated Over MOS: Predicting Core Technical Proficiency

	Mean Selection Validity				
	Unit Weights	Validity Weights	Foldback R	Adjusted R	Corrected R[a]
Full Equation: All predictors	.576	.697	.762	.701	(.772)
Reduced Equation: Selection	.668	.720	.739	.716	(.789)
Reduced Equation: Classification	.651	.716	.734	.714	(.786)

[a] Corrected for criterion unreliabilty.

SUMMARY

Consistent with the results from CVI, the analysis of the LVI data for the Batch A MOS indicated differential prediction across performance factors within MOS and differential prediction across MOS for the CTP factor but not for the other performance factors.

For this set of 12 unique equations. the best estimate of the population value for selection validity is the observed correlation corrected for restriction of range, unreliability in the criterion, and the fitting of error by the predictor weighting procedure. The resulting estimates of the population validity are very high, both for the equations using all predictors and for the equations using a reduced set of predictors ($k < 10$).

Predictor sets that were selected to maximize classification efficiency also yield high selection validity. To the extent that it exists, differential validity across MOS is judged to be a function of the ASVAB, the computerized perceptual and psychomotor measures, and the AVOICE. Estimates of actual classification gains for the various equations will be presented in a subsequent chapter of this report (Chapter 8).

Chapter 5
DIFFERENTIAL PREDICTION ACROSS RACIAL AND GENDER GROUPS: OPTIMAL BATTERY ANALYSES

Teresa L. Russell, Norman G. Peterson, Scott H. Oppler, and John P. Campbell

This chapter describes differential prediction analyses conducted to investigate differences between groups in the Project A/Career Force sample. Cleary's (1968) psychometric model of differential prediction is currently accepted by both the Uniform Guidelines (EEOC, 1978) and the Society for Industrial and Organizational Psychology (SIOP, 1987). The Cleary model defines bias in terms of prediction systems: "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup" (Cleary, 1968, p. 115).

This procedure involves comparing variances of errors of prediction, slopes of regression lines, and intercepts of regression lines for the two subgroups. Technically, test bias occurs when there is a difference in any one of the parameters, but operationally, differences occur more frequently for intercepts than for the other two parameters. Intercept differences yield either over- or underprediction of subgroup performance--differential prediction.

Overprediction of the performance of a protected group, when a common regression line is used, fits the model's definition of bias but is generally not considered a problem in the fair use of a test (SIOP, 1987).[1] Based on a review of differential prediction research, SIOP concluded that "there is little evidence to suggest that there is differential prediction for the sexes, and the literature indicates that differential prediction on the basis of cognitive tests is not supported for the major ethnic groups" (p. 18).

Other studies, however, suggest that differential prediction and fairness issues are not so easily discarded. Linn (1994) reviewed several studies citing evidence of differential prediction, particularly underprediction of female performance (e.g., Dunbar & Novick, 1988; Gamache & Novick, 1985; Houston & Novick, 1987; Valentine, 1977). That led him to suggest that SIOP's dismissal of differential prediction was premature.

The goal of the differential prediction analyses presented in this chapter was to explore group differences in prediction systems for predictor scores formed by three different methods:

(1) Three predictor batteries that have "traditionally" been used in Project A/Career Force validation analyses

[1] Antidiscrimination legislation prohibits discrimination against "protected" groups, including women and minorities.

(2) An "optimal" battery designed to maximize selection validity

(3) An "optimal" battery designed to maximize classification efficiency

The traditional predictor batteries included three predictor sets: (a) the ASVAB only, (b) the ASVAB along with spatial and computer composites, and (c) the ASVAB combined with the noncognitive composites from the ABLE, AVOICE, and JOB. The optimal selection and classification predictor batteries were composed of predictors identified for each MOS using the procedures described in Chapter 4. Expert judgment was used to reduce the total number of predictors (i.e., ASVAB + the full Experimental Battery) to two batteries of $K \leq 10$. One of the reduced batteries was intended to maximize selection validity and the other was intended to maximize classification efficiency.

PROCEDURES

The overall approach was to compare the bivariate distributions for each subgroup for a selected set of predictor battery/criterion relationships. For each predictor/criterion relationship, the two subgroups were compared on a number of regression parameters.

Sample

Differential prediction analyses require sufficient sample sizes for each subgroup. We chose to include in the analyses all MOS with subgroups of 50, at a minimum. Six MOS had large enough sample sizes for White-Black comparisons, and four MOS contained sufficient numbers of males and females for those analyses. Sample sizes for each selected MOS are listed in Table 5.1.

Regression Model Analyses

The regression analyses involved four principal steps:

(1) Compute least squares weighted composite scores (predicted performance scores) for individuals based on the regression of the predictor composites (e.g., the four ASVAB composites) against the criterion variable (e.g., Core Technical Proficiency) for the total group (e.g., all MOS 13Bs).

(2) Regress the least squares weighted composite scores against the criterion for each subgroup.

(3) Compute F-tests for regression slope homogeneity and for intercept differences.

(4) Graph the regression results.

Table 5.1
Subgroup Sample Sizes for Differential Prediction Analyses

MOS		Whites	Blacks
13B	Cannon Crewmember	263	218
19K	M1 Armor Crewman	350	71
63B	Light Wheel Vehicle Mechanic	282	90
71L	Administrative Specialist	115	113
88M	Motor Transport Operator	145	66
91A	Medical Specialist	362	116
		Males	Females
71L	Administrative Specialist	57	194
88M	Motor Transport Operator	168	53
91A	Medical Specialist	451	84
95B	Military Police	219	51

The steps were repeated for four criteria: Core Technical Proficiency (CTP), Effort and Leadership (ELS), Maintaining Personal Discipline (MPD), and Physical Fitness and Bearing (PFB). CTP analyses were conducted by MOS because the CTP criterion contains MOS-specific measures. MOS were pooled for ELS, MPD, and PFB analyses because those criteria are based on variables that are common across MOS.

As an example, Figure 5.1 shows the results of White-Black differential prediction analyses for the ASVAB composites against CTP for MOS 88M. It provides a number of pieces of information:

- The effect size for both the test score difference and the criterion score difference. The White test score mean was 1.2 SD higher than the Black test score mean, and the White criterion score mean was .6 SD higher than the Black criterion score mean.

- The p value for the level of significance of the difference between (a) subgroup slopes and (b) intercepts. Both p values are based on the Type I Sum of Squares (Cohen & Cohen, 1983, p. 145). Neither the slope nor the intercept differences are significant (at the .05 level) in the example shown in Figure 5.1.

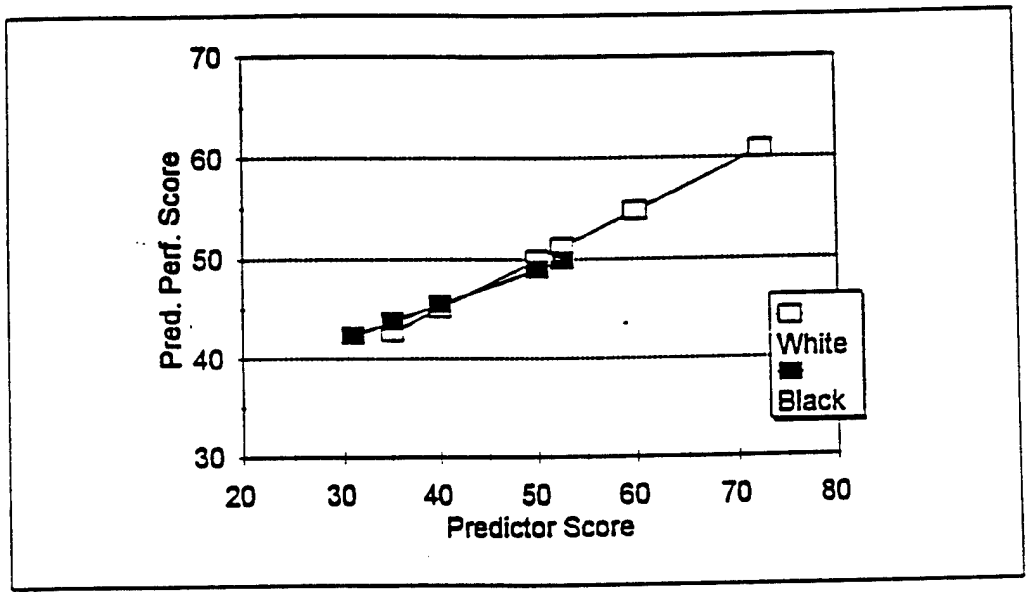
- Predictor scores for key points in the graph.

Fairness Analyses — **White/Black Subgroups**
MOS = 88M — Composite: 4 ASVAB Factors

Criterion: Core Technical Proficiency (Raw Score)

Group	N	Test MN	Test SD	Criterion MN	Criterion SD	Slope	Intercept	R-Square	R
Total	211	50.000	10.000	50.000	10.000	0.470	26.489	0.221	0.470
White	145	53.745	9.348	51.773	9.790	0.487	25.587	0.216	0.465
Black	66	41.773	5.427	46.104	9.397	0.347	31.608	0.040	0.200
Effect Size		-1.197		-0.567					
P value						0.521	0.971		

Predictor Score	Predicted Performance Score		Standard Error		Score Difference	
	White	Black	White	Black	Value	Under/Over
30.919	NA	42.337	NA	2.534	NA	NA
35.049	42.656	43.770	1.610	1.804	-1.114	Under
40.000	45.067	45.488	1.280	1.192	-0.421	Under
50.000	49.937	48.958	0.775	2.058	0.979	Over
52.627	51.216	49.870	0.725	2.534	1.347	Over
60.000	54.807	NA	0.866	NA	NA	NA
72.441	60.866	NA	1.610	NA	NA	NA



Notes.

Effect Size — The difference between subgroup means in standard deviation units. Negative values indicate high male (or, for black/white comparisons, white) means.

Predictor Score — The seven predictor scores listed are: (1) the majority test mean minus two standard deviations, (2) the minority test mean minus two standard deviations, (3) the total group test mean minus one standard deviation (i.e., 40), (4) the total group mean (i.e., 50), (5) the total group mean plus one standard deviation (i.e., 60), (6) the minority test mean plus two standard deviations and, (7) the majority test mean plus two standard deviations. The values are ordered from lowest to highest, without regard to subgroup.

Figure 5.1. White-Black differential prediction analysis for MOS 88M.

- Predicted performance scores [i.e., the subgroup intercept + (the subgroup slope*the predictor score)] for the same key points.

- The standard error of the predicted performance score (Hays, 1963, p. 522, formula 15.22.3).

- The value of the difference between subgroup predicted performance scores. "Under" indicates that performance of the protected group would be underpredicted by the White group line. "Over" indicates overprediction when White and Black lines are compared.

RESULTS

Recall that three different types of predictor scores were used--the traditional regression-weighted basic scores that the project has used in all of its basic validation analyses, and two regression-weighted composite scores obtained from prediction equations, one designed to optimize selection and one designed to optimize classification. The traditional basic scores included three predictor sets--the ASVAB only, the ASVAB with spatial and computer composites, and the ASVAB with the noncognitive composites from the ABLE, AVOICE, and JOB. Optimal selection and classification scores were composed of predictors identified for each MOS through the optimal battery analyses. Development of both sets of optimal weights was described in Chapter 4.

Traditional Basic Scores

The results of the differential prediction analyses based on the traditional composite scores are summarized in Tables 5.2 and 5.3. Table 5.2 shows the White-Black and Male-Female results for the Core Technical Proficiency criterion, and Table 5.3 shows the results for the three "will do" criteria (i.e., ELS, MPD, PFB). The symbols in the cells of the table indicate significant ($p < .05$) slope and intercept differences and over- or underprediction. An "s" indicates significant slope differences; "i" indicates significant intercept differences; "u" and "o" indicate under- or overprediction (respectively). A "u" is assigned when there is underprediction at two or more of the following predictor scores: (a) the total group mean, (b) the total group mean plus one standard deviation, and (c) the total group mean minus one standard deviation. An "o" indicates overprediction at two or more of the three points.

With regard to Core Technical Proficiency (Table 5.2), only one out of 18 White-Black comparisons yielded a significant difference. That is, for the four ASVAB factor composites, there was a significant intercept difference leading to overprediction of Black performance for MOS 91A. When the noncognitive variables are included in the analysis, the prediction systems are not significantly different even for 91A.

Table 5.2
Slope and Intercept Comparisons for Traditional Predictor Composites Against Core Technical Proficiency for White-Black and Male-Female Groups

MOS	Composite: ASVAB (4) Only	Composite: ASVAB (4) + Spatial (1) + Computer (8)	Composite: ASVAB (4) + ABLE (7) + AVOICE (8) + JOB (3)
White-Black			
13B	--	--	--
19K	--	--	--
63B	--	--	--
71L	--	--	--
88M	--	--	--
91A	i,o	--	--
Male-Female			
71L	i,u	--	--
88M	--	--	--
91A	i,u	i,u	i,u
95B	s,u	--	s,u

Note. "s" = slope; "i" = intercept; "u" = underprediction; "o" = overprediction. Numbers in parenthesis indicate the number of scores in the composite.

Table 5.3
Slope and Intercept Comparisons for Traditional Predictor Composites Against Will-Do Criteria[a] for White-Black and Male-Female Groups

Criterion	Composite: ASVAB (4) Only	Composite: ASVAB (4) + Spatial (1) + Computer (8)	Composite: ASVAB (4) + ABLE (7) + AVOICE (8) + JOB (3)
Effort and Leadership (ELS)			
White-Black	s,u	s,u	--
Male-Female	--	--	--
Maintaining Personal Discipline (MPD)			
White-Black	s,o	s,o	--
Male-Female	--	--	--
Physical Fitness and Bearing (PFB)			
White-Black	i,u	s,i,u	i,u
Male-Female	i,o	i,o	i,o

Note. "s" = slope; "i" = intercept; "u" = underprediction; "o" = overprediction. Numbers in parenthesis indicate the number of scores in the composite.

[a] Effort and Leadership, Maintaining Personal Discipline, and Physical Fitness and Bearing.

Six out of 12 Male-Female comparisons yielded significant differences (two slope and four intercept), all leading to underprediction of Female performance. It is important to note here that MOS 71L and 91A females scored higher on the criterion than their male counterparts did (and lower than males on most predictors). Inclusion of non-cognitive measures alleviated the Male-Female prediction differences for 71L but not for 91A.

As noted earlier, for the three Will-Do criteria, analyses were conducted for the full sample across MOS (Table 5.3). In predicting ELS and MPD, White-Black slope differences were significant when the ASVAB alone or the ASVAB-spatial-computer composites were used. However, inclusion of non-cognitive measures alleviated Black-White prediction differences on those variables.

The differences on PFB are consistent and logical. Mean scores for Blacks are considerably higher than those for Whites on the PFB criterion (i.e., about one-third of an SD higher). Consequently, Black performance on PFB was underpredicted by all three composites. Female performance on PFB was overpredicted by all three composites.

Optimal Selection and Optimal Classification Predictor Composites

With regard to CTP, for the optimal selection composites, none of the six White-Black comparisons yielded a significant difference (Table 5.4). For the optimal classification composites, one of the six White-Black comparisons was significant (i.e., a significant intercept difference leading to overprediction of Black performance for MOS 91A).

Both equations yielded a significant intercept difference underpredicting female CTP performance for both MOS 71L and 91A. Again, females did score higher on the CTP criterion than males in those two MOS.

In the Will-Do criteria comparisons, as shown in Table 5.5, there were no significant White-Black or Male-Female prediction differences on ELS and MPD. As in previous analyses (Table 5.3), Black performance on PFB was underpredicted by both composites.

CONCLUSIONS

Three main points can be drawn from the fairness analyses:

- Predictive differences between Whites and Blacks were observed infrequently. Out of the many White-Black comparisons, very few suggested significant predictive differences between Whites and Blacks. For Core Technical Proficiency, when significant differences arose they yielded overprediction of Black performance, which is not considered unfair test use (SIOP, 1987).

Table 5.4
Slope and Intercept Comparisons for Optimal Selection and Classification Composites Against Core Technical Proficiency for White-Black and Male-Female Groups

	Composite	
MOS	Optimal Selection	Optimal Classification
White-Black		
13B	--	--
19K	--	--
63B	--	--
71L	--	--
88M	--	--
91A	--	i,o
Male-Female		
71L	i,u	i,u
88M	--	--
91A	i,u	i,u
95B	--	--

Note. "i" = intercept; "u" = underprediction; "o" = overprediction.

Table 5.5
Slope and Intercept Comparisons for Optimal Selection and Classification Composites Against Will-Do Criteria for White-Black and Male-Female Groups

	Composite	
Criterion	Optimal Selection	Optimal Classification
Effort and Leadership (ELS)		
White-Black	--	--
Male-Female	--	--
Maintaining Personal Discipline (MPD)		
White-Black	--	--
Male-Female	--	--
Physical Fitness and Bearing (PFB)		
White-Black	i,u	i,u
Male-Female	--	--

Note. "i" = intercept; "u" = underprediction.

- Importantly, White-Black differences were often alleviated when non-cognitive measures were added to the composite.

- Underprediction of female performance in two jobs was salient across a variety of weighting strategies. Females in those two MOS scored higher than males on the Core Technical Proficiency criterion and lower than males on most predictors. That situation virtually dictates underprediction. It is also important to note that underprediction of female performance has been observed in other studies (e.g., Dunbar & Novick, 1988; Gamache & Novick, 1985; Houston & Novick, 1987).

Chapter 6
THE SEQUENTIAL PREDICTION OF INDIVIDUAL JOB PERFORMANCE: THE "ROLLUP"

Rodney A. McCloy

The Project A/Career Force research projects have provided the opportunity to study the determinants of individual soldier performance from the time of enlistment through the second tour of duty. A primary goal was to develop measures of soldier job performance at several points in each individual's career. To this end, Project A/Career Force assessed performance (a) at the end of training, (b) toward the end of the first tour of duty, and (c) toward the end of the second tour of duty.

Taken one at a time, the performance assessments constitute snapshots of the performance of a cohort of soldiers in selected MOS at specific points in time. Taken together, the measures provide a performance history for each soldier in the Career Force database. When the goal is to predict future job performance, these longitudinal data make it possible to assess the cumulative usefulness of the information collected prior to enlistment (e.g., ASVAB test scores) and the information gathered sequentially throughout a soldier's career. For the Project A/Career Force database, this means the sequential summation of the ASVAB, the Project A Experimental Battery, measures of performance in training, and performance as a job incumbent in the first tour for the purposes of predicting second-tour performance. This sequential combination of information is termed the "rollup" analysis.

The analyses described in this chapter demonstrate the degree to which incrementally "rolling up" the antecedent information into subsequently richer predictor sets results in more accurate predictions of future performance. In addition, the results presented below suggest which pieces of information are best for predicting various dimensions of job performance at each stage.

METHOD

Actually, two "rollups" are analyzed. One is the rollup to first-tour performance and the other is the rollup to second-tour performance.

The Samples

Soldiers were selected from the Career Force predictor database (N = 33,000). To be included in the LVI rollup analyses, soldiers were required to have complete data on (a) the four ASVAB composites, (b) nine Experimental Battery predictors, (c) the "can-do" and "will-do" end-of-training criterion composites, and (d) four LVI criteria. For the LVII rollup analyses, soldiers were further required to have complete data on five LVII criteria.

Listwise deletion (as explained in the Year Two summary in Chapter 1) was favored over pairwise deletion for all analyses because of the possibility of ill-conditioned covariance matrices (e.g., not positive definite). When the cell sample sizes vary greatly throughout a pairwise covariance matrix (as they would have in these analyses), the potential for ill-conditioned matrices increases. In light of the multivariate adjustment applied to the primary covariance matrix (i.e., correction for range restriction), the loss of sample size was considered less detrimental to the analyses than the possibility of a poor covariance structure. Only a subset of Experimental Battery predictors was included so that the predictor/sample size ratio remained reasonable for the LVII analyses. The variables used in the rollup analyses are presented in Table 6.1.

Both the LVI and LVII rollup analysis samples used Batch A MOS. However, two MOS did not appear in the LVII analyses--19E (too few soldiers) and 31C (no LVII criterion scores). Sample sizes for each of the MOS constituting the LVI and LVII analytic samples are given in Table 6.2.

Analysis Procedure

The basic analytic approach was to calculate the multiple correlations for a selected set of hierarchical regression models. The correlations reflect (a) correction for range restriction using the procedure developed by Lawley (1943) and described by Lord and Novick (1968, p. 147), and (b) adjustment for shrinkage using Rozeboom's formula 8 (1978, p. 1350).[1]

Because the measures of interest are ordered temporally (ASVAB, Experimental Battery, EOT criteria, LVI criteria, LVII criteria), models adding information over time were logical choices for analysis. Other models of substantive interest, but not conforming to the strict temporal ordering of predictors, were also examined.

Computation of the Aggregate Covariance Matrix

To compute the desired multiple correlations, the first step was to generate the relevant covariance matrix aggregated across MOS and corrected for range restriction. Two major issues were addressed during this step: (a) when the correction for range restriction was to be applied to the covariance matrix, and (b) what the target population matrix should be for the correction.

---

[1] Rozeboom's adjusted R^2 is

$$R^2_{adj} = 1 - \left(\frac{N+k}{N-k}\right)(1 - R^2_{yx})$$

where N is the size of the sample used to estimate the prediction equation, k is the number of predictors, and R^2_{yx} is the sample coefficient of determination.

Table 6.1
Variables Used in the Rollup Analyses

LVI Rollup Analyses

Predictors	Criteria
ASVAB	
Quantitative	Core Technical Proficiency (CTP)
Speed	Can-Do Composite (CTP + GSP)
Technical	Effort and Leadership (ELS)
Verbal	Will-Do Composite (ELS + MPD + PFB)
Experimental Battery	
Spatial	
Rugged/Outdoors Interests (AVOICE)	
Achievement Orientation (ABLE)	
Adjustment (ABLE)	
Physical Condition (ABLE)	
Cooperativeness (ABLE)	
Internal Control (ABLE)	
Dependability (ABLE)	
Leadership (ABLE)	
End of Training	
"Can-do" (Basic + Technical)	
"Will-do" (ETS + MPD + PFB + LDR)	

LVII Rollup Analyses

Predictors	Criteria
All predictors and all criteria from the LVI analyses (except LVI CTP and ELS)	Core Technical Proficiency
	Can-Do Composite (CTP + GSP)
	Leadership
	Effort/Achievement
	Will-Do Composite (LDR + EA + MPD + PFB)

Note. CTP = Core Technical Proficiency; GSP = General Soldiering Proficiency; ELS = Effort and Leadership; MPD = Maintaining Personal Discipline; PFB = Physical Fitness and Military Bearing; LDR = Leadership; EA = Effort and Achievement; ETS = Effort and Technical Skill.

Table 6.2
Sample Size for Rollup Analyses by MOS

MOS	Name	Sample Size	
		LVI	LVII
11B	Infantryman	486	23
13B	Cannon Crewmember	591	20
19E	M60 Armor Crewman	102	--
19K	M1 Armor Crewman	400	17
31C	Single Channel Radio Operator	117	--
63B	Light Wheel Vehicle Mechanic	206	10
64C	Motor Transport Operator	204	8
71L	Administrative Specialist	214	12
91A	Medical Specialist	495	31
95B	Military Police	258	9
Total		3,073	130

Because the database contains several MOS (10 for the LVI analyses and 8 for LVII), there were two procedures that could have been followed. One approach would be to (a) obtain a covariance matrix for each MOS, (b) weight each by its sample size, (c) average them to form a single average covariance matrix, and (d) correct this average covariance matrix for range restriction. An alternative approach would be to (a) obtain a covariance matrix for each MOS, (b) correct each MOS-specific matrix for range restriction, (c) weight each corrected covariance matrix by its sample size, and (d) average the corrected matrices to form the final covariance matrix.

Given adequate sample sizes, the second alternative would be preferred because there is arguably differential range restriction across MOS. Table 6.2, however, highlights the difficulties with this approach for the LVII analyses. Although calculating covariance matrices on single-digit sample sizes is questionable, we retained any MOS having more than five soldiers to maximize the total sample size. However, applying the range restriction correction formula to these MOS matrices would be totally inappropriate. For this reason, the correction for range restriction was applied only after averaging the MOS-specific matrices in both the LVI and LVII analyses. However, it should be kept in mind that, even for the total sample, the confidence intervals for R will be relatively large. The standard error is approximately 0.09.

The second issue concerned what data should be used as the basis for the population matrix in the correction for range restriction. When predicting LVI performance, the population matrix could be based upon (a) the applicant population

(i.e., correlations among the ASVAB subtests from the 1980 youth population, Mitchell & Hanser, 1982), or (b) soldiers completing AIT (i.e., correlations among the end-of-training measures). The two target matrices would result in different amounts of correction to the observed covariances because the amount of range restriction from applicants to LVI soldiers is greater than the amount of range restriction from AIT soldiers to LVI soldiers. Not all soldiers who enlist make it through AIT. Therefore, the range of scores on ASVAB and other predictors for the AIT soldiers is restricted relative to that of the applicant population. Correction to the youth population matrix is appropriate if the decision of primary interest is the selection of new accessions from the applicant pool. Therefore, for the LVI analyses, the youth population matrix served as the target for the range restriction correction.

For the LVII analyses, in addition to the restriction occurring at selection and at the end of AIT, there is further restriction regarding the number of soldiers who (a) qualify for reenlistment and (b) actually decide to reenlist. The distributions of the performance criteria (both EOT and LVI) for LVII soldiers are affected, because only the better performers are eligible for reenlistment. Further, self-selection due to the decision to reenlist further restricts the range of ASVAB and Experimental Battery score distributions.

For the prediction of LVII performance, the procedure that was used to correct for range restriction is a function of the specific prediction, or personnel decision, that is being made, which in turn governs how the population parameter to be estimated is defined. There are two principal possiblities. First, we could be interested in predicting second-tour performance at the time of hire. In this case the referent population would be the applicant sample. Second, we could be interested in the reenlistment decision and in predicting second-tour performance from information available during an individual's first tour. In this case, the referent population would be all first-tour job incumbents.

Consequently, for the LVII rollup analyses, a covariance matrix containing the ASVAB composites, the selected Experimental Battery predictors (see Table 6.1), LVI Can-Do and Will-Do composites, and the five basic LVI criteria served as the target matrix. The EOT measures were treated as incidental, rather than explicit, selection variables. The matrix was calculated using scores from all LVI soldiers having complete data on the specified measures (N = 3,702). This sample size is larger than the total sample size given in Table 6.2 for the LVI analyses because the EOT measures were not included in the listwise deletion of variables for the target matrix.

Alternative Rollup Models

As mentioned above, the temporal ordering of the measures used in the rollup analyses suggested certain hierarchical regression models. Hierarchical predictor sets allow investigation of the incremental validity provided by additional information. Variables were rolled up both forward (ASVAB → Experimental Battery → EOT → LVI [when predicting LVII]) and backward (LVI [when predicting LVII] → . . . → ASVAB).

Other models having breaks in the temporal order of predictors but of substantive interest (e.g., ASVAB + EOT when predicting LVI criteria) were also examined. The LVI and LVII regression models that were analyzed are provided in Table 6.3.

Note that only the can-do and will-do criterion composites from the EOT and LVI performance models were used as predictors. In addition, only one of the composites was used in any given prediction equation, depending upon whether the criterion to be predicted was a can-do or will-do measure.

RESULTS

The results for LVI and LVII will be discussed in turn. The reader should keep in mind that (a) the sample size for the LVII analysis is much smaller than for LVI, and (b) the decision being modeled in LVII is selection for reenlistment whereas the decision for LVI is selection into the Army.

LVI Analyses

The results of the LVI analyses are presented in Table 6.4. Both unadjusted and adjusted multiple correlations are reported. Underlined values represent multiple correlations that have been corrected for attenuation due to criterion unreliability. The values discussed below will be those adjusted for shrinkage and corrected for attenuation, unless otherwise specified.

There are three primary findings from Table 6.4. First, additional predictive validity can be obtained from sources other than the ASVAB. For example, EOT information increments ASVAB predictive validity for all criteria, with ASVAB and EOT constituting the best two-group predictor battery (especially for will-do criteria). Similarly, the Experimental Battery provides modest gains over the ASVAB in the adjusted multiple correlation for all four criteria, with increases ranging from .02 (for CTP and ELS) to .06 (for Will-Do). Further, the addition of Experimental Battery information increments the predictive validity of the available predictor information (i.e., ASVAB and EOT) for all criteria except ELS. For example, Table 6.4 demonstrates a 5-point increase in the adjusted multiple correlation realized by adding the Experimental Battery and EOT measures to the ASVAB when predicting CTP (adjusted R increases from .65 to .69 to .70). Similar improvements obtain for the other criteria, with the gains for the Will-Do criterion composite being particularly large (adjusted R increases from .31 to .48 to .50).

Second, a large amount of incremental validity over ASVAB is provided by the "Will-do" EOT variable for the will-do criteria (i.e., ELS and Will-Do: increases from .38 to .49 and from 31. to .48, respectively). Although EOT provides incremental validity over ASVAB for the can-do criteria (i.e., CTP and Can-Do) as well (increases from .65 to .69 and from .72 to .77, respectively), the effects for the will-do criteria are particularly striking. Even so, EOT cannot be declared the only useful predictor, because incremental validity over the EOT measure alone is obtained by both the ASVAB and

Table 6.3
Rollup Analysis Regression Models

For LVI Criteria

y = A4 [4]
= A4 + EXP [13]
= A4 + EXP + EOT [14]

= A4 + EOT [5]

= EOT [1]
= EOT + EXP [10]

For LVII Criteria

y = A4 [4]
= A4 + EXP [13]
= A4 + EXP + EOT [14]
= A4 + EXP + EOT + LVI [15]

= A4 + EOT [5]
= A4 + LVI [5]
= A4 + EXP + LVI [14]
= A4 + EOT + LVI [6]

= LV1 [1]
= LVI + EOT [2]
= LVI + EOT + EXP [11]

= EOT + EXP [10]
= LVI + EXP [10]

Note. Numbers in brackets are the number of predictor scores entering the prediction equations.

Legend -- A4 - ASVAB composites (Quantitative, Speed, Technical, Verbal)

EXP - Experimental Battery composites (Spatial, Rugged/Outdoors Interests, Achievement Orientation, Adjustment, Physical Condition, Internal Control, Cooperativeness, Dependability, Leadership)

EOT - "Can Do" (Basic + Technical); "Will Do" (ETS + MPD + PFB + LDR)

LVI - Can-Do (CTP + GSP); Will-Do (ELS + MPD + PFB)

Table 6.4
Rollup Validity Analyses: Multiple Correlations for Predicting First-Tour Job Performance (LVI) Criteria From ASVAB and Various Combinations of ASVAB, Selected Experimental Battery Predictors, and End-of-Training Performance Measures

LVI Criterion	Type	Predictor Composite					
		A	A+X	A+X+T	A+T	T	T+X
Core Technical Proficiency (CTP)	Unadj	.59	.61	.63	.62	.56	.62
	Adj	.58	.60	.63	.62	.56	.62
	Unadj	.66	.68	.70	.69	.63	.69
	Adj	.65	.67	.70	.69	.63	.69
Can-Do	Unadj	.68	.70	.73	.71	.63	.71
	Adj	.67	.70	.72	.71	.63	.71
	Unadj	.74	.76	.79	.77	.68	.77
	Adj	.72	.76	.78	.77	.68	.77
Effort and Leadership (ELS)	Unadj	.36	.38	.46	.45	.38	.44
	Adj	.35	.37	.45	.45	.37	.43
	Unadj	.39	.41	.50	.49	.41	.48
	Adj	.38	.40	.49	.49	.40	.47
Will-Do	Unadj	.30	.36	.48	.47	.42	.46
	Adj	.30	.35	.48	.46	.42	.45
	Unadj	.31	.38	.50	.49	.44	.48
	Adj	.31	.37	.50	.48	.44	.47

Note: Unadj and Adj reflect raw and shrunken (by Rozeboom, 1978, formula 8) multiple correlations, respectively. Values that are underlined have been disattenuated for criterion unreliability.

Key: A = ASVAB factors (Quantitative, Speed, Technical, Verbal)
X = Experimental Battery (Spatial, Rugged/Outdoors Interests, Achievement Orientation, Adjustment, Physical Condition, Internal Control, Cooperativeness, Dependability, Leadership)
T = End-of-Training Performance ("Can-do" or "Will-do")

the Experimental Battery for all criteria, with these increases being of about equal magnitude (compare the T column with both the A+T and T+X columns).

Third, ASVAB appears to be a better predictor of ELS than of Will-Do (adjusted correlations of .38 and .31, respectively), whereas the "Will-do" EOT measure predicts LVI Will-Do better than it predicts ELS (adjusted correlations of .44 and .40, respectively).

In summary, the Experimental Battery predictors do increment the predictive power of the ASVAB for predicting LVI job performance. These gains are smaller when taking into account all available preliminary information (i.e., ASVAB and EOT measures). EOT performance information is more useful for the prediction of will-do performance than can-do performance, although it provides incremental prediction for both types of criteria and better supplements ASVAB than does the Experimental Battery. The LVI will-do criteria show slightly different patterns of prediction, with ELS being better predicted by ASVAB but Will-Do being better predicted by EOT.

LVII Analyses

The LVII rollup results are given in Table 6.5. It should be remembered that the results for LVII must be viewed with some caution because of the small sample sizes (both MOS-specific and combined samples) upon which the analyses are based.

Looking first at the two can-do LVII criteria (CTP and the Can-Do composite), Table 6.5 shows that neither EOT nor LVI provides incremental validity to the ASVAB when predicting CTP (compare column A to columns A+T and A+1). Indeed, the adjusted correlations drop from .64 to .63. EOT also does not increment validity for the Can-Do criterion composite (multiple Rs of .68 and .67), but LVI performance data provide incremental validity for this criterion, with the validity increasing from .68 to .72. The Experimental Battery, however, adds substantial predictive power for the can-do criteria, incrementing the prediction provided by the ASVAB composites from adjusted multiple Rs of .64 to .69 for CTP and from .68 to .74 for Can-Do (compare column A to column A+X). These increments are larger than those provided by either the EOT or the LVI measures. The best predictor set for CTP is the ASVAB composites combined with the Experimental Battery measures. Adding LVI data to these two groups of predictors provides the best predictor set for Can-Do. Nevertheless, the tables demonstrate that the ASVAB scores are the best single predictor of LVII can-do performance.

For the will-do criteria (Leadership, Effort and Achievement. and the Will-Do composite), Table 6.5 demonstrates rather different findings from the can-do results. Specifically, LVI performance is the dominant predictor of LVII will-do performance. Incremental prediction over LVI performance is realized only for the Leadership criterion, however; this gain is generated primarily by the Experimental Battery (compare column 1 to column 1+X. with values of .53 and .67, respectively). Adding the "Will do" EOT measure as a predictor increases the multiple correlation another point to .68.

Looking across the will-do criteria, one unusual result occurs for the ASVAB composites. The multiple correlation between ASVAB and the Will-Do composite is smaller than the correlation of the ASVAB with the other two will-do criteria (*unadjusted* multiple R of .18 as compared to .43 and .23, respectively; the adjusted values for EA and Will-Do are .00). This result is somewhat strange given that the Leadership and the Effort and Achievement criteria are both components of the LVII Will-Do composite. Further, the adjusted multiple correlations for the latter two will-do criteria reduce to zero.

Table 6.5
Rollup Validity Analyses: Multiple Correlations for Predicting Second-Tour Job Performance (LVII) Criteria From ASVAB and Various Combinations of ASVAB, Selected Experimental Battery Predictors, and End-of-Training and First-Tour (LVI) Performance Measures

LVII Criterion	Type	Predictor Composite												
		A	A+X	A+X+T	A+X+T+1	A+T	A+1	A+X+1	A+T+1	1	1+T	1+T+X	T+X	1+X
Core Technical Proficiency (CTP)	Unadj	.58	.67	.67	.67	.58	.58	.67	.58	.29	.37	.60	.60	.60
	Adj	.54	.58	.57	.56	.53	.53	.57	.53	.29	.34	.49	.50	.50
	Unit	.44	.33	.35	.37	.45	.46	.35	.46	.29	.37	.27	.23	.23
	Unadj	.69	.80	.80	.80	.69	.69	.80	.69	.35	.44	.71	.71	.71
	Adj	.64	.69	.68	.67	.63	.63	.68	.63	.35	.40	.58	.60	.60
	Unit	.52	.39	.42	.44	.54	.55	.42	.55	.35	.44	.32	.27	.27
Can-Do	Unadj	.62	.72	.72	.76	.62	.66	.74	.66	.47	.51	.70	.66	.70
	Adj	.59	.64	.63	.65	.58	.62	.66	.61	.47	.49	.62	.59	.63
	Unit	.52	.43	.45	.49	.53	.58	.47	.57	.47	.51	.37	.31	.33
	Unadj	.72	.83	.83	.88	.72	.76	.86	.76	.54	.59	.81	.76	.81
	Adj	.68	.74	.73	.75	.67	.72	.76	.71	.54	.57	.72	.68	.73
	Unit	.60	.50	.52	.57	.61	.67	.54	.66	.54	.59	.43	.36	.38
Leadership (LDR)	Unadj	.40	.56	.63	.72	.52	.60	.70	.62	.49	.54	.70	.61	.68
	Adj	.33	.40	.51	.62	.46	.55	.60	.57	.49	.52	.63	.51	.62
	Unit	.37	.40	.44	.49	.45	.50	.46	.54	.49	.53	.42	.36	.37
	Unadj	.43	.61	.68	.78	.56	.65	.76	.67	.53	.58	.76	.66	.73
	Adj	.36	.43	.55	.67	.50	.59	.65	.62	.53	.56	.68	.55	.67
	Unit	.40	.43	.48	.53	.49	.54	.50	.58	.53	.57	.45	.39	.40
Effort and Achievement (EA)	Unadj	.21	.28	.35	.54	.30	.52	.54	.52	.50	.50	.52	.29	.52
	Adj	.00	.00	.00	.33	.14	.46	.35	.44	.50	.47	.37	.00	.39
	Unit	.15	.12	.15	.21	.21	.30	.19	.33	.50	.44	.19	.11	.16
	Unadj	.23	.30	.38	.58	.32	.56	.58	.56	.54	.54	.56	.31	.56
	Adj	.00	.00	.00	.36	.15	.50	.38	.48	.54	.51	.40	.00	.42
	Unit	.16	.13	.16	.23	.23	.32	.21	.36	.54	.48	.21	.12	.17
Will-Do	Unadj	.17	.32	.41	.60	.35	.55	.60	.56	.55	.56	.59	.38	.59
	Adj	.00	.00	.00	.44	.23	.50	.45	.50	.55	.54	.48	.00	.49
	Unit	.13	.14	.18	.25	.22	.30	.22	.35	.55	.52	.24	.16	.20
	Unadj	.18	.33	.43	.62	.36	.57	.62	.58	.57	.58	.61	.39	.61
	Adj	.00	.00	.00	.46	.24	.52	.47	.52	.57	.56	.50	.00	.51
	Unit	.14	.15	.19	.26	.23	.31	.23	.36	.57	.54	.25	.17	.21

Note: Unadj and Adj reflect raw and shrunken (by Rozeboom, 1978, formula 8) multiple correlations, respectively. Values that are underlined have been disattenuated for criterion unreliability. Unit indicates unit weighted.

Key: A = ASVAB factors (Quantitative, Speed, Technical, Verbal)
X = Experimental Battery (Spatial, Rugged/Outdoors Interests, Achievement Orientation, Adjustment, Physical Condition, Internal Control, Cooperativeness, Dependability, Leadership)
T = End of Training Performance ("Can do" or "Will do")
1 = LVI Performance (Can-Do or Will-Do)

One explanation for this result is that the relationship between the ASVAB composites and second-tour Will-Do performance stems almost entirely from the relationship between ASVAB and Leadership. The relationship with Effort and Achievement is small (unadjusted R of .23). Further, it is likely that the multiple correlations between ASVAB and the MPD and PFB LVII criteria are quite small. If this were so, the three components of the Will-Do composite other than Leadership would provide additional criterion variance that the ASVAB does not predict, thus decreasing the observed multiple correlation. Although the multiple correlation for Will-Do exceeds that for Effort and Achievement for every other combination of predictors, all other predictor combinations contain at least some measures that have been shown to be good predictors of will-do criteria (EOT "Will do", LVI Will-Do, the temperament and interest measures in the Experimental Battery). The ASVAB, by comparison, is primarily a predictor of can-do performance.

Analyses Using Unit-Weighted Predictors

For small samples, the shrinkage in R, as estimated by the Rozeboom formula, is substantial when the number of predictors is large and/or the population value for R is small. Under these conditions it is much more likely that empirical maximization procedures such as multiple regression will capitalize on chance factors and produce inflated estimates of the population parameter. Evidence of this effect appears in Table 6.5 for the Effort and Achievement and the Will-Do criteria, where multiple correlations obtained from large, optimally weighted predictor composites shrink to values of .00. For example, when using an optimally weighted composite comprising the EOT performance measure and the Experimental Battery predictors (column T+X, with 10 measures in all), multiple correlations of .29 and .38 (not corrected for attenuation) for Effort and Achievement and for Will-Do, respectively, both shrink to .00. Even when using only four predictors (the ASVAB factors), the small values of .21 and .17 for EA and Will-Do, respectively, shrink to .00.

In light of these rather severe shrinkage adjustments, analyses using unit weights for the predictors were conducted for the LVII rollup prediction equations. Because the weights are non-optimal, no adjustment for shrinkage is required. The validity estimates from these analyses are presented in Table 6.5 in the rows designated "Unit" under the Type column. Again, values corrected for criterion unreliability are underlined.

Examining the cases where the Rozeboom formula provides adjusted multiple correlations of .00 (EA and Will-Do criteria -- the first three columns and the next-to-last column), the unit-weighted predictor composites yield disattenuated values ranging from .12 for EOT and Experimental Battery predicting Effort and Achievement (column T+X) to .19 for ASVAB, EOT, and Experimental Battery predicting Will-Do (column A+T+X). The values for the unit-weighted composites are not large, but they are greater than .00.

A second finding of note is the large decreases that occur in the multiple correlations for composites containing the Experimental Battery predictors when going from optimally weighted to unit-weighted composites. In many instances, the additional

information provided by the Experimental Battery measures is not utilized by the unit weights, typically resulting in lower composite correlations than those obtained before adding the Experimental Battery predictors to the composite. For example, the unit-weighted composite correlation of ASVAB with CTP is .52, whereas the value obtained from the unit-weighted composite of ASVAB and Experimental Battery is only .39. Even larger decrements occur for the will-do criteria. For example, the unit-weighted multiple correlations for EA and Will-Do predicted from LVI are .54 and .57, respectively -- values that drop to .17 and .21, respectively, for the unit-weighted composite of LVI and Experimental Battery. This pattern recurs for most criterion/predictor combinations in Table 6.5.

Summary

In summary, the Experimental Battery provides substantial incremental validity over ASVAB to the prediction of LVII can-do performance criteria. By comparison, neither EOT nor LVI performance measures provide much incremental validity. However, for the will-do criteria, the result is much different, in that LVI performance data are the most effective predictor of LVII will-do performance criteria. The Experimental Battery does add incremental validity for predicting the Leadership criterion, but LVI is by far the most efficient predictor of the other will-do criteria.

Unit weighting circumvents the severe shrinkage of the multiple correlations obtained from optimally weighted predictor composites when samples are small, the number of predictors is relatively large, and the population value for R is not large. However, even for the conditions prevailing for the LVII rollup, optimal weights are better than unit weights for predicting the can-do criteria.

DISCUSSION AND SUMMARY CONCLUSIONS

The above analyses presented a detailed examination of the usefulness of information collected at various points in a soldier's military career for predicting his or her job performance during their first or second tour of duty. The predictors range from measures that are, or likely would be, collected prior to enlistment (scores on the ASVAB composites and the various Experimental Battery measures), to performance measures gathered at the end of MOS-specific training (EOT "Can-do" and "Will-do" composites), to first-tour performance measures ("Can do" and "Will do" composites). Although the results for second-tour job performance must be interpreted with caution, certain conclusions seem warranted.

First, the validity of prior performance information for predicting will-do performance at the next point in time was highlighted by the rollup analyses. EOT measures provided incremental validity when added to the ASVAB composites for predicting all of the LVI criterion measures, with large increments obtained for the will-do criteria. However, for the prediction of second-tour will-do performance, the validity of EOT measures for predicting the will-do criteria was low for Leadership and non-existent for Effort and Achievement and the Will-Do composite. Instead, the LVI

Will-Do composite score was the most valid measure for predicting second-tour will-do performance.

For can-do performance, the relationships between measures of prior performance and subsequent performance were not as strong. The EOT measures added a small amount of predictive validity to the ASVAB for LVI but not for LVII. LVI performance measures supplemented ASVAB and the ASVAB/Experimental Battery when predicting LVII Can-Do but not for LVII CTP.

Second, the incremental prediction provided by the Experimental Battery supplied one finding that did hold across LVI and LVII for predicting can-do criteria. The Experimental Battery provided considerably more incremental validity to the prediction of can-do criteria than to will-do criteria. Although it added approximately the same incremental validity to all LVI criteria and the LVII can-do criteria, the Experimental Battery did not contribute enough variance to increase the adjusted multiple correlations above zero for two of the three LVII will-do criteria (Effort and Achievement and Will-Do). The only measure that seems to predict these latter two criterion factors is the LVI Will-Do performance composite score. However, the Experimental Battery does provide a substantial increment in validity for the prediction of Leadership in LVII.

Overall, the rollup analyses suggest the following conclusions:

- When predicting can-do criteria, it is difficult to improve upon the ASVAB and the Experimental Battery, although EOT data are helpful for predicting first-tour can-do performance.

- When predicting will-do criteria, the best information one can obtain is the immediately preceding will-do performance information.

- Both the Experimental Battery and prior performance make major contributions to the prediction of second-tour performance.

  In general, the validities of the rollup composites are quite high, even when predicting leadership as a second-tour NCO from test battery data obtained 5 to 6 years earlier.

Chapter 7

PERSONNEL CLASSIFICATION AND DIFFERENTIAL JOB ASSIGNMENTS: A REVIEW OF THE ISSUES AND ALTERNATIVE MODELS

John P. Campbell, Mark Fellows, and Paul Sticha

The objectives of this chapter are to (a) summarize the major options for modeling the classification problem and the critical issues they encompass; (b) review alternative procedures for estimating classification efficiency; and (c) outline the major alternative strategies for making differential job assignments.

Since there is probably no such thing as a "pure" classification problem that can be identified in a real-world context, any attempt to model personnel selection and classification fully must deal with various sets of constraints. However, given some agreed-upon way to represent the problem, three major issues remain. The first concerns how to choose and/or weight the variables in a prediction equation so as to maximize the potential gains from using a classification strategy. The second concerns how to estimate the degree of classification efficiency that is potentially obtainable; this is analogous to estimating the population value for the selection validity coefficient. The third issue deals with how differential job assignments can actually be made so as to realize the gains that are possible from classification.

It is reasonable to think of potential gain and realized gain for classification as well as for selection. For example, in the selection situation, the predictive accuracy represented by the population validity coefficient can be realized only if all future applicants are a random sample from the appropriate population, and there is no systematic bias that distinguishes high scorers on the predictors (X) who are offered a job and accept it from those who are offered the job and do not accept it. Similarly, individuals who are to be classified may not actually take the job assignment that is optimal for the organization.

The major sections of this chapter will attempt to address each of these issues in turn. First, a background section outlines the properties of a selection/classification system that can potentially influence the system's payoffs. Next, the major issue of predicting group membership versus predicting criterion outcomes is discussed in some detail. Following on this major distinction, a number of specific representations of each type of goal are summarized. The next sections are devoted to describing the available methods for selecting predictors for classification and the methods for estimating classification gains. Finally, the last two sections summarize the current job assignment systems in each Service and contrast them to two new prototype procedures.

THE COMPONENTS OF SELECTION AND CLASSIFICATION

The more the Army can learn about the benefits and costs of alternative methods for selecting and classifying the individuals who apply, the more effective its personnel management systems can be. Ideally, personnel management would benefit most from a

complete simulation of the entire system that would permit a full range of "what if" questions focused on the effects of changes in (a) labor supply, (b) recruiting procedures, (c) selection and classification measures, (d) decision-making algorithms, (e) applicant preferences, (f) various organizational constraints, and (g) organizational goals (e.g., maximizing aggregate performance, achieving a certain distribution of individual performance in each job, minimizing attrition, minimizing discipline problems, or maximizing morale). Further, it would be desirable to have a good estimate of the specific costs involved when each parameter is changed.

Describing, or "modeling," effective selection and classification in a large organization is a complex business. When all the variations in all the relevant components are considered, there may indeed be dozens, or even hundreds, of alternative models that could be considered. Also, there are always one or more constraints on personnel decision-making that are specific to the organization, which complicates the decision model even further. The overall complexity of any real-world personnel management situation is such that an operational process for making selection and classification decisions probably cannot be fully modeled by currently available analytic methods.

It may not even be possible to describe all the potential parameters of a real-world selection/classification model. However, while one may not be able to fully account for all the parameters in a complex selection/classification system, it may still be possible to capture the most relevant properties and to use Monte Carlo methods to estimate their effects. If such a simulation could reflect the most critical parts of the system, it could serve as a useful test bed for evaluating the effects of making changes. For purposes of thinking about future attempts to model the classification problem, one can start by simply listing the major parameters of selection and classification decision-making, and the principal implications of each.

The Goal of Selection/Classification

A selection and classification decision model would be implemented to achieve a particular objective, or set of objectives. Identifying the objective(s) is the most critical ingredient of the model because it directly determines what information and procedures are appropriate for use in decision-making.

Some possible alternative objectives are to (a) maximize the mean individual performance across jobs, (b) maximize the number of people above a certain performance level in each job, (c) maximize the correspondence of the actual distribution of performance in each job to a desired distribution, (d) minimize attrition across all jobs, (e) minimize the number of "problem" employees across all jobs, (f) fill all jobs with people who meet minimal qualifications, or (g) maximize the utility, or value, of performance across jobs.

There are many important implications relative to these alternative decision-making objectives. For example, the procedure for maximizing average expected performance would not be the same as for maximizing average expected utility if the

utility of performance differs across jobs and/or the relationship of performance to performance utility within jobs is not linear.

The way in which performance is to be defined and measured is also critical. For example, if major components of performance can be identified, then which component is to be maximized? If the objective is to maximize some joint function of multiple goals (e.g., maximize average performance and minimize attrition), then deciding on the combination rules is a major issue in itself. For example, should multiple goals be addressed sequentially or as a weighted composite of some kind?

Selection Versus Classification

A personnel decision-making system could give varying degrees of emphasis to selection versus classification. At one extreme, individuals could be selected into the organization and then assigned at random to k different jobs (i.e., MOS). At the other extreme, no overall selection would occur and all available information would be used to make optimal job assignments until all available openings were filled.

In between, a variety of multiple-step models could emphasize different objectives for selection and classification. For example, selection could emphasize minimizing attrition while classification could emphasize those aspects of individual performance that are the most job specific. For classification to offer an advantage over selection plus random assignment, jobs must in fact differ in terms of their requirements, predictability, difficulty level, or relative value (utility). Many other factors as well will affect the decision process in varying degrees.

Job Differences. As noted above, jobs can differ in a number of ways that are relevant for modeling selection/ classification decisions. That is, gains from classification (over selection plus random assignment) can be greater to the extent that (a) jobs differ in the knowledges, skills, and abilities (KSAs) required, and consequently a greater degree of differential prediction is possible; (b) jobs differ in terms of the accuracy with which performance can be predicted and higher ability people are assigned to the more predictable jobs; (c) jobs differ in terms of the value or utility of performance; or (d) jobs differ in terms of the variance of performance or the variance in performance utility (e.g., SDy is different across jobs) and, other things being equal, higher ability people are assigned to jobs with higher SDy's.

The Number of Jobs. Other things being equal, the gains from classification are greater to the extent that the number of distinct jobs, or job families, is greater.

Selection Ratio. The gains from both selection and classification are greater to the extent that the number of applicants exceeds the number of openings.

Applicant Population. The gains from both selection and classification are greater to the extent that the mean qualification level of the applicant pool is greater.

Predictor Battery. Gains from selection are directly proportional to increases in the validity coefficient (R). Gains from classification are a joint function of the average R across jobs and the level of differential prediction across jobs that can be obtained by using a different predictor battery for each job or each job family. The nature of this joint function is perhaps a bit more complex than the conventional wisdom implies.

Individual Preferences. Other things being equal (e.g., subsequent motivation), the gains from classification will be less to the extent that individual preferences across jobs do not correlate with individual profiles of predicted scores across jobs, and individual preferences play a significant role in job assignments. Under such conditions, preferences become a constraint on optimal assignments.

Quotas. Other things being equal (e.g., the total number of people to be assigned), the gains from classification are less to the extent that each job has a specified number (quota) of openings that must be filled.

Real Time Versus Batch Decision-Making. Other things being equal, to the extent that job assignments must be made in real time and the characteristics of future applicants during specified time periods must be estimated, the gains from classification will be reduced. The decrement will be greater to the extent that the characteristics of future applicants cannot be accurately estimated.

Additional Constraints. In all organizations, the selection and classification decision-making process must operate under one or more additional constraints (e.g., budget limitations, training "seat" availability, hiring goals for specific subgroups, management priorities). In general, the existence of constraints reduces the gains from selection and classification. These effects must be taken into account.

Gains Versus Costs. The gains from selection and classification obtained by recruiting more applicants, recruiting higher quality applicants, improving the assessment of qualifications (e.g., a better predictor battery), enabling more informed individual preferences, and improving the assignment algorithm are partially offset by increases in related costs. The primary cost factors are recruiting, assessment, applicant processing, and system development (R&D). It is possible that a particular gain from improvements in classification could be entirely offset by increased costs.

Any attempt to fully model the selection and classification decision process in a real-world organization must take the above issues into account.

MULTIPLE REGRESSION VERSUS MULTIPLE DISCRIMINANT ANALYSIS

At the most general level, two types of personnel classification models have been researched. The first is the regression-based model, which begins by deriving least square estimates of performance separately for each job, using all predictors. The regression models have been labeled "maximization methods" to indicate that job performance is the criterion to be maximized by optimal assignment of applicants to different positions. The

second is the discriminant-based model, which derives least square estimates to maximally predict membership (or perhaps performance above a certain level) in a job. Group membership itself is the criterion of interest.

Differences Between the Two Approaches

Differences between the two basic approaches to classification are fundamental in the sense that they are not easily evaluated in the same terms. Their differences can be summarized as follows:

(1) Classification Goal, or Criterion: The efficacy of a regression-based classification procedure is determined by an aggregate outcome index such as mean performance level of assignees across all jobs. For discriminant approaches the appropriate measure of efficacy is a probability of group membership or some related index. Discriminant analysis is not concerned with estimating performance and regression procedures do not estimate probability of group membership.

(2) Estimation Methods: Both are least-squares maximization procedures. In regression, a linear combination of predictors is formed to maximally differentiate among individuals within each group. In discriminant analysis, a linear combination of predictors is formed to maximally differentiate among individuals from different groups.

(3) Application of Equations: Once the coefficients are estimated, in both approaches the equations are applied to new applicant groups. In the regression approach, performance is predicted for each applicant for each job, using the equations from each job, and yielding a vector of predicted performance scores for each applicant. For multiple discriminant analysis, the probability of group membership is derived from applying the discriminant functions to each applicant's profile of predictor scores, yielding a vector of probabilities for each applicant.

(4) Assignment: In the case of regression methodology, the applicant vectors of performance estimates can be entered into a linear programming algorithm to distribute applicants among jobs in such a way that aggregate performance is maximized. Constraints such as quotas expressing the number of openings can also be considered by the linear programming procedure. Linear programming could be similarly used to make assignments based on the vectors of job membership probabilities provided by discriminant methods. In this case the algorithm could maximize the mean membership probability in the assigned job across jobs, given the same sort of constraints.

(5) Evaluation: The most natural index for evaluating regression-based assignments is an aggregated performance measure achieved in a new sample. For discriminant analysis it is the average probability of group

membership in the assigned jobs achieved in a new sample, or the number of correct classifications relative to a base rate (e.g., random assignment). It could be informative, however, to apply the evaluation criteria for each method to the other. If the proper criteria are available, either classification procedure could be evaluated with reference to any classification goal.

Advocates of regression methodology do not view jobs as natural groupings in the same sense as classes or species in biology, the discipline within which the discriminant methodology was developed: "The model [discriminant analysis] is much less appropriate in personnel decisions where there is no theory of qualitatively different types of persons" (Cronbach & Gleser, 1965, p. 115). Advocates of discriminant methods, however, take exception to fundamental assumptions of regression methodology. Specifically, they question whether it is appropriate to predict performance for an individual without knowing if the individual is a member of the same population on whom the regression equation was estimated. Rulon, Tiedeman, Tatsuoka, and Langmuir (1967) note, "We seldom bother to estimate an individual's similarity to those who are in the occupation before using the multiple regression relationship for the occupation" (p. 358). In the Army context, this raises the issue of whether the effects of self-selection for different MOS are strong enough to produce different subpopulations within the applicant pool.

The problem of predicting performance for individuals who may not come from the same population as the sample on which the equations were derived does not have an entirely adequate empirical solution. Horst (1954) noted that the misestimated regression weights that are a symptom of this problem can be largely corrected with corrections for restrictions in range using applicants for all jobs as the applicant population. Consider the case, however, where the incumbents in a job (e.g., mechanic) are uniformly high on an important predictor in the applicant population (e.g., mechanical interests and aptitude). The within-job regression equation developed to predict performance would not include this predictor because of its severely restricted range. When this equation is then applied to the general population of applicants, the important predictor would not be weighted and inaccurate performance estimates would result. Especially troubling would be the potential for a high predicted performance score for an applicant with low scores on important but range-restricted (and thus unweighted) predictors (e.g., mechanical interests and aptitude) and high scores on predictors that are less important for the specific job but that vary across a much larger range in the incumbent population.

There has been concern for the accuracy of range restriction correction formulas when the selection ratio is extreme (Lord & Novick, 1968). This concern has been found in Monte Carlo studies to be largely unfounded, but only when the assumptions of the correction (linearity of regression and homoscedasticity) are not violated (Greener & Osburn, 1980). In the case of violations of one or both assumptions and extreme selection, the corrected correlation can be very inaccurate estimates of the population correlation (Greener & Osburn, 1980), and most often they are still underestimates.

Restriction of range problems are of concern not only to regression models. Restriction in range of predictor variables violates the assumption of equal within-group covariance matrices in discriminant analysis. Correction of the within-group matrices may reduce the extent of violations of this assumption, but application of the formula is subject to the same concerns as in the case of extreme selection discussed above.

Discriminant functions calculated on corrected and uncorrected matrices could be evaluated in terms of percentage of correct classifications in a cross-validation sample. In regression-based classification, there is no directly analogous check of classification accuracy. This difference between the two classification models is fundamentally a function of the different criteria they employ. In the regression situation we do not have criterion data for all persons in all jobs. For the discriminant function we do have criterion data for all persons because job membership itself is the criterion. This allows for direct cross-validation checks on the accuracy of classification in discriminant assignment.

In general, the choice of the appropriate classification model is at one level a purely theoretical one. The advocates of each type of model rely on different basic assumptions about incumbents in a job as a discrete class of individuals. However, the two methods have different practical implications as well. They rely on different types of data, maximize different parameters, and hold different implications for assigning personnel. It is on these grounds that the two models might be most profitably compared.

Regression Models

Using regression procedures in classification of personnel was first discussed in theoretical terms in work by Brogden (1946a, 1951, 1954, 1955, 1959) and by Horst (1954, 1956). The framework developed by these two researchers consists of an established battery of tests optimally weighted to predict performance separately for each job. The key assumptions behind this methodology are that the relationships between predictors and performance are multidimensional and that the ability (and other personal characteristic) determinants of performance vary across jobs or job families.

Although Brogden and Horst shared a common framework, they focused on different aspects of the classification problem. Horst's concern was with establishing procedures to select tests to form a battery that maximizes differential prediction across jobs. Horst (1954) described test selection by the method of least squares, which is essentially a stepwise regression procedure that maximizes an index of differential validity at each step. Horst called this index "differential predictive efficiency" or $\mathbf{H_d}$, which is defined as the sum of the covariances between the pairwise differences among predictors and among criteria. In other words, it is the prediction of differences among criterion scores across jobs and within individuals in which Horst is interested.

Specifically, Horst's procedure operates on the matrix of correlations between all of the predictors and all of the criteria. Let us call this matrix **V** and specify its dimensions as $(p \times j)$ where p is the number of predictors and j is the number of jobs.

The selection of variables proceeds by calculating the variance of each row of V and selecting the variable with the largest variance in predicted criterion correlations across jobs. Next, the variance of the selected predictor is partialled from the V matrix, creating V_2 matrix of dimensions (*p-1 x j*). The variances of criterion correlations across jobs are again calculated and the predictor with the largest variance is chosen as the second predictor in the differential validity equations. This variable maximizes H_d (or the prediction of differences among predicted criterion scores) for the two-variable composite. The next step is to partial this variable's variance from V_2 to create V_3 of dimensions (*p-2 x j*) and continue extracting variables.

Given this subset of best differential by predicting variables, Horst specifies the computation of the regression weights and the multiple correlations with each criterion. Using Horst's method, then, it is feasible to select and weight variables with differential prediction as a criterion. It is an empirical question how well these equations perform when absolute levels of validity (perhaps averaged across criteria) are considered. Simulation procedures, to be mentioned later, facilitate comparison of classification methods designed with absolute or differential validity priorities.

Brogden assumed that the battery of tests to be used is a given and concentrated on estimating the increase in assignment efficiency achieved by classification methods. Brogden (1959) outlined an approach to estimating classification gains based on a model analogous to Spearman's two-factor theory. Seen in this way, variance in job performance is partly shared with all other jobs and partly unique to each job. The shared component of variance is represented in Brogden's work by the intercorrelations among least square estimates of performance for a set of jobs. This corresponds to the general component or factor of performance, much like Spearman's "g". The remaining predictable variance in job performance is attributable to a set of unique factors (one for each job or job family) uncorrelated with the general factor and with each other.

While Horst did not address the issue of evaluating classification batteries and assignment procedures, Brogden developed tabled estimates of classification efficiency for the two-job two-predictor case under certain simplified assumptions (Brogden, 1951). The efficiency metric in his procedure is standardized mean predicted performance (MPP) calculated as the average predicted performance across all job assignments.

Recently, Johnson and Zeidner have rekindled interest in classification by integrating the Brogden and Horst theoretical traditions. Work by Johnson and Zeidner (1990, 1991; Johnson, Zeidner, & Scholarios, 1990) and their colleagues (Statman, 1992) is most relevant here because it provides simulation methodology for the estimations of MPP as a function of various parameters of interest in classification situations (e.g., number of job families, number and kind of predictors), and because they have developed simulations based on Project A data.

By extracting factors from this residual matrix, the hypothesis of multidimensionality in the joint predictor/criterion space can be examined. If multidimensionality exists and the jobs fall into a relatively simple structure, the components can be used to define job families for classification purposes.

Discriminant Models

The majority of research on classification in the discriminant analysis tradition has used non-cognitive predictors such as vocational interests and personality constructs. Research on interests using the discriminant model dates back to Strong's (1931) publication of the Strong Vocational Interest Blank and the demonstration through his research that different occupational groups reliably differ on patterns of interests.

One classification methodology emerging from this tradition is Shoenfeldt's (1974) assessment classification method. This development can be seen as a union of contemporaneous efforts in the taxonomy of jobs (e.g. McCormick, Jeanneret, & Mecham, 1972) and of persons (Owens, 1978; Owens & Schoenfeldt, 1979). Although the original conception was developmental in nature and the first empirical study involved college student interests and their subsequent choice of major fields of study (Schoenfeldt, 1974), later empirical investigations attempted to link interest-based clusters of persons with jobs or job clusters (Brush & Owens, 1979; Morrison, 1977).

Brush and Owens formed 18 groups of persons based upon cluster analyses of biographical information and 18 job families in a large organization based upon a clustering of job analysis information. Their results showed that 11 of the 18 person clusters displayed job family memberships that were "clear departures from base rates." Post hoc examinations of the characteristics of matching person and job clusters revealed several matches that displayed intuitively sensible combinations of person and job characteristics.

Research by Skinner and Jackson (1977) took a similar approach in classifying persons by clustering profiles of personality scale scores and by using discriminant functions to analyze the personality characteristics of incumbents in certain jobs. *Post hoc* observations also showed interesting and reasonable allocations of persons in certain "modal profile" types into certain military occupations. In the discriminant analysis, four significant functions were found that were interpreted as: risk taking vs. dominance, independence vs. social orientation, impulse expression, and aggressive orientation vs. dependence. The authors concluded that this type of information would be useful for counseling purposes.

The other major theoretical analysis of discriminant-based classification is the work of Rulon and colleagues (Rulon et al., 1967). This approach calculates probabilities of group membership based upon similarities of patterns of predictors between applicants and means of job incumbents. Rulon et al. assume multivariate normality in the distribution of predictor scores and derive a statistic (distributed as chi square) to represent the distance in multivariate space between an individual's pattern of scores and the centroids for the various jobs. These distances can be converted into probabilities of group membership and assignment can be made on this basis. In an attempt to link the discriminant and regression approaches, Rulon et al. include a derivation of the joint probability of membership and success in a group. The measure of success is limited to a dichotomous acceptable/not acceptable performance criterion.

The implications for practical application of this procedure are that the applicant should be assigned to the job for which the probability of membership (or membership and success) is highest. In cases where the probabilities for numerous groups are similar in size, or when all probabilities are negligible, the allocation choice becomes relatively arbitrary.

The most recent application of discriminant-based methods used the large Project TALENT database of high school students to develop equations that predict occupational attainment from ability, interest, and demographic information (Austin & Hanisch, 1990). This research involved dividing occupations into 12 categories and then calculating discriminant functions to predict occupational membership 11 years after high school graduation. Prediction in this case is very impressive; the first five discriminant functions accounted for 96.8 percent of the variance between groups with the majority (84.5%) accounted for by the first two functions. The first function can be interpreted as a general ability composite while the second function weighted mathematics ability and gender heavily. The final three functions accounted for relatively little variance and had much less clear interpretations.

Conclusions

Given the above discussion, the following points seem relevant.

Group Membership as a Criterion

One critical issue in comparing the regression and discriminant approaches is the nature of the membership criterion. Humphreys, Lubinski, and Yao (1993) express enthusiasm for group membership as an "aggregate criterion" of both success in and satisfaction with an occupation. They note that the composite nature of the membership criterion addresses method variance and temporal stability concerns that have been problematic for regression methodology. The quality of the group membership criterion, however, is critically dependent on the nature of the original classification system and subsequent opportunities to "gravitate" within the organization to reach the optimal job, as well as on the forces that produce attrition from the original sample of incumbents. The more gravitation that occurs and the more that the causes of attrition are related to performance, the more valuable the job membership criterion becomes. Discriminant-based classification is most useful when the employment situation involves free choices and/or free movement over a period of time and attrition occurs for the "right" reasons. When these conditions do not exist, discriminant procedures do not provide interesting new information.

Wilk, Desmarais, and Sackett (1993) have provided limited support for within-organization gravitation. Specifically, they note the movement of high general ability employees into jobs of greater complexity. Additionally, they found that the standard deviation of ability scores tended to decrease with longer experience in a job providing further indirect evidence for gravitation. Discriminant analysis has been most effective in research that considers broader occupational criterion groups and longer periods that allow for crucial gravitation to create meaningful groupings (e.g., Austin & Hanisch,

1990). However, despite the popularity of discriminant methodology and its success in predicting occupational membership, without stronger evidence for substantial gravitation and/or appropriate attrition, it seems inappropriate for classification research within organizations.

Another use of the discriminant functions would simply be to use them to predict performance itself. Comparison of MPP values obtained from alternative weighting schemes (such as the application of discriminant weights) in a regression equation might demonstrate favorable MPP results when compared with full regression methodology.

While results in terms of probability of correct classification may produce little or no information for the organization in terms of quantifying improvements in performance, this type of information could be very useful to applicants for advising and counseling purposes, even if it is not formally used in making assignments.

Assigning Jobs to Families

Cronbach and Gleser (1965) lamented the loss of dimensional information in job family designations, and yet the inclusion of all jobs in a large organization or discrete assignments can be prohibitive.

One possible way to preserve more information on between-job differences would be to estimate regression models hierarchically. Bryk and Raudenbush (1992) present a statistical model that allows for the estimation of a performance criterion at one level and the estimation of variation in regression weights across jobs as a function of characteristics of the job at another level. If job analysis information is available for a large number of jobs, this can constitute the dimensional information for the second level of the hierarchical model. By this (or some other similar) method it is possible to escape from the situation of allocating all of the variance of a job to one family or cluster.

McCloy, Hedges, and Harris (1991) present a derivation and an application of a multilevel regression model to military data. At the first level, job performance was predicted by AFQT and ASVAB technical scores and by time in service and education scores. At the second level, regression parameters for these predictors were modeled to vary across jobs as a function of four principal component scores calculated from job analysis data (working with things, cognitive complexity, unpleasant working conditions, and fine motor control). If job families are not true types or natural entities, a dimensional conception of their differences might be appropriate.

The question also remains of how to assign MOS to the new job families. One solution is synthetic validation (see Mossholder & Arvey, 1984 for a review). Prediction equations have been synthetically developed for the 18 MOS in Project A using content and job component validity information (Peterson, Rosse, & Owens-Kurtz, 1990). It is possible that more global assignment judgments of MOS into existing families might be a more efficient procedure.

A fairly extensive synthetic validation research effort with possible relevance to classification used the Position Analysis Questionnaire (PAQ) job analysis instrument (McCormick, DeNisi, & Shaw, 1979; McCormick, Jeanneret, & Mecham, 1972). These researchers attempted to predict mean scores on different predictor constructs using least-squares weighted combinations of job analysis dimension scores. While conceptually interesting, this methodology suffers from criterion problems similar to those in the discriminant-based methods mentioned earlier. Specifically, there is a critical assumption that persons gravitate to jobs in which they are satisfied and successful. If this is not true, then the prediction of mean test scores of incumbents is a misleading exercise. This research is intriguing, however, if only for its treatment of jobs as differing along dimensions and the use of this dimensional information for classification purposes.

It is tempting to consider adapting regression-based methodologies such as those used by Johnson and Zeidner (1990) to consider dimensional information. The factor solution which results from their procedure is dimensional in nature but is used to make discrete categorizations of jobs into the derived families. Any new or additional jobs to be considered would have to be similarly categorized.

An alternative to this procedure would be to take advantage of the rich dimensional information contained in the factor solution and have expert judges place newly considered jobs into the reference factor solution. This could be achieved via a judgment task or through a geometric representation of the solution (possible in limited dimensionality through multidimensional scaling) into which the experts place the new job to best fit in the configuration of the old ones. Mean predicted performance for this new job could be calculated as a combination of predicted performances from the reference jobs weighted by their proximity to the new job. In this manner, even a job created to accommodate new technology or organizational structure that does not fit into the reference job families could be more faithfully represented. If the sampling of MOS used in Project A is in fact representative (which it was designed to be), these data can be used to determine reasonable reference factors (or job families) for this kind of dimensional exercise.

Using Computer Technology and Simulation

A severe limitation for the optimal implementation of assignment procedures is the requirement of batch processing for linear programming algorithms. The classification situation must be "frozen" over a relatively large period of time to provide enough applicants to run through the procedure. This situation is at odds with the requirement of on-line decision making as applicants enter the organization. Computer-based expert systems trained and tuned with simulation methodology provide a potential alternative to batch-oriented processing.

More recent developments in expert systems have incorporated concepts of fuzzy logic to handle problems of higher complexity. Fuzzy logic allows for graded degrees of certainty of membership instead of all-or-nothing formulations of classic logic. This allows fuzzy logic-based systems to deal with ill-defined concepts or variables measured with substantial error (see Zadeh, 1972). Variables in the classification and assignment

decision problem which might profitably be expressed as fuzzy variables are probabilities of group membership, graded statements about MOS quota situations or fill criticality, MOS-specific minimum predictor requirements, and even predictor-criterion relationships.

For example, if the variable "MOS fill need" (or x) took on fuzzy values of "*none, low, medium, high, critical*" (each taking on a specified range of values) and the variable "probability of MOS membership" (or y) took the fuzzy values "*very low, low, medium, high, nearly certain*" (each representing a range of probabilities), rules in the fuzzy expert system would be conditional statements that link the two variables and express implications for a decision variable. An example is: "IF x is *high* and y is *medium*, THEN give z MOS as an option to applicant i."

Using certain distributional models of fuzzy variable values, the development of decision rules and the "defuzzification" of outcome decisions can occur. Although these procedures entail loss of precision, they also allow for the modeling of very complex systems and for a useful level of quantification. If a reasonable simulation of a personnel system could be used to calibrate the fuzzy variable "metric" and to evaluate their joint relationships with outcomes, then it might be possible to closely approximate batch-processed linear programming results with an on-line expert system approach. Expert system-based decision processes also seem to blend well with current Army operational assignment procedures in terms of computer implementation and the provision of choice to applicants, given certain constraints.

Another potential application of new technology to selection and classification involves alternative prediction models such as the neural networks that have been successfully applied in classification and prediction situations in many domains for several years. These techniques have proven especially valuable in approximating solutions to problems with a degree of complexity that forbids exact analytical solution. Very recently, neural network prediction models have been formulated in a personnel selection situation and compared with traditional regression procedures (Collins & Clark, 1993). Although neural networks are able to take advantage of stable nonlinearities in relationships, they also capitalize more on chance variation and their predictions are likely to shrink more on cross-validation than regression predictions do. Even after cross-validation, however, Collins and Clark's neural network models showed modest improvements over regression prediction.

Neural networks can be configured to perform either regression-like performance prediction or discriminant-like battery classification. In decision situations in other domains, neural networks have been formally embedded inside expert systems as a part of the knowledge base (for an example in medical diagnosis, see Cohen & Hudson, 1992). It would be possible to "train" a neural network to make on-line classification decisions by presenting it with input data (a vector of predictor scores; vectors of job characteristics, importance ratings, and fill status values; and possibly a matrix of validities) and output data in the form of the assignments made by a batch-processed procedure. After each assignment made by the neural network, the fill status variables would be accordingly adjusted. It is possible that this type of prediction system (perhaps

embedded in an expert system containing other decision rules) could provide on-line assignment performance nearly equal to batch processing.

Finally, Cronbach and Gleser (1965) point to the possibility of using adaptive testing procedures to facilitate classification. Ideally, rough assignment could be made for an applicant on the basis of a general battery of tests. At the next stage, predictor tests could be individually chosen that would provide the most relevant information to make more fine-grained classifications. If all of the predictor tests were computer administered, the selection of tests could occur on-line and save a great deal of testing time, while simultaneously preserving much of the potential gain from classification.

Summary

In one very real sense the regression model and the discriminant function do not have the same goals and cannot be compared directly. However, in any major evaluation of a selection/classification system, the design might best be thought of as having a repeated measures component that evaluates the classification gain for several kinds of objectives. Also, it would be useful to determine how much is lost, or gained, when the predictor weights estimated by one model are used to make job assignments intended to maximize a goal appropriate for the other.

CLASSIFICATION OBJECTIVES AND CONSTRAINTS

Efficient applicant assignment procedures should be controlled by the objectives and constraints of military selection and classification. Objectives define the functions to be maximized by the classification process. Constraints define the minimum standards that must be met by any acceptable classification solution. When a problem has both objectives and constraints, the constraints are of primary importance, in the sense that maximization of the objectives considers only candidate solutions that meet all constraints. On the other hand, additional capability beyond the minimum standard can add value to an assignment system if the additional capability helps to satisfy a desirable, and specifiable, objective.

When an optimal procedure, such as linear programming, is used to solve a classification problem, the objectives are represented as continuous functions to be maximized, while constraints are represented by inequalities among variables that must be satisfied. However, because of the duality between objectives and constraints in optimal assignment methods, it is in some sense arbitrary whether a particular factor is considered an objective or a constraint.

For example, we could easily frame a classification problem as one of minimizing the total cost required to meet specific performance standards. In this case, minimizing cost would be the objective, while the performance standards would be the constraints on the classification process. Alternatively, the problem could be formulated as one of maximizing performance, subject to cost constraints. This approach reverses the role of objective and constraint from the first formulation. A third approach would maximize a

function that combines cost and performance, such as a weighted average (the weight assigned to cost would be negative). This approach has no constraints, because both of the relevant variables are considered part of the objective.

It is possible for all three methods to arrive at the same solution, if the constraints, objective functions, and weights are set appropriately. However, in general, optimal solutions will satisfy the constraints exactly, or very nearly so, and then attempt to maximize (or minimize) the objective function. In the previous example, additional performance generally requires additional cost. Thus, if performance is considered to be a constraint in a classification problem, then the optimal classification will barely meet the performance standards, while minimizing cost. Thus, it is important to ensure that a classification that meets each constraint exactly is truly acceptable. If the constraints are set too low, then the optimization procedure may reach a solution that, in reality, is unacceptable. Alternatively, if the constraints are set too high, there will be little room for the optimization of the objectives to occur.

Wise (1994) has identified a number of potential goals that could be addressed by selection and classification decisions, depending on the organization's priorities.

Maximize percentage of training seats filled with qualified applicants. When the classification goal is stated in this way, the specifications for qualified applicants are a constraint. If the fill rate is 100 percent, the quality constraint could be raised.

Maximize training success. Training success could be measured via course grades, peer ratings, instructor ratings. Success could be represented with a continous metric or a dichotomy (e.g., pass/fail).

Minimize attrition. Any index of attrition must be defined very carefully and would ideally take into account the time period during which the individual attrited, as noted by McCloy and DiFazio (1994).

Maximize aggregate job performance across all assignments. In terms of a multifactor model of performance, the maximizing function could be based on any one of the factors, or some weighted composite of multiple factors.

Maximize qualified months of service. As used in previous research, this term refers to the joint function of attrition and performance when performance is scored dichotomously as qualified/not qualified.

Maximize aggregate total career performance. This goal would be a joint function of individual performance over two or more tours of duty.

Maximize the aggregate utility of performance. In this instance, the performance metric would be converted to a utility, or value of performance, metric which could change assignment priorities as compared to making job assignments to maximize aggregate performance.

Maximize percentage of job assignments that meet specific performance goals. The current personnel assignment system's "quality goals" fall in this category. The assignment rules could use one cut score as a minimum standard for each job or they could define maximum and minimum proportions of individuals at each of several performance levels.

Maximize the social benefit of job assignments. Potential indicators of such a goal could be things such as the percentage of minority placements or the potential for civilian employment after the first tour.

This list illustrates the variety of goals that may be served by selection and classification processes. The individual goals are not mutually exclusive, but neither are they totally correspondent with each other, and no organization could be expected to try to optimize all of them at once. Some of these goals, such as maximizing training seat fill rates, are in close chronological proximity to the classification process, and can be easily measured. Others, such as total career performance, cover a period of time that may be many years removed from the classification process. Finally, some are "nested" within others, such as maximizing total performance utility which is really maximizing total performance, where levels of performance have been evaluated on a utility metric.

One important feature is that most of the goals on the previous list could be stated as either objectives or constraints. In addition, there are other constraints under which the classification system must operate. Probably the most obvious of these is cost. Other constraints are quotas for total accessions and for individual jobs, and minimum performance standards.

Different classification procedures focus on different objectives and constraints, and employ different methods to determine the optimal allocation of applicants to jobs. No existing method addresses all of the goals described above.

PREDICTOR SELECTION FOR CLASSIFICATION

Brogden (1955) and others have argued that the maximum potential gain from classification will occur when the most valid predictor battery for each individual job is used to obtain predicted criterion scores for that job for each person, and these predicted criterion scores are what are used to make job assignments. However, this means that each individual must be measured on each predictor, no matter what subset of them is used for each job. Consequently, the total number of predictors aggregated across jobs could get large. If the number becomes too large, the necessary database becomes too expensive to generate.

As a consequence of the above considerations, one major issue in selecting predictors for classification is how to choose a battery that will maximize potential classification efficiency for a battery of a given length, or more generally, for a given cost. The available methods for making such a battery selection are limited. We know of only two totally empirical methods, and they are discussed in the following subsections.

The Horst Method

The value of the predictors used for classification depends upon their ability to make different predictions for individuals for different jobs. That is, it should be possible, using the predictors, to predict with some accuracy that one job assignment is better than another for an individual. Differential validity refers to the ability of a set of predictors to predict the difference between criterion scores, such as performance on different jobs. Horst (1954) defined an index of differential validity (H_d) as the average variance in the difference scores between all pairs of criteria accounted for by a set of tests. It is not feasible to calculate H_d directly, because criterion scores are not available for more than one job for any individual. Consequently, Horst suggested substituting least squares estimates (LSEs), or the predicted criterion scores for the actual criterion scores.

When H_d is calculated based on LSEs of criterion performance, then the index may be calculated from the matrix of covariances between the LSEs, denoted C, and the average off-diagonal element of C, as shown in the following equation:

$$H_d = tr\ C - 1'C1/m$$

where $tr\ C$ is the sum of the diagonal elements (or trace) of C, 1 is a vector with each element equal to one, and m is the number of jobs.

The Abrahams et al. Method

An alternative to the Horst stepwise procedure that chooses tests for a battery to maximize differential validity would be to use a stepwise procedure to select predictors that maximize the mean selection validity across jobs. This is congruent with goals of the Brogden model but it provides a means to eliminate tests from the total pool such that all applicants don't take all tests.

This approach was used by Abrahams, Pass, Kusulas, Cole, and Kieckhaefer (1993) to identify optimum combinations of ten ASVAB and nine ECAT subtests to maximize selection validities of batteries ranging in length from one to nine subtests. For each analysis, the steps were as follows: (a) The single test that had the highest weighted average validity across schools (jobs) was selected, (b) multiple correlations were computed for all possible combinations of the first test with each of the remaining tests, and (c) the average multiple correlation was then computed for each test pair and the combination with the highest weighted average was selected. This process was repeated until all remaining tests were included. The idea is that at each step, the subtests currently in the prediction equation represent the test battery of that length that maximizes average absolute selection validity.

Three limitations are associated with this method. First, because this method does not allow subtests to be dropped at later stages of accretion (i.e., the procedure uses forward stepwise regression only), the particular combination of subtests at the n^{th} step may not be the battery of n tests that maximizes mean absolute validity. The relative

contribution of a particular subtest may diminish as other subtests are added to the battery.

Second, this method seeks only to identify the optimal battery relative to the mean absolute validity across jobs; it does not account for potential classification efficiency or subgroup differences. One implication of this difficulty is that for a test battery of any length, the combination of subtests identified by this method may produce a relatively low level of differential validity compared to another combination of subtests with a similar level of mean absolute validity, but a higher level of differential validity.

A final limitation of this method is that while the number of subtests in a test battery is related to the cost of administering the battery, the actual administration time for the test battery might provide a more accurate assessment of cost.

Abrahams, Alf, Kieckhaefer, Pass, Cole, and Walton-Paxton (1994) addressed this latter concern in a subsequent analysis. They used an iterative procedure described by Horst (1956) and Horst and MacEwan (1957) to adjust test lengths, within a total time constraint, so as to maximize the differential validity of a battery for a specific time allotment. The tradeoff is between validity and reliability through successive iterations. Changes in predictor reliabilities as a function of estimates of changes in test length (via the Spearman-Brown) are used to recompute the necessary predictor intercorrelation and predictor validity matrices. These are then used in turn to recompute predictor weights that maximize differential validity.

Evaluating All Possible Combinations

It is not possible to choose a combination of predictors that simultaneously optimizes both selection validity and classification efficiency, or absolute validity and differential validity. It is possible, however, to calculate all the indices of test battery performance for every possible combination of subtests that falls within a given test administration time interval.

For example, this method can be used to compute the absolute validity, differential validity, and Brogden index of classification efficiency for every combination of subtests that requires from 134 to 164 minutes of administration time. These combinations of subtests can then be rank-ordered according to each index of test battery performance. For instance, the top 20 test batteries ranked on the basis of maximum absolute validity can be compared to the top 20 test batteries ranked on the basis of maximum differential validity.

An advantage of this method is that it provides explicit information necessary to evaluate tradeoffs: If subtests are included to optimize test battery performance on one index, how will the battery perform on the other indices?

This approach was one of those followed in the Joint Services Enhanced Computer Administered Test validation study designed to evaluate different combinations of 19 ASVAB and ECAT subtests in predicting end-of-training performance (Peterson,

Oppler, Sager, & Rosse, 1993). Data were collected from 9,037 Air Force, Army, and Navy enlistees representing 17 military jobs. The analysis procedures used three time intervals, included two ASVAB subtests (Arithmetic Reasoning and Word Knowledge) in every potential test battery in each of the three time intervals, and evaluated absolute validity, differential validity, classification efficiency, and three types of subgroup differences (White/Black, White/Hispanic, and Male/Female) for each battery in each time interval.

Results indicated that no single test battery (within each time interval) simultaneously optimized all the test battery indices examined. The researchers identified tradeoffs associated with maximizing absolute validity or classification efficiency vs. minimizing all three types of subgroup differences, and with minimizing M-F subgroup differences vs. minimizing either W-B or W-H subgroup differences.

The same approach was used to obtain the results reported in the FY1993 Career Force Annual Report (Peterson et al., 1994). These analyses were done for each of three different Project A/Career Force criterion measures and the sample was limited to Army personnel only.

METHODS FOR ESTIMATING CLASSIFICATION GAINS

Given that a maximizing function, or classification goal, has been decided upon and that a predictor battery has been identified, the next issue concerns how the level or magnitude of classification efficiency can actually be estimated. For example, under some set of conditions, if the full Project A Experimental Battery were used to make job assignments in an optimal fashion, what would be the gain in mean predicted performance as compared to random assignment after selection, or as compared to the current system?

As is the case in other contexts, there are two general approaches to this issue. The first would be to use a statistical estimator that has been analytically derived, if any are available. The second is to use Monte Carlo methods to simulate a job assignment system and compute the gains in mean predicted performance that are produced in the simulation. Each of these approaches is discussed below.

Brogden's Analytic Solution

Work on estimating the value of selection and classification procedures starting with Brogden demonstrated that significant gains from classification are possible, even using predictors of moderate validity. Brogden (1955) provided the rationale for making classification decisions based on mean predicted performance, and for using a classification procedure based on full least square estimates (LSEs) of job performance from a battery of tests. He showed that, given certain assumptions, MPP will be equal to the mean actual performance for such a classification procedure, and that classification based on the full LSE composites produces a higher MPP than any other classification procedure.

The assumptions that limit the generalizability of this result are the following:

(1) The regression equations predicting job performance for each job are determined from a single population of individuals. In practice, this assumption is infeasible because each individual has only one job. Consequently, as Brogden states (1955, p. 249), "Regression equations applying to the same universe can be estimated through a series of validation studies with a separate study being necessary for each job."

(2) There is an infinite number of individuals to be classified. Simulation research by Abbe (1968) suggests that the result is robust with respect to this assumption. However, as discussed later in this report, if the samples of applicants to be assigned are not relatively large, cross-validation of MPP does become an issue.

(3) Relationships between test scores and criterion performance are linear.

Brogden (1951, 1959) provided a method of estimating the MPP of a full LSE classification procedure, based on the number of jobs, the intercorrelation between job performance estimates, and the validity of the performance estimates. The development of this measure is based on the following assumptions:

(1) There is a constant correlation (r) between each pair of performance estimates.

(2) The prediction equations for each job have equal validity (v).

(3) The population of people being assigned is infinite. This assumption is used to avoid consideration of job quotas.

Because the development of the method uses the results of Brogden (1955), those assumptions also apply.

From these assumptions Brogden (1959) showed that the mean predicted performance, expressed as a standard score, is given by the following equation:

$$MPP = v\sqrt{1-r}\ f(m)$$

where f(m) is a function that transforms

$$v\sqrt{1-r} \quad ,$$

the Brogden Index of Classification Efficiency (BCE), into the mean predicted performance standard score as a function of the number of jobs (m) and the selection ratio. The function, f(m), is based on the range of standard scores that would be expected in a sample from a normal distribution as a function of the sample size.

This result has several important implications regarding the determination of MPP. First, MPP is directly proportional to test validity. This result indicates that there can be much more value from tests of limited validity than was indicated by earlier estimates, such as the index of forecasting efficiency or coefficient of determination. Second, since MPP depends on $(1 - r)^{1/2}$, substantial classification utility can be obtained even when predictors are positively correlated. For example, Brogden (1951; adapted by Cascio, 1982) illustrated that using two predictors to assign individuals to one of two jobs can increase MPP substantially over the use of a single predictor even when the intercorrelation between the predictors is 0.8. Third, the results indicate that MPP increases as the number of jobs (or job families) increases. The increase will be a negatively accelerated function of the number of jobs; for example, going from two to five jobs will double the increase in MPP, while going from two to 13 jobs will triple the increase in MPP (Hunter & Schmidt, 1982).

The assumptions of equal predictor validities and intercorrelations are simplifications that allow for an easy, analytical determination of MPP. For more realistic cases in which validities and intercorrelations vary, MPP may be estimated using simulations.

Estimates of Gains From Simulations

The development of Brogden (1959) provides an analytic approach to predicting MPP when certain simplifying assumptions are met. When these assumptions are relaxed, analytical estimation of MPP becomes more complicated. Following methods used by Sorenson (1965), Johnson and Zeidner (1990, 1991), along with several of their colleagues (e.g., Statman, 1992), have applied simulation methods to examine the gains in MPP for a variety of classification procedures. They have used these simulation procedures, which they call synthetic sampling, to examine a number of issues pertaining to estimates of classification efficiency.

There have been several summaries of their work, most notably those of Johnson and Zeidner (1990, 1991); Johnson, Zeidner, and Scholarios (1990); Statman (1992); and Zeidner and Johnson (1991).

Estimating MPP With Synthetic Sampling

Brogden's (1959) formula allows one to estimate the MPP of a classification procedure that is based on full least square estimates (LSE). However, this formula is not appropriate for estimating the MPP of classification policies that are not based on full LSEs. Furthermore, the formula will be inaccurate when the assumptions of equal validities and intercorrelations of the composites are not met. To estimate the MPP of a wider variety of classification procedures using more realistic assumptions, Johnson and Zeidner have relied on a Monte Carlo approach termed synthetic sampling.

Synthetic sampling can estimate the MPP associated with any number of potential selection and classification policies. The basic approach is to evaluate classification methods based on random samples from a theoretical (multivariate normal) distribution

representing the overall population of test scores and job performance criterion scores. Three classes of distributions are generated:

(1) One sample is used to develop the prediction equations that form the basis of the assignment procedures. The prediction equation may be based on LSEs, Aptitude Areas, or other combinations of the predictor variables. Several alternative prediction functions can be developed from this sample, depending on the experimental design.

(2) A second class of samples is used to apply one or more selection and classification assignment procedures. The assignment procedures could range from simple random assignment to optimal assignments using linear programming methods. These samples represent applicants who must be assigned to individual jobs, based on the procedures developed using the first sample. In effect, they are cross-validation samples for the predictor weights computed on the first sample. Usually several samples are used. Each assignment procedure is used for each sample, producing a repeated measures design.

(3) A third sample (or population distribution) provides the weights used to estimate MPP for each of the assignment methods after the assignments have been made. The weights are calculated directly from the parameters of the distribution that is used to generate the samples used for developing and applying the selection and classification methods. The population parameters, in turn, are inferred from empirical predictor intercorrelations and validity measures after correction for restriction in range. In effect, the corrected covariance structure of the complete Project A/Career Force database is used as a model for a simulated population. The "population" weights are applied to the scores of each simulated applicant to determine the performance in the job assigned by each classification procedure. In this way the MPP was calculated for each of the samples (among the second class of samples) for each of the candidate classification procedures.

Analyses of synthetic sampling data compare the MPP for different assignment methods. The greatest MPP would occur if the population weights themselves were used to make optimal assignments using linear programming. Other assignment strategies produce lower MPP values for two reasons. First, the assignment weights are based on a sample from the population rather than from the actual population parameters. Second, all of the assignment strategies except those based on full LSEs and linear programming are special cases of the optimal assignment strategy. That is, they reduce the number of factors considered or otherwise restrict the values for some of the weights of the composites used to predict performance. The variance of individual performance levels around the full LSE for the population enters into the analysis indirectly. If the variability is high, then the samples generated for development and application of the assignment methods will be dissimilar to the population values and to each other, thus producing lower MPP values. The synthetic sampling method assumes that the linear model is accurate. That is, there are no nonlinearities, and the distributions are all

normal. Evidence reviewed by Hunter and Schmidt (1982) suggests that these assumptions are reasonable.

Results of Simulation Runs

All of the empirical comparisons that have been made used the Project A database and show the improvement in MPP in standard units, that is, as a proportion of the standard deviation of the MPP distribution. Each experimental condition was investigated with several synthetic samples (usually 20). The researchers used the means and standard deviations of the MPP values, calculated over the 20 samples, to form the basis of statistical tests of the significance of improvements in MPP resulting from the experimental conditions, usually compared to current assignment methods. Standard errors were typically very small, and nearly all differences were significant.

Full LSE composites lead to an increase in MPP of about 0.15 SD when compared to current methods. Enforcing current Army quality distribution goals has little impact on MPP. Increasing the number of predictor tests, the number of job families, and the number of factors in composites, as well as decreasing the selection ratio, all improve MPP substantially. The test selection method, job clustering method, and overall selection and classification strategy have much smaller effects.

These results represent substantial potential improvements in classification efficiency, given a particular set of parameter values for things such as the number of jobs or job families, the dimensionality of the joint predictor criterion space, and the level of criterion intercorrelations for pairs of jobs. The actual improvement obtained by implementation of specific classification procedures will be less because of the damping influence of various constraints. For example, specific quotas, such as the number of training seats available, will limit the extent to which individuals can be placed in jobs that maximize MPP, or assignment to the optimal jobs may leave training seats unfilled, thus increasing training cost. Alternately, if applicants have significant latitude in job choice, they may not elect to take their best person-job match. The extent to which system constraints will reduce the expected performance gains below the maximum possible or increase the cost required to obtain these gains is not known, and should be a future research topic. Further illustration of these issues, in the context of using Army accessions data to simulate specific job assignment strategies, is given by Nord and Schmitz (1991).

ALTERNATIVE DIFFERENTIAL JOB ASSIGNMENT PROCEDURES

The preceding section reviewed methods for estimating the degree of classification efficiency, or the degree to which a particular classification goal (e.g., MPP) can be increased as a result of a new classification procedure. The methods used to estimate the expected payoff in the population and the procedures actually used to make job assignments encompass two different sets of issues. That is, the estimation methods portray the maximum potential gain that can be achieved, given certain parameters, but

the ability of the real-world decision-making procedures to realize the gain is another matter.

This section briefly describes the methods the Services currently use to assign applicants to jobs, and then discusses the development of two future systems for making differential job assignments: the Army's Enlisted Personnel Allocation System (EPAS) and the Air Force's revised Processing and Classification of Enlistees (PACE) system. The overall goal of the new systems is to realize more of the potential maximum gain than is captured by the current systems.

Current Methods

All Services assign applicants to either an occupational area or a specific job at the MEPS. Although the process differs somewhat across the Services, generally a career counselor, or classifier, reviews the recruit's aptitude scores, medical history, and educational records. The counselor uses a computer system to obtain a list of current and future technical school vacancies and specialties, in order of Service priority, that match the applicant's records. Applicants and counselors discuss the job options, and the applicant makes the final decision about enlistment (Camara & Laurence, 1987).

Aptitude scores are an important component in each Service's assignment/ classification system. Each Service has established minimum cut scores for each of its jobs or occupational areas on one or more of its composites to ensure a minimum level of aptitude for each job. Additionally, each Service uses aptitude scores to match people to jobs. However, the way in which this "match" is made and the type of information that goes into the "matching" process vary considerably by Service. The actual assignment of recruits to occupational areas or jobs is accomplished via computerized Person Job Match (PJM) algorithms. Each Service has its own algorithm, which reflects its current policies toward the relative priorities of filling jobs at any point in time. A brief overview of each algorithm is provided below.

Air Force Allocation Systems

The Air Force has two PJM systems. At the MEPS, the Procurement Management Information System (PROMIS) is used to make pre-enlistment assignments into either (a) specific jobs, Air Force Specialties (AFSs), through the Guaranteed Training Enlistment Program (GTEP), or (b) one of four occupational areas: Mechanical, Administrative, General, or Electronic (MAGE). Currently, about 30 to 40 percent of recruits are assigned into AFSs at the MEPS; 60 to 70 percent enter the Air Force with a guaranteed MAGE area. During Basic Military Training (BMT), recruits originally classified by PROMIS into one of the four MAGE areas are classified by the Processing and Classification of Enlistees (PACE) system into a specific AFS within the pre-assigned MAGE area.

The Air Force assignment variables differ from those used by other Services in two ways. First, minimum physical strength requirements exist for many AFSs. Second, recruits indicate occupational preference by weighting (on a 0 to 9 scale) the M, A, G,

and E areas. After all data are input to PROMIS, the program checks to ensure the applicant is eligible for the Air Force, identifies AFSs for which the applicant is eligible, and generates a relative payoff index (with a maximum of 1.000 points) that reflects the value of assigning the recruit to each AFS. PROMIS then compares the payoff index with the Air Force's current need to fill AFSs (based on training seat vacancies) and develops an ordered list of up to 16 AFSs. The first AFS is the "best choice" for both the individual and the Air Force (Pina, 1988). The specific functions that lead to the ordered list are summarized below.

Five components enter the PROMIS payoff algorithm to form the payoff index: (a) variable fill versus aptitude/difficulty, 600 points; (b) predicted technical school success, 50 points; (c) M, A, G, and E area preference, 180 points; (d) minority/non-minority, 70 points; and (e) constant fill, 100 points (Pina, 1988). Variable fill is an index of the Air Force's needs at a particular point in time (i.e., number of personnel needed and the time remaining to fill the AFS). The aptitude/difficulty subcomponent matches individual aptitude to the level of aptitude required by the job (i.e., job difficulty). Variable fill and aptitude/difficulty interact so that aptitude/difficulty receives a larger allocation of the 600 points if the Air Force's need for recruits is being met, and vice versa. The technical school success component is based on regression equations for predicting technical school grades from AFQT, M, A, G, and E composites, and binary variables representing high school courses taken. The area preference component assigns points to M, A, G, and E areas in proportion to the applicant's preference. When PROMIS was originally developed, the minority/nonminority component was designed to help meet the Air Force minority representation goals set for each AFS.

The current PACE is a simple, nonoptimal system that processes recruits in batch (i.e., non-sequential) mode (Pina, 1988; Pina, Emerson, Leighton, & Cummings, 1988). It sorts recruits into available training seats on the basis of the recruit's (a) preference for the AFS, (b) ASVAB scores, and (c) gender.

Army Allocation Systems

The Army currently uses a computerized reservation, monitoring, and PJM system labeled REQUEST (Recruiting Quota System). A new assignment procedure, the Enlisted Personnel Allocation System (EPAS), was developed in a research effort known as Project B, but has not yet been implemented. The Army has no post-enlistment PJM system because specific jobs -- Military Occupational Specialties (MOS) -- are guaranteed to all enlistees prior to enlistment.

The Army does not typically use job-occupational preference for assignment and, aside from the gender-exclusive policy prohibiting women from combat jobs, it has no minority fill component. Data on physical condition (as portrayed by the PULHES measures) may be considered for some Army jobs; for example, superior physical condition or color blindness may enter into selection for specialized MOS.

Using functions related to these goals, REQUEST computes an MOS Priority Index (MPI) that reflects the degree of match between the applicant and the MOS and uses the MPI to produce a list of MOS in order of Army priority. The functions involved in the MPI computation can be grouped into two broad categories: (a) MOS Status (MS) functions that define the Army's need to fill a particular MOS, and (b) Applicant Qualification (AQ) functions that define the degree to which the applicant is matched to the MOS. The program first lists the five MOS that are highest in priority, and the classifier encourages the applicant to choose one of them. If the applicant is not interested in these jobs the next five high priority jobs are shown and so on until the applicant chooses a job (Camara & Laurence, 1987; Schmitz, 1988).

Marine Corps Allocation Systems

Like the Air Force, the Marine Corps has two PJM systems for assignments. ARMS (Automated Reservation Management System) is used at the MEPS to assign applicants to either specific MOS or one of 35 occupational areas. Currently, only about 2 percent of recruits enter the Marine Corps with a guaranteed MOS. Nearly 85 percent are guaranteed an occupational area, and about 14 percent enter under an "open contract," with no occupational assignment. Most Marines are assigned to specific MOS after BMT; the Recruit Distribution Model (RDM) is used to make these post-enlistment assignments.

The Marine Corps uses its assignment system differently from the way in which the other Services use theirs. For most Services, occupational preference, if considered at all, is a piece of information in the algorithm with a known weight; the algorithm produces a list of options from which the applicant selects an occupational area or job. The Marine Corps relies more heavily on its counselors to assess job interests.

Marine Corps applicants and classifiers talk about the applicant's interests. The classifier obtains a list of the applicant's preferences and calls the ARMS operator who, in turn, enters the applicant's data into ARMS. The ARMS operator checks to see whether the applicant can be assigned to his/her first preference. If not, the process is repeated until either a match is made or the applicant decides to enter the Marine Corps under an open contract. In short, occupational preference starts the assignment process. The ARMS algorithm ensures that applicants meet minimum qualifications for chosen MOS/occupational areas, fills available training seats according to Marine Corps priorities, and ensures that minority representation goals are met.

RDM is a batch-mode system used at Recruit Training Centers (RTCs) to assign recruits to job categories. RDM first fills jobs in accordance with the Marine Corps needs while meeting minority representation goals for jobs. After these two constraints are satisfied, the algorithm maximizes (a) the average probability of success in training, and (b) the number of recruits assigned to the highest prerequisite levels within each job category (Kroeker, 1989).

Navy Allocation Systems

The Navy's pre-enlistment assignment system, the Classification and Assignment within PRIDE (CLASP), works much like the Air Force's PROMIS, from which it was derived. At the MEPS, CLASP is used to make pre-enlistment assignments either into specific Navy jobs (or ratings) or into apprenticeship or general detail assignments. Currently, about two-thirds of Navy recruits are assigned to guaranteed training slots for specific Navy ratings. About one-third of the new recruits receive an apprenticeship or a General Detail assignment as Seaman, Airman, or Fireman. The Navy uses a post-enlistment system, Computer Assisted Assignment System II (COMPASSII), to assign recruits to ratings during BMT.

After data are input to CLASP, the components of the payoff algorithm are (a) predicted training success, (b) technical aptitude/job complexity, (c) Navy priority/ individual preference, (d) minority fill rate, (e) fraction fill rate, and (f) probability of attrition. School success is the predicted final grade based on ASVAB composite scores. Technical aptitude/job complexity is a numeric value for the expected relative utility of matching the level of individual aptitudes to the level of job complexity; assignments that match on these two variables receive a higher value, and the value is proportionally higher if the match is for more complex jobs. Navy priority/individual preference is an index of the relative value of assigning a recruit to ratings that vary in terms of the correspondence between the rating's Navy priority and the individual's preference. The minority fill rate component is designed to help the Navy meet minority representation goals for each rating. The fraction fill rate component evens the flow of allocations into ratings over the course of the recruiting month; that is, it gives utility points to ratings that have below average assignment rates. The attrition component is an estimate, based on demographic data, of the probability of retention during the initial service term and costs to the Navy for personnel loss (risk) for each rating (Kroeker, 1988, 1989; Kroeker & Folchi, 1984; Kroeker & Rafacz, 1983).

During the fifth week of recruit training, the Navy uses COMPASSII, in conjunction with a classification interview, to assign recruits to ratings. The interviewer recommends five occupational groups based on the recruit's ASVAB test scores, job experience, background, and preferences. After data are entered, COMPASSII conducts a series of optimizations, each one constraining its predecessors. COMPASSII goals, in order, are to (a) maximize the utilization of training seats, (b) minimize transportation costs, (c) match the interviewer's recommendations, and (d) maximize the probability of success in training schools (Hatch, Pierce, & Fisher, 1968; Kroeker, 1989).

Summation

There are similarities among the classification systems used by the Services. They all ensure adherence to minimum aptitude standards for each job, and all are designed to maximize the utilization of training school vacancies across jobs. Pina (1974) and Kroeker (1989) distinguish classification systems in terms of how they fill training seats. Systems that fill training seats (or vacancies) from available resources (within the constraint that each individual meets minimum job requirements) are "fill" oriented. "Fit"

oriented systems match individual aptitudes and/or preferences to the jobs/available seats. The batch-mode, post-enlistment processing systems used by the Marine Corps (RMD) and the Navy (COMPASSII), for example, are driven primarily by fill policy (Kroeker, 1989). PROMIS and CLASP are examples of fit-oriented systems.

NEW METHODS FOR MAKING JOB ASSIGNMENTS

None of the current systems represent "true" classification in the sense that an entire set of job assignments is made as a batch such that the goal of classification (e.g., MPP) is maximized. All current systems seek to insure that one or more limited goals for each job are met even though the resulting assignments are suboptimal in terms of maximizing total gain. However, two new experimental systems have been developed which attempt to incorporate a true classification component as part of the assignment algorithm. They are the EPAS system developed by the Army and the new PACE system developed by the Air Force.

Enlisted Personnel Assignment System (EPAS)

The Army's recently developed Enlisted Personnel Assignment System (EPAS) optimizes several functions simultaneously. The system is designed to (a) maximize expected job performance across MOS, (b) maximize expected service time, (c) provide job fill priority, and (d) maximize reenlistment potential. EPAS was designed to support Army guidance counselors and personnel planners (Konieczny, Brown, Hutton, & Stewart, 1990; Rudnik & Greenston, in press).

The following maximization problem provides a heuristic for understanding the view of the classification process taken by EPAS:

Maximize

$$Z=\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}X_{ij}$$

subject to

$$\sum_{i=1}^{n} X_{ij} = 1$$

$$\sum_{j=1}^{n} X_{ij} = 1$$

where the variables, i and j index the applicants and jobs, respectively. If the matrix $X_{ij} = 1$, then applicant i is assigned to job j. The two constraints specify that each job is filled by a single applicant, and each applicant is assigned to a single job, respectively. The variable c_{ij} is a weight that represents the value of assigning applicant i to job j.

However, there are many factors that make the problem more complex than is indicated in the equation, including sequential processing of applicants, an applicant's

choice of suboptimal assignments, complications caused by the Delayed Entry Program (DEP), and temporal changes in the characteristics of the applicant population. Consequently, the optimization approach taken by EPAS is considerably more complex than the simple formulation shown above.

For example, in "pure" classification the optimal allocation of individuals to jobs requires full batch processing, but in actual applications applicants are processed sequentially. EPAS attempts to deal with this complication by grouping applicants into "supply groups" defined by their level of scores on the selection/classification test battery and by other identifiers, such as gender and educational level. For a given time frame the forecasted distribution of applicants over supply groups is defined, and network or linear programming procedures are used to establish the priority of each supply group for assignment to each MOS. For any given period, the actual recommended job assignments are a function of the existing constraints and the forecast of training seat availability.

Consequently, the analyses performed by EPAS are based on the training requirements and the availability of applicants. EPAS retrieves the class schedule information from the Army Training Requirements and Resource System (ATRRS), and provides this information, along with the number of training seats to be filled over the year, to the decision algorithm. It then forecasts the number and types of people who will be available to the Army in each supply group over the planning horizon (generally 12 months). The forecasts are based on recruiting missions, trends, bonuses, military compensation, number of recruiters assigned, characteristics of the youth population, unemployment rates, and civilian wages.

Based on the requirements and availability information, EPAS performs three kinds of analysis: (a) planning and policy analysis, (b) simulation analysis, and (c) operational analysis. The first two of these analyses are designed to aid personnel planners, while the third primarily supports Army guidance counselors.

The planning and policy analysis allocates supply group categories to MOS over a 24-month planning horizon, including both direct enlistment and delayed entry. The allocation is based on large-scale linear programming that sets a priority on MOS for each supply group. The analysis is used primarily for evaluating alternative recruiting policies, such as changing recruiting goals or delayed entry policies. The alternative maximizing functions that can be used to determine the optimal allocation include expected job performance, the utility of this performance to the Army, and the length of time that the person is expected to stay in the job. Other goals include minimizing DEP costs, DEP losses, training losses, and recycles. Constraints include applicant availability, class size bounds, annual requirements, quality distribution goals, eligibility standards, DEP policies, gender restrictions, priority, and prerequisite courses.

The simulation analysis mode provides a more detailed planning capability than is possible with the policy analysis mode. The simulation analysis makes individual MOS assignments for the current month and produces detailed output describing the flow of applicants through the classification process. The simulation analysis is based on the

same linear programming optimization that is used for policy analysis; after the current month's assignments are simulated, the 24-month planning window is then moved ahead one month and another aggregate supply group allocation is made.

The operational analysis component is similar in structure to the planning made. In the operational mode, EPAS output feeds into the REQUEST system and provides counselors with a list of the MOS that are best suited to each applicant and the training dates for each supply group. The primary differences between the operational analysis and the policy analysis components are that the operational analysis allocates individual applicants to jobs, rather than supply groups, and performs sequential allocation of applicants. The operational module uses the lists of MOS provided by the planning analysis as the basis of its allocation procedure.

The ability of EPAS to "look ahead" derives from the interactions between the planning/policy analysis over the planning horizon and the operational analysis. The planning analysis provides an optimal allocation over a 24-month period. This solution is one input to the sequential classification procedure used by the operational analysis. Individual assignments of MOS to an individual are scored according to how close they are to the optimal solution. Highly ranked MOS are those that are in the optimal solution. MOS that are lower ranked would increase the cost (reduce the utility) of the overall solution. The MOS are ranked inversely according to this cost.

The New PACE

The Air Force has developed a new microcomputer-based version of the PACE classification algorithm. The PACE algorithm includes the components of PROMIS plus some additions. Supplementary PACE functions are designed to (a) improve the fit between occupational preferences and assignment by improving occupational interest measurement, (b) take training costs into account, (c) minimize unproductive lag time (also called casual time) between BMT graduation and technical school entry, and (d) minimize first-term attrition.

The ten components of the PACE algorithm are (a) aptitude (M, A, G, and E composites), (b) job difficulty, (c) predicted technical school grade (based on ASVAB composite scores), (d) academic background (the percentage of desirable high school courses completed), (e) occupational interest (based on the Air Force Vocational Interest Career Examination, VOICE), (f) restricted interest (the recruit's rankings of available jobs), (g) training cost, (h) the probability of retention during the first term of enlistment, (i) casual time (the number of days between BMT graduation and technical school entry), and (j) fill priority (the relative urgency of filling the AFS) (Pina, 1988; Pina et al., 1988).

The PACE payoff algorithm was developed using the "policy specifying" (Ward, 1977) approach that was also used to develop the payoff algorithm for PROMIS. The technique uses SMEs, classification experts, and policy makers, to define a post-enlistment classification policy for non-prior service airmen. The methods were designed to be similar to PROMIS and to avoid generating new data requirements.

The algorithm is based on a person-job match (PJM) metric that combines the ten fundamental classification criteria organized into a hierarchical taxonomy. Six of the criteria address the effectiveness issues, that is, aptitude, interest, trainability, and so forth. The other four criteria are concerned with efficiency issues, such as cost, time, and fill priority.

The first-level criteria are combined using the agreed-upon combinatory functions to produce composite measures of effectiveness and efficiency. The effectiveness and efficiency measures are combined linearly to produce the individual's predicted score for each job; this score is used to make assignments that optimize the PJM. The relative weights given to effectiveness and efficiency in this combination are determined "by management at run time" (Pina et al., 1988, p. 8). The assignment rules used to maximize the overall classification payoff (i.e., the PJM) are then computed, using linear programming optimization methods.

Since all information about both predicted goal outcomes (e.g., predicted training success) and constraints has been combined into one composite score, the solution for the optimal PJM becomes the familiar linear programming assignment problem.

PACE and EPAS differ in a number of respects. Perhaps the most distinctive is that EPAS attempts a simultaneous solution for the maximizing functions and constraint equations. PACE uses a much more compensatory model and combines almost all predictor and constraint information into one index before optimization takes place.

Cost-Performance Tradeoff Model

Research by McCloy and associates (1992) developed a cost-performance tradeoff model that combines selection and classification functions. The goal of this model is to minimize the cost required to obtain a specified performance level. Thus, in contrast to other evaluation or allocation methods, predicted performance is considered a constraint in the model, and cost minimization is the objective.

The model considers costs involved in recruiting, basic training, initial skill training, and compensation over the first term of enlistment. Performance estimates are weighted by the predicted survival probability, by month, over the first term. Thus, predicted attrition is incorporated in the estimation of both cost and performance. Separate submodels predict performance based on individual and job characteristics, calculate recruiting cost, estimate survival rates, and assess training and compensation costs.

The model uses quadratic programming methods to determine the selection and allocation strategy that minimizes the cost required to meet the required level of expected performance. One advantage of this method is that it does not require that performance be measured on a monetary scale. The optimization method can consider accession limits and quality distribution requirements. It does not consider other constraints, such as training seat fill requirements or casual time between basic training and initial skill training.

McCloy et al. compared the prescriptions of the model to actual FY 1990 accessions for the Army and the Navy. They found that actual accessions were close to the values prescribed by the model. The estimated cost for the optimal policy was about 1 percent lower than that for the actual policy, leading to a predicted cost savings of $72 million for the Army and $31 million for the Navy. Considering quality goals had little impact on the cost of the optimal solution.

Comparison of Models

Both EPAS and PACE represent many of the goals of the classification process. However, the two methods represent these goals in different fashions. The PACE algorithm combines all goals into a single objective function which it then maximizes. EPAS treats predicted performance as the objective function, and other variables enter the model as constraints. The cost-performance tradeoff model of McCloy et al. (1992) does not consider the range of goals addressed by the other two allocation methods. However, it addresses cost specifically, in a way the other models do not. The many specific differences between the methods preclude a comparison of the pros and cons of the approaches. However, it would be possible to compare the methods using simulation studies.

SUMMARY

The overall goal of this chapter has been to present a full consideration of all the major issues involved in modeling the personnel classification problem, estimating its critical parameters, and evaluating the results of using specific job assignment procedures. At least three major research questions emerged:

(1) To achieve maximum gains from a classification strategy, how should the prediction equations for each job assignment be developed? To what extent is the appropriate procedure a function of the goal of classification or the specific set of constraints within which the classification strategy must operate?

(2) How can the aggregate payoff, or maximum potential gain, from classification best be estimated? Is it possible to make such estimates analytically, or are Monte Carlo/simulation procedures the only realistic alternative?

(3) How much of the potential gain from classification, as compared to alternatives, can be realized by specific job assignment methods in an operational setting? What components of an assignment system are the most critical for being able to capture the maximum potential gain? What situational constraints have the most serious effect on the capabilities of job assignment procedures?

The Project A/Career Force database presents a unique opportunity to address these research questions. In the next chapter the issue of estimating the gain from classification is addressed and several alternative estimators are compared, including a hybrid method that has recently been developed. The last chapter in this final report will propose the building of an Army personnel system test bed, based on the Project A/ Career Force database, that can be used to evaluate systematically the effects of alternative (a) prediction equations, (b) system constraints, (c) cost/performance tradeoffs, (d) assignment models, (e) operational assignment procedures, and (f) labor market conditions on alternative classification goals.

Chapter 8
ESTIMATING CLASSIFICATION GAINS: DEVELOPMENT OF A NEW ANALYTIC METHOD

Rodney L. Rosse, Norman G. Peterson, and Deborah Whetzel

Classification efficiency is defined as the improvement in expected job performance that could be obtained by optimal classification of applicants to jobs (in contrast to random assignment) using predicted job performance scores derived from the ASVAB and Experimental Battery test scores. Two metrics, or objective functions, that can be used to define classification efficiency are mean predicted performance and mean actual performance.

Mean Predicted Performance (MPP) is the mean predicted performance score for a sample of applicants assigned to a set of jobs using a particular classification strategy (e.g., assigning individuals to jobs using highest predicted performance score or optimizing use of linear programming). Mean Actual Performance (MAP) is the mean actual performance score for a sample of applicants actually assigned to a set of jobs using a particular classification strategy. In a validation context MAP is unknown because every individual cannot work on every job. However, once the prediction functions are known, MPP can be calculated for each person on each job. MAP could be estimated directly via simulation by defining a population, sampling applicants from the population, and calculating the "actual" performance scores for each applicant on each job. MAP is the value of real interest; MPP serves as an estimate of it.

An assumption underlying much classification research is that mean predicted performance and mean actual performance are equal when (a) least-squares equations are used to develop the predicted performance scores on very large samples, (b) assignment is made on highest predicted scores, and (c) no quotas for jobs are used (Brogden, 1959). Since these conditions rarely exist in actual practice, this chapter describes an alternate method for estimating classification efficiency.

The remainder of this chapter is divided into three sections that serve distinctly different purposes:

(1) Statistical development of two important estimators, eMPP and reMAP.
(2) Empirical tests of eMPP and reMAP.
(3) Application of the estimates to weighting and variable selection issues pertinent to the Career Force project.

The first section pertains to the statistical definition of two estimators of classification efficiency. One estimator (reMAP) is for estimating actual means of performance for a future group of applicants that would be assigned to the nine Batch A MOS jobs. The second (eMPP) is a refined estimator of the means of predicted performance for the same applicants. Both estimators purport to be indices of classification efficiency.

The second section describes a set of empirical, Monte Carlo experiments designed to test the accuracy of the two estimators, eMPP and reMAP. Such a demonstration is necessary because the estimators are novel and have been recently developed by the authors. Accordingly, they cannot be assumed to be acceptable without such a demonstration.

The third purpose bears on the important practical issues of the Career Force project. Specifically, it uses the developments described in the first two sections to estimate the potential classification efficiency under various conditions that could plausibly be chosen for implementation by the Army.

STATISTICAL BASIS OF eMPP AND reMAP

Situation

A sample consisting of one or more job applicants may be randomly selected from a population of applicants. For each applicant, the observed column vector of predicted performance scores,

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ \hat{y}_P \end{bmatrix} \quad (1)$$

has a corresponding vector of future actual performance scores.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_P \end{bmatrix} \quad (2)$$

None of the actual performance scores are observable at the point of application.

The vector of scores, $\hat{Y}$, contains predicted performance scores $\hat{y}_i$ (i.e., $\hat{y}_i ... \hat{y}_p$), with one score for each of P jobs. Commonly, the values of $\hat{y}_i$ will have been determined in previous validation research by empirically fitting individual prediction functions with a function for each of the P jobs. The prediction functions may use a common set of predictor variables such as the scores from a battery of tests. (Note: The observed value

of $\hat{y}_i$ is only defined as a predictor of the criterion, y_i, and not necessarily an optimal predictor or one that was obtained from any particular methodology, such as least-squares regression.)

Given the situation in which predicted performance scores for several jobs are available for applicants, one may elect to assign each applicant to the job where the highest predicted performance is observed. This simple assignment strategy was proposed by Brogden (1955). More complex strategies may involve the simultaneous assignment of individuals in a sample of applicants. For instance, optimizing techniques such as linear programming have been proposed to accomplish assignment under constraints such as incumbency quotas for individual jobs, differential job proficiency requirements for specific jobs, and minimum work-force standards (McCloy et al., 1992).

The purpose of this section is to clarify the issues of how group means of predicted performance relate to the actual criterion performance of the corresponding groups when an assignment strategy is employed.

Relevant Parameters of the Applicant Population

In the population of applicants, the vector of predicted performance, $\hat{Y}$, has a multivariate distribution. Its variance-covariance matrix, $\boldsymbol{\Sigma}_{\hat{y}}$, is symmetric about the diagonal, that is,

$$\boldsymbol{\Sigma}_{\hat{y}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdot & \sigma_{1P} \\ \sigma_{21} & \sigma_{22} & \cdot & \sigma_{2P} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{P1} & \sigma_{P2} & \cdot & \sigma_{PP} \end{bmatrix} \qquad (3)$$

so that $\sigma_{ij} = \sigma_{ji}$, and the correlation between the ij-th pair of predictors is

$$r_{ij} = \sigma ij \, [\sigma_{ii} \, \sigma_{jj}]^{-1/2}$$

The population means of $\hat{Y}$ are

$$E(\hat{Y}) = \begin{bmatrix} E(\hat{y}_1) \\ E(\hat{y}_2) \\ \cdot \\ \cdot \\ E(\hat{y}_P) \end{bmatrix} \qquad (4)$$

The unobservable vector of actual performance, Y, also has a multivariate distribution. For purposes here, each value, y_j, in this actual performance vector is a standard score with the expectation of zero and variance of one.

Additionally, predicted performance, $\hat{Y}$, is linearly related to actual performance, Y. The term most commonly applied to characterize the magnitude of the relationships is "validity," which is the correlation between each predicted performance score, $\hat{y}_i$, and each actual performance score, y_i. Thus, there is a symmetrical matrix, V, of (P x P) validities so that

$$V = \begin{bmatrix} \nu_{11} & \nu_{12} & \cdot & \nu_{1P} \\ \nu_{21} & \nu_{22} & \cdot & \nu_{2P} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \nu_{P1} & \nu_{P2} & \cdot & \nu_{PP} \end{bmatrix} \tag{5}$$

where each row represents the validity of the i-th predicted performance score for predicting the actual performance for the j-th job in the applicant population. The covariance of the i-th predicted with the j-th actual performance score is $\nu_{ij}\ \sigma_{ii}^{1/2}$. The matrix of covariances, $\Sigma_{\hat{y}y}$, is

$$\Sigma_{\hat{y}y} = Dg\{\Sigma_{\hat{y}}\}^{1/2}\ V \tag{6}$$

where $Dg\{\Sigma_{\hat{y}}\}^{1/2}$ is a diagonal scaling matrix consisting of standard deviations on the diagonal and off-diagonal zeros.

Mean Values of Predicted Performance

Suppose that a sample of applicants has been assigned to the P number of jobs according to a chosen assignment strategy. There exists a matrix of means of the observed values of predicted performance scores for each job. For each job, there is a group of n_j individuals assigned so that

$$m_{ij} = \sum_{k=1}^{n_j} \hat{y}_{ijk} / n_j \tag{7}$$

where $\hat{y}_{ijk}$ is the i-th predicted performance score of the k-th individual assigned to the j-th job. Thus, the observed values of mean predicted performance consist of a P by P matrix, $M_{\hat{y}}$, of means where the rows represent predictors and the columns represent jobs:

$$M_{\hat{y}} = \begin{bmatrix} m_{11} & m_{12} & . & m_{1P} \\ m_{21} & m_{22} & . & m_{2P} \\ . & . & . & . \\ . & . & . & . \\ m_{P1} & m_{P2} & . & m_{PP} \end{bmatrix} \quad (8)$$

Brogden (1955) suggests that the mean of the diagonal elements of this matrix, called mean predicted performance (MPP), indicates the classification efficiency realized by applying the assignment strategy based on the predicted performance scores of applicants:

$$MPP = \sum_{j=1}^{P} n_j m_{jj} / \sum_{j=1}^{P} n_j \quad (9)$$

Classification Efficiency Defined in Terms of Actual Performance: Definition of the Re-estimate (reMAP)

The enhancement of predicted performance scores through an assignment strategy is of value to the extent that future actual performance is correspondingly enhanced.

With respect to actual performance, Brogden (1955) argued that MPP is a satisfactory approximation of the expected actual performance in standardized units (Mean = 0, SD = 1). Using a limiting case argument, he contended that, as the sample sizes on which least-squares prediction equations are estimated become very large, the resulting prediction composites asymptotically approach the expected values of actual criterion performance. Brogden gave no additional consideration of actual performance.

This argument has been recently cited by Scholarios, Johnson, and Zeidner (1994) with the implication that MPP is approximately the same as mean actual performance (MAP) where samples used to develop the prediction equations ranged in size from about 125 to 600. It is not clear that MPP is an unbiased estimator of MAP where developmental samples are so small.

A simple case in which an applicant is both randomly drawn and randomly (or arbitrarily) assigned to the j-th job demonstrates this issue. The expected actual performance in the j-th job of such an applicant is

$$E(y_j) = v_{jj} z_{\hat{y}_j} \quad (10)$$

where

$$z_{\hat{y}_j} = [\, \hat{y}_j - E(\hat{y}_j) \,] / \sigma_{jj}^{1/2}$$

and y_j is the actual performance of the applicant for the j-th job. $\hat{y}_j$ is the corresponding observed predicted performance which has the validity v_{jj} (from equation 5), a population mean of $E(\hat{y}_j)$, and a standard deviation of $\sigma_{jj}^{1/2}$.

Equation 10 expresses the extent of regression that is to be expected by conditioning a prediction of actual performance on the observed value of predicted performance. However, it is true only for an applicant who is randomly assigned to the j-th job. When an assignment strategy is applied, the observed values of $\hat{y}_j$ are not random but, rather, are conditioned on the observed values of predicted performance for all P of the jobs. Thus, additional conditions are placed on the expected value, $E(y_j)$.

To illustrate the issue, consider a simple case of assignment based on Brogden's strategy of assigning each applicant to the job with the highest observed $\hat{y}_{ij}$. For this hypothetical case, there are two predictors for two corresponding jobs. The variance of each predictor is .75 and the covariance is .375. Thus, each predictor has a standard deviation of .866 and the correlation between them is .50. Each of the two variates has a mean of zero.

Figure 8.1 is a bivariate scatterplot of 10,000 points in this population of applicants. The two predictors are normally distributed. The diagonal line in Figure 8.1 is drawn so that $\hat{y}_1$ is greater than $\hat{y}_2$ for all points that are below the line and $\hat{y}_2$ is greater than $\hat{y}_1$ for all points above the line. Thus, any selected point appearing below the line would be assigned to job 1 while any selected point appearing above it would be assigned to job 2.

Because this simple case meets all of Brogden's assumptions, the expected mean predicted performance for each of the two jobs is the same and may be obtained using Brogden's allocation average (Brogden, 1955). The allocation average is $R(1\text{-}r)^{1/2}$ A, such that R = .866, r = .5, and the tabled adjustment, A (referred to as f(m) in Chapter 7), is .564. The allocation average is .345. The expected values of randomly selected observations, $\hat{y}_1$ and $\hat{y}_2$, for the assigned applicants in each of the two jobs are as shown:

	Job	
Predictor	1	2
1	.345	-.345
2	-.345	.345

The expected mean of $\hat{y}_1$ for the applicants assigned to job 1 is .345. The corresponding expected mean of $\hat{y}_2$ for job 1 is less because of the conditions of the assignment strategy. In fact, in this simplified case, the expected mean of predicted performance for the predictor not targeted for each job is negative.

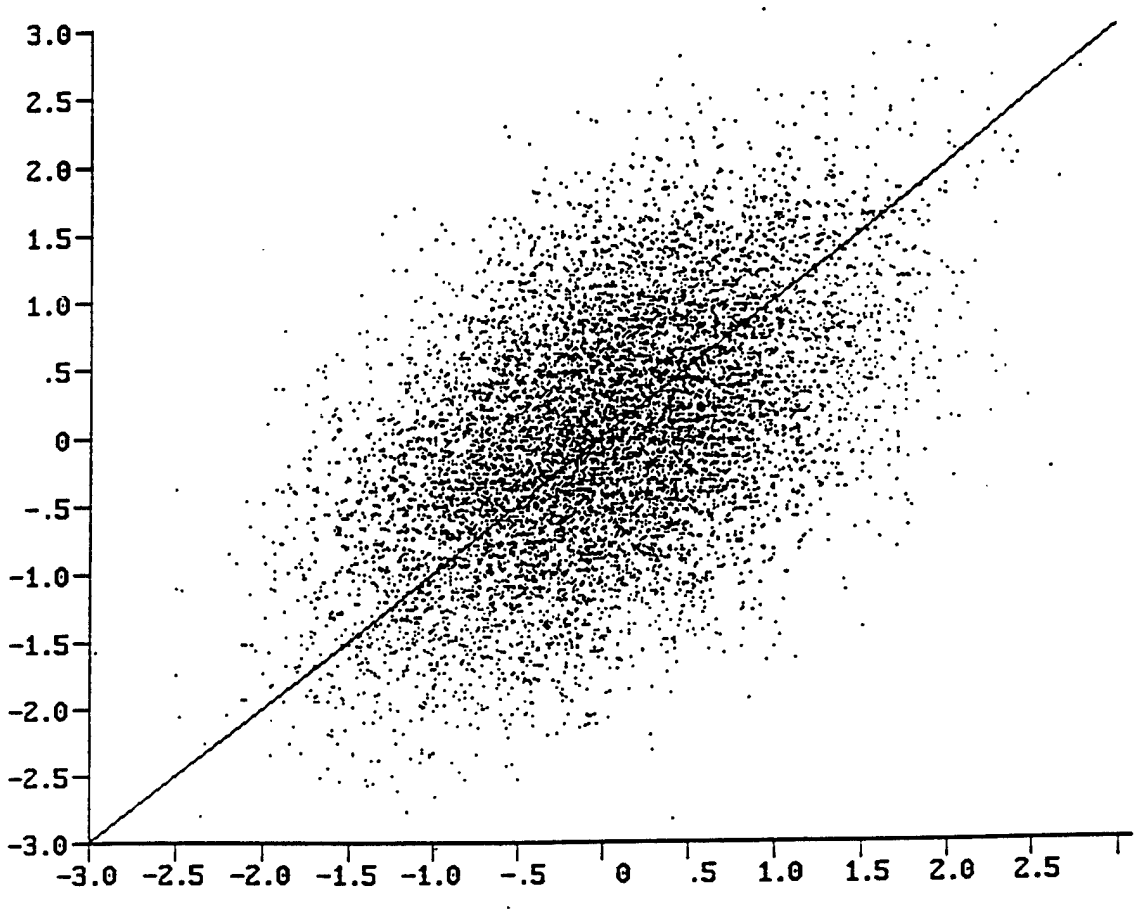


Figure 8.1. Bivariate scatterplot of 10,000 points in a population of applicants.

Unless the validity of $\hat{y}_2$ for predicting job 1 and the validity of $\hat{y}_1$ for predicting job 2 are both zero, there is a conflict. The conflict is that both $\hat{y}_1$ and $\hat{y}_2$ are valid predictors of job 1 (or job 2) and they make contradictory predictions for the same sample of applicants. That is, $\hat{y}_1$ predicts the mean performance to be +.345 standard deviation units and $\hat{y}_2$ predicts it to be -.345 standard deviation units.

Clearly, both predictions cannot be true. The apparently paradoxical situation arises because the assignment strategy introduces additional conditions on the observed predicted performance scores. Specifically, it selects extreme cases based on comparison of the observed values of predictors. In this example, it compares $\hat{y}_1$ and $\hat{y}_2$ for each applicant.

Since the paradoxical effect is introduced by conditional assignment, a linear equation reflecting the conditions may be defined which accounts for the effects of the assignment of extreme scores (often denoted as regression effects) as follows:

$$y_{ij} = \beta'_j [\hat{Y}_i - E(\hat{Y})] + \epsilon_{ij} \tag{12}$$

where $\hat{Y}_i$ is the P-vector of predicted performance for the randomly chosen i-th applicant, and β_j is the P-vector of regression weights which minimize the expected square of the error, ϵ_{ij}. This least-squares solution is determined by solving for β_j in the normal equations,

$$\boldsymbol{\Sigma}_{\hat{y}} \beta_j = Dg\{ \boldsymbol{\Sigma}_{\hat{y}} \}^{1/2} V_j \tag{13}$$

where the variance-covariance matrix, $\boldsymbol{\Sigma}_{\hat{y}}$, is defined in equation 3, V_j is the j-th column vector of validities in equation 5, and the matrix, $Dg\{ \boldsymbol{\Sigma}_{\hat{y}} \}^{1/2}$, is the scaling matrix from equation 6.

Accordingly,

$$\dot{y}_{ij} = \beta'_j [\hat{Y}_i - E(\hat{Y})] \tag{14}$$

defines $\dot{y}_{ij}$, the expected value of the actual criterion performance (y_{ij}) under the conditions imposed by an assignment process.

The value $\dot{y}_{ij}$ is hereafter referred to as a re-estimate of actual performance (reMAP). The term, re-estimate, was chosen because the y_{ij} have already been defined as estimates of actual performance. The re-estimation is necessary because of the use of the conditional assignment strategy.

The mean of the re-estimates, $\dot{y}_{ij}$, for a sample of n_j applicants assigned to the j-th job would constitute a measure of classification efficiency with respect to that job, that is,

$$m_{\dot{y}_j} = \sum_{i=1}^{n_j} \dot{y}_{ij} / n_j \tag{15}$$

Furthermore, the weighted mean of the appropriate re-estimates across jobs estimates the overall classification efficiency of an assignment strategy in terms of actual criterion performance, that is,

$$MAP = \sum_{j=1}^{P} n_j \, m_{\dot{y}_j} \Big/ \sum_{j=1}^{P} n_j \qquad (16)$$

Sample Re-estimation of Actual Performance

Unfortunately, the re-estimate, $\dot{y}_{ij}$, is defined in terms of population parameters that are not ordinarily known. The potential for practical use depends on obtaining satisfactory estimates of the variance-covariance and validity matrices ($\boldsymbol{\Sigma}_{\varphi}$ and $\mathcal{V}$) defined in equations 3 and 5, respectively. Furthermore, estimation of the unknown values of $E(y_i)$ is required.

It is beyond the scope of this report to summarize all issues regarding the estimation of the elements of these two matrices. It must be recalled that they are parameters of the population from which the applicants have been drawn. Generally, the only statistics available are obtained from previous validation research and, unfortunately, these statistics are frequently based on samples where covariances have been restricted in range and validity estimates are subject to "shrinkage." Details of how the estimates were obtained for the Monte Carlo studies are described in the next section of this chapter.

For now, suppose that satisfactory estimates

$$S_{\varphi} = Est(\boldsymbol{\Sigma}_{\varphi}) \qquad (17)$$

and

$$V = Est(\mathcal{V}) \qquad (18)$$

are available. Then, estimates of the P x P matrix of regression coefficients, B, required for the re-estimates is

$$B = S_{\varphi}^{-1} \, Dg\{S_{\varphi}\}^{1/2} \, V \qquad (19)$$

where the j-th column of elements, B_j, in B constitute estimates of β_j in equation 12.

Thus, using sample data, the re-estimate of actual performance for the i-th applicant assigned to the j-th job is

$$Est(\dot{y}_{ij}) = B'_j \, [\, \hat{Y}_i - G(\hat{Y}) \,] \qquad (20)$$

where $\hat{Y}_i$ denotes the P-vector of predicted performance scores for the applicant and $G(\hat{Y})$ constitutes a vector of "guesses" of the values of $E(\hat{Y})$.

With respect to these expected "guessed" values for the $E(\hat{y}_i)$, one might estimate them from the applicant sample if the sample is large. Also, one might reasonably assume them to be zero if regression methods were used to develop the prediction composites.

A re-estimate of the classification efficiency for the j-th job can then be written as follows:

$$\text{Est}(m_{\hat{y}_j}^{n}) = \sum_{i=1}^{n_j} \text{Est}(\hat{y}_{ij}) \tag{21}$$

Additionally, a re-estimate of overall classification efficiency may be computed as

$$\text{Est}(\text{MAP}) = \sum_{j=1}^{P} n_j \, \text{Est}(m_{\hat{y}_j}) \Big/ \sum_{j=1}^{P} n_j \tag{22}$$

Theoretical Expected Job and Overall Mean Predicted Performance

The re-estimates of expected actual performance for applicants assigned to jobs using an assignment strategy depend on the estimation of the elements of the matrix of means of predicted performance (equation 8). To obtain these estimates, one may go through the process of developing the predictor variates for each job and collecting a sample of applicants on which to base the re-estimates.

However, it is of practical value to be able to forecast the results of assignment at a point in time before the applicant samples are actually obtained. This capability may be expected to provide useful information regarding the selection among assignment strategies or assist in the development of predictors.

Building on the work of Tippitt (1925) and Brogden (1951, 1959), this subsection develops the rationale of the statistical expectation for the matrix of means of predicted performance. The development continues with a proposed method of estimating the values based on statistics that are often available from samples used in predictor development.

For a given situation where the predictors, $\hat{Y}$, have been developed for a given set of P number of jobs, the three parameters which define the expectation of the matrix of means of predicted performance are as follows:

(1) $E(\hat{Y})$ = the vector of applicant population means (equation 4).
(2) $\Sigma_{\hat{y}}$ = the variance-covariance matrix (equation 3).
(3) Q = a P-vector of quotas for the jobs.

Not previously mentioned is the vector Q, which consists of proportions. The element of Q, q_i, is the proportion of the applicant sample that is to be assigned to the i-th job. Thus, it is a value that is positive and less than or equal to 1.00. Also, the sum of the elements of Q must be less than or equal to 1.00.

The fact that these are the three relevant parameters becomes evident by examining a case where the exact statistical expectation can be defined. Consider the simple case that was illustrated in Figure 8.1. Under the rule of Brogden's assignment strategy, the expected values of $\hat{y}_1$ and $\hat{y}_2$ for a randomly selected point are completely determined by the bivariate normal distribution with the specified variance-covariance.

The specific function for the expected value of the mean, m_{11}, of $\hat{y}_1$ for those assigned to job one would be

$$E(m_{11}) = E(\hat{y}_1) + \int_{-\infty}^{\infty} \int_{\hat{y}_2}^{\infty} \hat{y}_1 \, f(\hat{Y},E(\hat{y}_1),\boldsymbol{\Sigma}_{\hat{y}}) \, d\hat{y}_1 d\hat{y}_2 \qquad (23)$$

where

$$f(\hat{Y},E(\hat{y}_1),\boldsymbol{\Sigma}_{\hat{y}}) = (2\pi)^{-1} |\boldsymbol{\Sigma}_{\hat{Y}}|^{-1/2} \exp\{-1/2[(\hat{Y}-E(\hat{Y}))' \boldsymbol{\Sigma}_{\hat{Y}}^{-1} (\hat{Y}-E(\hat{Y}))]\}$$

which is the density function for the point defined by $\hat{y}_1$ and $\hat{y}_2$ in the bivariate normal distribution of $\hat{Y}$. The expected values of all four means, m_{ij}, may be defined by appropriate substitutions. Moreover, the form may be readily generalized to incorporate the Brogden assignment strategy for any number of jobs by adding an integral for each added job and augmenting the matrices with an added predictor for each job. Thus, for the Brogden assignment strategy, it is clear that the first two of the three parameters listed above completely determine the expectation of the means of predicted performance.

The third parameter of quotas, Q, does not affect the definition where the Brogden assignment strategy is applied because incumbency quotas are not invoked by the Brogden strategy.

The type of modification required for equation 23 to incorporate quotas depends on whether the quota for each job applies as a proportion of the population of applicants or as a proportion of a particular sample of applicants. If the quota for job 1 is a proportion of the population of applicants, the lower limit of the inside integral would become a value, z, such that

$$q_1 = \int_{-\infty}^{\infty} \int_{z}^{\infty} f(\hat{Y},E(\hat{y}_1),\boldsymbol{\Sigma}\hat{y}) \, d\hat{y}_1 d\hat{y}_2 \qquad (24)$$

In practice, z may be a random variable. Job incumbency quotas ordinarily would be applied to a particular sample of applicants because any particular sample would be assigned to jobs with a fixed number of vacancies at the time the assignments are made.

It is beyond the scope of this paper to exhaust the complexities of the variations that may arise in defining a function of the form of equation 23 for varied general applications. The purpose of introducing the definition here is limited to providing the basis for contending that the three parameters that determine the expectation of $\hat{Y}$ are as listed above. Examination of the equations supports the contention.

For purposes here, it is sufficient to state that a function exists which determines the expectation of any element, m_{ij}, in the matrix, $M_{\hat{y}}$, as follows:

$$E(m_{ij}) = F_s(i, j, P, Q, E(\hat{Y}), \Sigma_{\hat{y}}) \tag{25}$$

which is the expected value for the mean of the j-th predictor in the group assigned to the i-th job among P number of jobs and where the predictors, $\hat{Y}$, have the expectation and covariance of $E(\hat{Y})$ and $\Sigma_{\hat{y}}$, respectively.

The subscript of F_s in equation 25 denotes a function for a particular assignment strategy, that is, the s-th strategy. For instance, F_B may denote the Brogden strategy or F_{LP} may denote the strategy that employs a linear programming algorithm to assign the applicants.

Thus, the function F_s belongs to a family of functions dependent on (a) the specific assignment strategy that is applied, and (b) the multivariate distribution of $\hat{Y}$.

At this point, it is clear that the parameters that determine the expected result of applying an assignment strategy make it cumbersome, if not prohibitive, to attempt mathematical definition of the general range of forms which F_s may incorporate.

Also, the function F_s depends on parameters that are ordinarily unknown in practice. Accordingly, an estimator based on sample data is proposed as

$$Est(m_{ij}) = F_s(i, j, P, Q, G_{\hat{y}}, S_{\hat{y}}) \tag{26}$$

where $S\hat{y}$ is defined (equation 17) as the sample estimator of $\Sigma_{\hat{y}}$ and $G_{\hat{y}}$ is a vector of "guesses" of the values of $E(\hat{Y})$ in the population of applicants. As in equation 20, the "guesses" are "reasonable" guesses if empirical estimates are not available.

Numerical Evaluation of the Estimator of Mean Predicted Performance

The calculus of analytically evaluating a form such as equation 23 as a general case is formidable and impractical. Moreover, since the form has limited generality, the aforementioned task of rewriting the form to fit more than a nominal range of situations and assignment strategies is prohibitive.

A numerical solution for the function F_s, which appears in equations 25 and 26, is necessary. The approach proposed in what follows is the Monte Carlo method. Although researchers are accustomed to Monte Carlo methods as empirical investigative tools, it is a viable numerical analysis technique for the problem at hand. Implementing software can be prepared even for desktop, personal computers with the capability of rapid computation with real numbers.

The required Monte Carlo method is a process of applying the selected assignment strategy to simulated, random instances from the known or assumed population of applicants. The needed statistics to be used as the estimates of mean predicted performance, such as those defined in equation 7, are simply computed from the simulated samples in the same way as they would have been computed from a real sample.

The capability of obtaining random observations, $\hat{Y}$, from the simulated applicant population is the first requirement of obtaining the numerical evaluation by Monte Carlo method. Obtaining random observations of the vectors of predicted performance may be difficult. One option may be to sample a very large, finite population of predictors that has been derived from an existing database of applicants. Conceivably, one might also contrive a large finite database that could be assumed to have the distribution properties of the future applicant population. For general purposes, it is often reasonable to assume that the distribution of $\hat{Y}$ in the population of applicants is multivariate normal. This provides a convenient source of pseudo-random observations in the form of a multivariate normal, pseudo-random number generator.

Accordingly, for purposes here, the population will be assumed to be multivariate normal. Specifically, the distribution of $\hat{Y}$ is assumed to be multivariate normal with assumed expectation of either $E(\hat{Y})$, for equation 25, or $G_{\hat{Y}}$, for equation 26. Also, the multivariate distribution has the variance-covariance of either $\Sigma_{\hat{Y}}$, for equation 25, or $S_{\hat{Y}}$, for equation 26.

Next, the capability of providing an algorithm for implementing the chosen assignment strategy for samples of the simulated applicants is presumed. For practical application, this choice could be critical in that the precise method of assignment very directly affects the numerical values of the resulting estimates of mean predicted performance.

Finally, the procedure is simply that of repeating the following steps enough times to obtain the desired accuracy of the numerical results:

(1) Obtain a simulated sample size N from the assumed population of applicants.
(2) Apply the assignment algorithm to the sample, imposing the operational quota constraints.
(3) Compute the P x P matrix of sample means of predicted performance scores.
(4) Aggregate the resulting matrix into a final matrix of estimates.

The more times the process is repeated, the more accurate the resulting matrix of estimates of mean predicted performance will become.

The following section describes the implementation of the method characterized here, using linear programming assignment for each of 100 samples of approximately 500 applicants.

MONTE CARLO DEMONSTRATION OF ACCURACY OF eMPP AND reMAP

A demonstration of the accuracy of the sample estimators of mean predicted performance (eMPP) and mean actual performance (reMAP) was performed by Rosse, Whetzel, and Peterson (1993). The accuracy of both estimators based on sample data was considered satisfactory.

Since the conditions of the Career Force validation studies vary somewhat from those of the previous evaluation, further demonstration of the accuracy of eMPP and reMAP, under conditions specific to the Career Force validation effort, was performed.

Several Monte Carlo investigations were performed for this demonstration. They consist of simulating the process of:

(1) Gathering predictor and criterion data.
(2) Developing weights for predictor composites.
(3) Computing the estimates of mean predicted performance (eMPP) and mean actual performance (reMAP) for groups that might be assigned using an assignment algorithm.
(4) Sampling again for applicants from the population and assigning them to jobs using the chosen assignment algorithm.
(5) Computing the mean predicted performance (MPP) and mean actual performance (MAP) for the assigned groups for comparison to the estimates.

Defining the Simulated Population

An artificial population that approximates the population from which the Army draws recruits was needed. The objective was to simulate the sampling of applicants (or recruits) from the population consisting of predictor and criterion scores. To accomplish this, the covariances of this population were defined so that multivariate observations drawn from the population could be generated using a pseudo-random number generator.

To conduct the Monte Carlo investigations, the actual sample statistics obtained for the Longitudinal Validation sample (LVI) data for the Career Force project were used to simulate a population of applicants. This simulated population, while not exactly the same as the population from which the Army actually draws applicants, was assumed to have covariance properties very similar to the actual population from which the LV1

samples arose and from which Army recruits will be drawn in the future. The specific predictor variables (of which there were 33) and the criterion variable (Core Technical Proficiency for each of the nine MOS included in the investigation) are described in the final section of this chapter.

The simulation of all nine criterion variables, one for each MOS, for each observation provides an opportunity to observe a separate criterion for each MOS with the appropriate covariances with the predictor variables. Obviously, this kind of observation cannot be obtained in the "real world" because it would require the simultaneous assignment of each applicant to all nine MOS. However, as will be seen, the presence of an "actual" criterion score on all nine criteria for each simulated applicant provides the capability of comparing the proposed estimators to the simulated "actual" performance.

To construct a plausible covariance matrix with the required 42 variables (the 33 predictors common across MOS and a set of nine criteria, one for each of the nine MOS), a series of steps were performed. It was decided that the covariances should approximate those found in the 1980 Youth Population (Mitchell & Hanser, 1984), which contains only the nine ASVAB variables. The remainder of the variables are experimental variables from the Career Force project for which available data were gathered during the course of the research. Since these data were based on selected samples, they are assumed to have been directly or indirectly restricted in range on all variables.

The first step consisted of correcting the nine covariance matrices and the corresponding mean vectors of predictors for range restriction. The nine corrected matrices were then pooled to form a single matrix of covariances for predictors. The pooling was a weighted sum of the covariances with the between-groups variances (based on the corrected means) added into the final matrix. This provided covariances that were assumed to approximate the covariances that would be found in the 1980 Youth Population if the experimental variables had been included in the 1980 data. The resulting matrix of covariances became the assumed population of predictor variables for the Monte Carlo investigations.

To include the criterion variables in the correction, each of the sample matrices of predictor covariances for the nine MOS was corrected for restriction of range using the assumed population described above. Then, the criterion variable for each group was appended to the matrix so that the covariances of predictor variables with the criterion were corrected for restriction of range. (All range restriction corrections were made using the Lawley method cited in Lord and Novick [1968].)

Each row of covariances of predictors with the criterion was appended to the population matrix in order to provide the needed covariances of predictor variables with all nine of the criterion variables. The covariances among the nine criterion variables were not included because they were not involved in any of the computations.

The final matrix contained covariances between all 42 variables considered plausible for the 1980 Youth Population. This matrix was used in the Monte Carlo investigations to generate 42 scores (33 predictors and 9 criteria) for each simulated applicant (or recruit), using a pseudo-random number generator capable of sampling the multivariate normal population.

Simulated Validation Studies

The investigations were designed to simulate the approximate conditions of the actual Career Force validation analyses. However, the validation studies were repeatable because they used simulated data. In all, the simulated validation studies were repeated 30 times for each particular investigation.

In the "real" Career Force validation studies, predictor and criterion data were obtained from incumbents in each of nine MOS. The covariances of the actual incumbent data were corrected for restriction of range. The correction of each criterion score and 24 of the 33 predictor variables was based on the nine ASVAB variables found in the 1980 Youth Population.

The simulated validation studies approximated the same conditions as the "real" study. To conduct a complete validation study, nine samples of predictor and criterion variables, one for each of the nine MOS, had to be obtained from the simulated population.

Each sample consisted of the appropriate predictor variables and the appropriate criterion drawn from the simulated population. The samples were restricted by using cutoff scores (on the appropriate operational ASVAB composite) supplied by the Army. To accomplish this, the ASVAB composite score for each MOS was computed for each simulated observation; if an observation fell below the cutoff for the MOS corresponding to the simulated sample, it was discarded. Sampling continued until the number of observations for each MOS was equal to the actual developmental sample sizes in the LV1 sample data.

The covariance matrices for each of the nine simulated samples were then corrected for restriction of range using only the ASVAB variables from the 1980 Youth Population. This corresponds to the practice used in the actual Career Force validation studies, where 1980 Youth Population information for the experimental variables was not available.

These nine covariances constituted developmental samples comparable to those realized in the LV1 data.

Using the covariance matrix for each of the nine MOS, least-squares regression weights for forming predicted performance composites were obtained. The validity of each composite was estimated by adjusting the foldback multiple correlation for shrinkage, using the Rozeboom formula 8 (1978). Additionally, the validity of each composite for predicting the criterion in each of the eight remaining MOS was computed.

This was accomplished by applying the predictor weights in each of the simulated samples for all MOS except the one in which the weights were developed. Thus, these validity estimates consisted of the Pearsonian correlation between the predictor composite and the criterion.

Finally, an estimate of the covariances of the predictor variables was obtained by pooling the nine simulated MOS covariance matrices.

At this point, the information needed to compute eMPP and reMAP was available, that is:

(1) $S_{\mathcal{P}} = Est(\Sigma_{\mathcal{P}})$ = the matrix of estimates of covariances of predictor composites (see equation 17).
(2) $V = Est(\mathcal{V})$ = the estimate of the validities of each predictor composite for predicting each criterion (see equation 18).

Numerical Evaluations for Estimated Mean Predicted Performance (eMPP)

The numerical evaluations for estimating mean predicted performance followed the approach and procedures described earlier in this chapter.

For this demonstration, the incumbency quotas were assumed to be proportional to the developmental sample sizes. Thus, there were 3,083 observations distributed across the nine MOS in the LV1 data. Since 551 of them were in MOS 13B, the quota assumed for 13B was 551/3,083 = .179. These proportions were calculated for each of the nine MOS to represent the proportions of future applicants that would be assigned to each of the nine MOS.

The numerical analysis consisted of application of Monte Carlo numerical evaluation to equation 25 in order to obtain a 9 by 9 matrix of estimated means of predicted performance score--that is, one mean for each combination of predictor composite and criterion.

The assignment method used here was chosen because of its simplicity and speed of computation. It was considered a satisfactory approximation to the results that would be obtained using a linear programming algorithm which would maximize overall MPP.

The simplified algorithm assumes that the future applicant samples would contain some fairly substantial number of applicants. The assignment would consist of finding the highest predicted performance among a group of applicants, assigning the corresponding applicant to the MOS for which the highest of the nine predicted performance scores was realized, and then removing the applicant from further consideration. The process is repeated for the applicants still remaining for consideration and continues until the incumbency quota for any one MOS is filled. At that point, the MOS which is filled to quota is removed from further consideration. The process is continued until all applicants are assigned.

Using the Monte Carlo method, the process was simulated 50 times by randomly simulating 496 points in the multivariate normal distribution which has a covariance matrix of S_9. The result was 496 times 50 = 24,800 simulated points which were assigned according to the assignment algorithm. The number of points in each of the nine groups of points was proportional to the assigned quota. For instance, the number of points in the second group representing MOS 11B was .179(24,800) = 4,439.

Each point had nine values corresponding to predicted performance on each of the nine predictor composites. A 9 by 9 matrix of means was then computed by dividing the sum of the scores for each row of the matrix by the number of points in the group of points. Each element of this 9 by 9 matrix constitutes an evaluation of the function defined in equation 26 for this particular method of assignment.

The diagonal elements of this matrix constitute the estimate of the predicted performance, eMPP, for each of the nine MOS. The weighted mean of the nine diagonal elements constitutes an overall estimate of mean predicted performance.

More importantly, the information for computing estimates of actual criterion performance (reMAP) is available at this point. Specifically, the reMAP is computed using equation 20 for each of the nine MOS. An overall estimate of reMAP is also obtained by a weighted mean of the estimates for each of the nine MOS.

Simulated Actual Assignment of 1,000 Applicants

All statistics described above are computed on developmental sample data of the type that was actually realized in the LV1 data. To demonstrate that the statistics eMPP and reMAP are accurate in forecasting the results that would be realized if the Army actually used the predictor composites as developed and assigned recruits based on the same assignment algorithm, it is necessary to continue the simulation process to include the gathering of simulated, "future" applicant data.

To accomplish this, after each simulated validation study was completed, a sample of 1,000 simulated applicants was drawn from the same population of applicants from which the developmental samples were drawn. However, unlike the "real world," criterion scores for all nine MOS were known for every applicant. Thus, when the mean predicted performance (MPP) is computed for each simulated group of assigned applicants, it is also possible to compute the mean actual performance for each group.

The means of predicted performance can be compared to the means of the estimates, eMPP, and the means of actual performance can be compared to the means of the estimates, reMAP. If the two estimators, eMPP and reMAP, are accurate, they should be the same as the MPP and MAP obtained from the simulated group of "real" applicants (allowing for sampling error).

Results of Three Monte Carlo Investigations

Three Monte Carlo investigations were conducted which included all of the processes described above (i.e., simulating validation studies, obtaining eMPP and reMAP, and assigning 1000 "future" applicants). All three studies used the same simulated population as the source of data for both developmental sample and applicant data. For each investigation, the complete process was repeated 30 times; the results reported below were based on the averages across the 30 repetitions.

The three investigations differed in the predictor variables used in predictor equations as follows:

(1) Nine ASVAB subtests, for the first investigation.
(2) Nine ASVAB subtests, the spatial composite, and eight computer-administered composites, for the second investigation.
(3) Nine ASVAB subtests, eight ABLE composites, and seven AVOICE composites, for the third investigation.

Tables 8.1 through 8.6 present summaries of the results of each of the three Monte Carlo investigations. Tables 8.1, 8.3, and 8.5 contain summary information regarding the 30 simulated validation studies. This information includes:

- The sample sizes for each of the groups for each repetition.
- The foldback correlation (multiple correlation in the sample).
- The adjusted validity estimate based on Rozeboom's formula 8 (1978).
- The "true" validity computed on the population.
- The selection ratio representing the proportion of simulated applicants accepted according to the AFQT cutoff scores.

Tables 8.2, 8.4, and 8.6 contain the average of the 30 estimates of mean predicted performance (eMPP) and estimated mean actual performance (reMAP) computed from developmental sample data.

The results of these three Monte Carlo investigations suggest that the estimators, eMPP and reMAP, for assessing classification efficiency are quite accurate. The critical comparisons are eMPP with MPP and reMAP with MAP. For the eMPP-MPP comparison, the discrepancies are .002, .002, and .003 across the three investigations. For the reMAP-MAP comparisons, the discrepancies are .006, .02, and .013.

Importantly, note that the individual MOS comparisons are similarly close, especially in the pattern across MOS, though the absolute values of the discrepancies are, of course, a bit larger. Finally, note that mean MPP and mean eMPP consistently overestimate mean MAP and mean reMAP by sizeable amounts. MPP overestimates MAP by .026, .116, and .134 in these three investigations and eMPP similarly overestimates reMAP.

Table 8.1
Means of Developmental Sample Statistics for Monte Carlo Investigation Using ASVAB Only

MOS Group	Sample Sizes[a]	Foldback Correlation	Adjusted Validity[b]	True Validity	Selection Ratio
11B	235	.741	.716	.739	.679
13B	551	.447	.416	.417	.750
19K	445	.487	.454	.452	.672
31C	172	.661	.611	.610	.498
63B	406	.623	.600	.596	.665
71L	251	.833	.819	.813	.601
88M	221	.554	.496	.513	.679
91A	535	.679	.666	.666	.600
95B	270	.751	.730	.740	.596

Note. Across 30 repetitions.

[a] Sample sizes are the same for each of the 30 repetitions.
[b] Rozeboom formula 8 (1978).

Table 8.2
Means of Estimates of Mean Predicted and Mean Actual Performance (eMPP and reMAP) Compared to Simulated Results of Assigning 1,000 "Real" Applicants: First Investigation

MOS Group	Quota	Developmental Samples		1,000 Applicants	
		eMPP	reMAP	MPP	MAP
11B	.077	.924	.829	.929	.847
13B	.179	-.172	-.226	-.167	-.217
19K	.145	-.116	-.180	-.130	-.200
31C	.056	.914	.747	.956	.768
63B	.131	.092	.034	.084	-.009
71L	.081	1.414	1.373	1.405	1.357
88M	.071	.459	.289	.463	.273
91A	.173	-.128	-.147	-.119	-.147
95B	.087	.847	.766	.833	.772
Mean		.284	.216	.286	.210

Note. Across 30 repetitions.

Table 8.3
Means of Developmental Sample Statistics for Monte Carlo Investigation Using ASVAB, Spatial, and Computer Tests

MOS Group	Sample Sizes	Foldback Correlation	Adjusted Validity[a]	True Validity	Selection Ratio
11B	235	.802	.765	.771	.670
13B	551	.482	.424	.427	.753
19K	445	.517	.452	.453	.674
31C	172	.724	.641	.639	.498
63B	406	.664	.624	.623	.666
71L	251	.837	.809	.816	.604
88M	221	.608	.506	.527	.669
91A	535	.701	.675	.681	.594
95B	270	.794	.760	.773	.599

Note. Across 30 repetitions.

[a] Rozeboom formula 8 (1978).

Table 8.4
Means of Estimates of Mean Predicted and Mean Actual Performance (eMPP and reMAP) Compared to Simulated Results of Assigning 1,000 "Real" Applicants: Second Investigation

MOS Group	Quota	Developmental Samples		1,000 Applicants	
		eMPP	reMAP	MPP	MAP
11B	.077	1.177	1.037	1.197	1.106
13B	.179	-.134	-.251	-.140	-.220
19K	.145	-.016	-.172	-.011	-.177
31C	.056	1.179	.940	1.168	.932
63B	.131	.136	-.001	.137	.025
71L	.081	1.315	1.227	1.308	1.232
88M	.071	.573	.272	.562	.291
91A	.173	-.060	-.134	-.068	-.136
95B	.087	.946	.811	.927	.862
Mean		.366	.228	.364	.248

Note. Across 30 repetitions.

Table 8.5
Means of Developmental Sample Statistics for Monte Carlo Investigation Using ASVAB, ABLE, and AVOICE Tests

MOS Group	Sample Sizes	Foldback Correlation	Adjusted Validity[a]	True Validity	Selection Ratio
11B	235	.794	.738	.737	.675
13B	551	.510	.439	.441	.755
19K	445	.548	.468	.477	.673
31C	172	.747	.642	.653	.506
63B	406	.663	.608	.608	.669
71L	251	.857	.824	.827	.606
88M	221	.656	.537	.559	.679
91A	535	.699	.664	.675	.598
95B	270	.782	.732	.744	.602

Note. Across 30 repetitions.

[a] Rozeboom formula 8 (1978).

Table 8.6
Means of Estimates of Mean Predicted and Mean Actual Performance (eMPP and reMAP) Compared to Simulated Results of Assigning 1,000 "Real" Applicants: Third Investigation

MOS Group	Quota	Developmental Samples		1.000 Applicants	
		eMPP	reMAP	MPP	MAP
11B	.077	1.076	.870	1.061	.862
13B	.179	-.039	-.162	-.036	-.155
19K	.145	.130	-.025	.134	-.018
31C	.056	1.251	.941	1.323	1.032
63B	.131	.256	.100	.249	.091
71L	.081	1.406	1.311	1.430	1.327
88M	.071	.871	.557	.848	.602
91A	.173	-.114	-.182	-.098	-.163
95B	.087	.922	.775	.880	.764
Mean		.434	.281	.437	.294

Note. Across 30 repetitions.

Advantages and Disadvantages of Practical Application of eMPP and reMAP Estimators

The estimators for classification efficiency in terms of predicted performance are offered as preferred estimates to the Brogden Allocation Average. As shown just above, the eMPP estimators offer the advantage of estimating the mean predicted performance on the individual job level as well as the aggregated mean across all jobs among which assignments were made.

Moreover, the assumptions required for the application of eMPP are more consistent with actual circumstances than are the simplified assumptions of the Brogden Allocation Average. For instance, the eMPP estimators are not restricted to any simple assignment algorithm such as that outlined by Brogden but can be applied for any assignment algorithm that can be computerized. Additionally, the patterns of variances and correlations of predictor composites need not conform to the assumptions of equality made by Brogden.

An obvious disadvantage is that the eMPP estimators cannot be calculated from a simple formula. The numerical evaluation of the multivariate distributions of predictor composites requires a computerized process, but this is not a serious disadvantage now, as it was when Brogden developed the Allocation Average.

Both the Allocation Average and eMPP estimators are based on predicted performance. This is a major disadvantage because, as shown above, mean predicted performance is not a particularly good estimator of mean actual performance under conditions where assignment algorithms are likely to be applied. Thus, any estimate of the mean predicted performance based on validation statistics from finite samples is subject to being spuriously inflated by the conditions imposed by the assignment algorithm.

Accordingly, the reMAP estimators based on mean actual performance have been offered as the most useful estimator. As with eMPP estimators, the reMAP estimators may be applied at the level of individual jobs as well as for the aggregate of jobs. The metric of the estimators is the metric of actual performance and not the metric of predicted performance. The estimator is calculated from sample-based validation data, as is the Brogden Allocation Average.

APPLICATIONS OF CLASSIFICATION EFFICIENCY ESTIMATES IN LVI ANALYSES

In the previous two sections of this chapter, the statistical concepts for estimating mean predicted performance (eMPP) and mean actual performance (reMAP) were developed, and the accuracy of the estimates was demonstrated. In this section, the estimators are applied to the LVI sample results in order to investigate the implications of certain issues important to decisions about classification strategies for the Army. These decisions all involve applying several types of possible prediction systems for future operational use of classification strategies.

The specific issues addressed here are the effects of:

(1) Least-squares weighting compared to synthetic weighting.

(2) Various combinations of the experimental tests with the ASVAB.

(3) Classification on individual MOS (in terms of reMAP).

(4) The criterion (Overall Performance or Core Technical Proficiency) on classification efficiency.

The synthetic weighting systems resulted from the Army's Synthetic Validity Project (Wise et al., 1991). In this project, job components that were common across a range of jobs were identified and job incumbents rated the extent to which each component or task was an important and frequent part of their jobs. A group of psychologists judged the validity of predictor constructs included in Project A/Career Force for these job components; prediction equations can then be developed for any Army job by combining these judgments. The results of that research showed that the judgment-based approach produced validity coefficients very close to those from least-squares, sample-based methods.

The obvious advantage of the synthetic validation procedures is that they do not require the gathering of empirical criterion data but, rather, the application of relatively inexpensive experimental procedures in order to arrive at the weights. Accordingly, the generalization of the LVI results to additional MOS would be relatively easily accomplished if confidence in the effectiveness of the procedures exists.

As just noted, Wise et al. (1991) showed that the synthetic method of deriving weights for predictor composites yielded validity coefficients only slightly lower than least-squares weights for LVI data. Thus, the important question, described in this chapter, is whether they are satisfactory for estimating classification efficiency in terms of predicted and actual performance.

The second issue involves combining experimental tests with the ASVAB for operational use in selecting and assigning Army recruits. Decisions will be required regarding which of the experimental tests to use and how to combine them for satisfactory classification of the recruits among MOS. This section presents results about the classification efficiency of various predictor combinations.

The third issue involves reviewing the classification efficiency results for individual MOS and determining the extent to which gains, in terms of standard deviation units of mean predicted performance that are above zero (the classification efficiency of random assignment), in some jobs are offset by mean standard scores on performance that are below zero in other jobs. It is important for the Army to know the extent to which critical Army jobs are predicted to have low levels of job performance as a result of classification. This section presents classification efficiency results for individual jobs, as well as means across jobs.

The fourth issue involves the criterion used for the prediction of performance. Several criteria were identified during the analyses of first-tour job performance (Campbell, McHenry, & Wise, 1990). These included Core Technical Proficiency, General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing. In addition, a weighted combination of these five criteria, Overall Performance, was developed (Sadacca, Campbell, White, & DeFazio, 1988). This section compares the classification efficiency resulting from performance predictions of Core Technical Proficiency and Overall Performance.

Weighting Systems

The predictor scores used in these analyses were taken from the ASVAB operationally administered when the soldiers were inducted, and from the paper-and-pencil and computerized tests administered in the Project A LV Experimental Battery. As indicated above, least-squares weights and synthetic weights were computed on the synthetic validity variable set, described below.

Least-Squares Weights. In all, 108 sets of least-squares (LS) weights were computed for the eMPP and reMAP estimations: nine MOS, six sets of predictor variables (shown below), and two criteria. Each of the sets of LS weights (shown in Appendix F) was computed using the common matrix of correlation among predictor variables and the appropriate vectors of correlations with the criterion variables. Each of the 108 sets of LS weights was used to define a predictor composite corresponding to the appropriate MOS and criterion.

The validity of each predictor composite for predicting its own criterion was obtained by computing the multiple correlation and adjusting it for "shrinkage" using the Rozeboom formula 8 (1978). The validities of the predictor composite for predicting the criteria in other MOS were directly computed from the correlations of the predictors with the appropriate correlations with the criterion variables. These validity estimates did not require adjustment for "shrinkage." The standard deviations of the LS composites were determined by the multiple correlations. Specifically, the standard deviation of a composite of standardized regression weights (in the sample) is equal to the multiple correlation.

Synthetic Validity Weights. There also were 108 sets of weights based on the Synthetic Validation methods. The validities of these weights for predicting all criteria were directly estimated using the correlations of predictor test variables and the correlations with criteria. None of these validities required adjustment for "shrinkage" because they were not estimated from the existing sample. Each set of weights based on the synthetic validation methodology was adjusted so that the standard deviation of the corresponding composite was equal to the correlation of the composite with the appropriate criterion. This was done so that the definition of the metric of the synthetically derived weights would be comparable to the metric of the least-squares weights.

Predictor Variables

Shown below are the components of each of the five variable sets included in the investigations. Note that, for this investigation, the ASVAB component scores are slightly different that those traditionally seen. Four ASVAB factor scores were used, but computer-administered measures were used for perceptual speed and accuracy rather than the Numerical Operations and Coding Speed subtests. This was done because the synthetic validity methods used this approach and we wished to make direct comparisons of weighting systems on the same set of variables. The components were:

Attribute Set	Components
ASVAB	ASVAB: Verbal composite ASVAB: Quantitative Composite COMPUTER: Speed/Accuracy ASVAB: MC, EI, AS subtests
Spatial	SPATIAL: Reasoning SPATIAL: Assembling Objects, Map, Maze, Orientation, Object Rotation
Computer	COMPUTER: Basic Speed, Basic Accuracy COMPUTER: Perceptual Speed, Perceptual Accuracy COMPUTER: Short-Term Memory COMPUTER: Target Shoot: Mean Log Distance COMPUTER: One-Hand Tracking: Mean Log Distance, Two-Hand Tracking: Mean Log Distance COMPUTER: Cannon Shoot: Mean Time Discrepancy COMPUTER: Mean median movement time across 5 tests
ABLE	ABLE: Physical Condition scale ABLE: Work Orientation, Control scales ABLE: Cooperation, Stability scales ABLE: Energy scale ABLE: Dependability composite ABLE: Dominance, Self-Esteem scales
AVOICE	AVOICE: Structural/Machines composite AVOICE: Rugged/Outdoors composite AVOICE: Protective Services composite AVOICE: Computers, Electronics, Electronic Communications, Drafting, Audiographics scales AVOICE: Science scales AVOICE: Leadership/Guidance scale AVOICE: Aesthetics scale AVOICE: Clerical/Admin., Warehousing/Shipping, Food Service Prof., Food Service Employee scales

Subsets of the variables from these covariance matrices were used for all analyses:

ASVAB only
ASVAB + spatial
ASVAB + computer-administered psychomotor
ASVAB + ABLE
ASVAB + AVOICE
ASVAB + all experimental predictors

Criteria

Core Technical Proficiency, defined by Campbell, McHenry, and Wise (1990) and described by Oppler, Childs, and Peterson (1994), consisted of hands-on and job knowledge measures of technical skill. Components were unit weighted; that is, they were combined by standardizing them within MOS and then adding them together.

Overall Performance was a weighted combination of the five criterion components -- Core Technical Proficiency, as mentioned above, General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing, described by Campbell, McHenry, and Wise (1990). The weighting of the five criterion components, described by Sadacca, Campbell, DiFazio, Schultz, & White (1990), was conducted using officers' ratings of scenarios of performance (i.e., the conjoint method). For each MOS, the ratings of all five components summed to 100. The weights, from Sadacca, Campbell, White, & DiFazio (1988), are shown in Table 8.7.

Table 8.7
Weights Used for Calculating Overall Performance Across Five Criteria[a]

MOS	Core Technical Proficiency	General Soldiering Proficiency	Effort and Leadership	Personal Discipline	Military Bearing
11B Infantryman	22.9	18.5	29.1	17.2	12.3
13B Cannon Crewmember	22.7	19.2	27.7	18.3	12.1
19K Armor Crewman	29.4	21.1	20.5	17.9	11.0
31C Single Channel Radio Operator	29.0	20.3	22.0	17.3	11.4
63B Light Wheel Vehicle Mechanic	27.5	18.1	23.5	21.1	9.9
71L Administrative Specialist	24.1	19.9	22.7	21.0	12.3
88M Motor Transport Operator	26.1	22.8	21.8	15.4	14.0
91A Medical Specialist	26.9	16.6	23.1	22.5	11.0
95B Military Police	20.0	27.8	20.5	19.1	12.6

[a] From Sadacca, Campbell, White, and DiFazio (1988).

Sample

The sample information required for computing the estimators of mean predicted performance (eMPP) and mean actual performance (reMAP) consists of:

(1) Estimated covariances of predictor tests in the population from which future recruits will be drawn.

(2) Estimated validities of each of the predictor tests for predicting the criteria for all MOS among which assignments are to be made.

There were 32,075 soldiers in the Career Force data for whom all experimental test scores were available. These data also include the nine ASVAB subtest scores.

These data are presumed to have been restricted in range through selection, and thus are not representative of the population of recruits.

To obtain the estimated covariances of predictor tests in the population of recruits, the covariances based on the sample (N=32,075) were corrected for range restriction (Lord & Novick, 1968) based on the nine ASVAB subtest scores found in the 1980 Youth Population. These covariances were assumed to be representative of the population of recruits. And, at this point, the covariances involving the nine ASVAB subtests were discarded because these subtest scores were not further used in these analyses.

To obtain validity estimates for the two criterion variables that are the subject of these analyses, the LVI data for which criterion information exists were used. Thus, for each of the nine MOS, there exists a matrix of covariances between predictor tests and between each test and each of the criteria. The MOS included are:

11B	Infantryman (N=265)
13B	Cannon Crewmember (N=597)
19K	Armor Crewman (N=474)
31C	Single Channel Radio Operator (N=182)
63B	Light-Wheel Vehicle Mechanic (N=432)
71L	Administrative Specialist (N=265)
88M	Motor Transport Operator (N=241)
91A	Medical Specialist (N=564)
95B	Military Police (N=280)

The covariances in each of the nine matrices were also presumed to have been restricted in range by selection. At this point, the nine ASVAB subtest scores were not needed for the correction because the full matrix of corrected covariances based on the larger sample (N=32,075) could be (and was) used as a basis for correcting the covariances of tests with criterion variables.

The result of these steps was a matrix of covariances of predictor variables based on 32,075 observations presumed to be representative of the 1980 Youth Population. And, for each MOS, there was a set of covariances between the predictor tests and the criteria.

Then, all of the covariances were transformed to correlations so that the covariances were expressed in terms of standardized z-scores (mean = 0 and SD = 1).

Quota Conditions

In the following analyses, two quota conditions were compared: (a) Job assignments were made for all individuals, or (b) job assignments were made for the top 95% (5% were eliminated on the basis of the AFQT). Under both conditions MOS assignments were made in proportion to actual accessions for fiscal year 1993.

For the first condition, 50 samples were generated with 501 applicants in each sample. There were no rejections, quotas for each MOS were proportional to FY 93 accessions, and assignments were made by the linear programming strategy. In FY 93, 24,258 individuals entered Career Force MOS after passing the AFQT screen, in which there was a 5 percent failure rate. Table 8.8 shows that of the total that passed the AFQT screen, 7,320 (30.2%) were assigned to MOS 11B, 1,826 (7.7%) were assigned to MOS 13B, and so on.

Table 8.8
Proportion and Number of Soldiers Selected Into Nine Career Force MOS in Fiscal Year 1993

MOS	Proportion Assigned	Number Assigned
11B Infantryman	30.2	7,320
13B Cannon Crewmember	7.7	1,826
19K Armor Crewman	7.7	1,863
31C Single Channel Radio Operator	1.6	382
63B Light Wheel Vehicle Mechanic	12.5	3,046
71L Administrative Specialist	4.6	1,116
88M Motor Transport Operator	8.1	1,963
91A Medical Specialist	16.3	3,968
95B Military Police	11.4	2,774
Total	100.0	24,258

To simulate these quotas in the second condition, 50 sets of 527 applicants were generated. The lowest 5 percent were rejected, thus simulating the Army's initial selection prescreening ratio. The remaining 501 applicants were assigned to MOS according to proportions entering those MOS in FY 93. As shown in Table 8.9, of the 501 simulated applicants remaining after the 5 percent rejection, 151 were assigned to the 11B MOS (30.1%), 38 were assigned to the 13B MOS (7.6%), and so on. The proportions shown are slightly different from those used in FY 93 due to rounding.

Table 8.9
Proportion and Number of "Applicants" Assigned to Career Force MOS in Simulations

MOS	Proportion Assigned	Number Assigned
11B Infantryman	30.1	151
13B Cannon Crewmmember	7.6	38
19K Armor Crewman	7.6	38
31C Single Channel Radio Operator	1.6	8
63B Light Wheel Vehicle Mechanic	12.6	63
71L Administrative Specialist	4.6	23
88M Motor Transport Operator	8.2	41
91A Medical Specialist	16.4	82
95B Military Police	11.4	57
Total	100.0	501

Results

Estimates of Mean Actual Performance, reMAP, at the Level of Individual Jobs

As can be seen in Tables 8.10 through 8.21, the estimated classification efficiency gains indicated by the average across all nine MOS are not distributed equally for the individual MOS. Table 8.22 presents the "worst" and "best" cases for each of the MOS to be found among the six sets of predictors. For each MOS, by weighting method and criterion, the minimum and maximum reMAP estimates found among the predictor sets are given.

Table 8.22 is included to emphasize the lack of uniformity of gains in classification efficiency for individual jobs. For instance, the entries for MOS 11B indicate that almost any of the predictor sets and either weighting method would provide an increment in mean actual performance. It suggests that the "worst" case for 11B would be a gain of .16 standard deviation unit (over random assignment) using synthetic weights with the predictor set, ASVAB + computer-administered psychomotor, for predicting Core Technical Proficiency. That choice of predictors would also result in the highest gain for the MOS 95B. Unfortunately, as shown in Table 8.14, it would also result in decrements in performance for four of the nine MOS, substantially so for 13B and 19K.

It can also be seen that no predictor system results in much performance gain for some MOS. For instance, the "best" case for the MOS 13B for predicting Core Technical Proficiency is -.42 standard deviation unit.

Table 8.10
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB Only and Criterion = Core Technical Proficiency

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.651	.651	.433	.499	.287	.680	.680	.470	.524
13B	.076	-.579	-.579	-.538	-.680	.072	-.431	-.431	-.350	-.506
19K	.076	-.612	-.612	-.517	-.500	.072	-.468	-.468	-.405	-.383
31C	.016	-.027	-.133	-.087	-.206	.015	-.006	-.112	-.043	-.169
63B	.126	-.055	-.055	-.048	.048	.120	-.016	-.016	-.001	.095
71L	.046	.580	.532	.460	.389	.044	.630	.579	.484	.415
88M	.082	-.338	-.338	-.443	-.499	.078	-.278	-.278	-.371	-.431
91A	.164	-.063	-.063	-.083	-.216	.156	.000	.000	-.036	-.170
95B	.114	.745	.705	1.095	1.116	.108	.789	.748	1.147	1.169
All	1.000	.172	.163	.142	.132	.951	.231	.222	.200	.189

Table 8.11
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB Only and Criterion = Overall Performance

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.682	.666	.679	.683	.287	.706	.690	.701	.705
13B	.076	-.048	-.048	-.064	.040	.072	-.007	-.007	-.048	.056
19K	.076	.119	.084	-.065	-.055	.072	.140	.101	-.046	-.035
31C	.016	-.127	-.127	-.034	.049	.015	-.099	-.099	-.021	.062
63B	.126	.094	.094	.040	.101	.120	.111	.111	.056	.119
71L	.046	.134	-.026	.018	-.154	.044	.162	-.002	.031	-.147
88M	.082	-.162	-.162	-.060	-.089	.078	-.117	-.117	-.047	-.080
91A	.164	.065	.065	-.063	-.154	.156	.084	.085	-.047	-.136
95B	.114	.084	.084	.084	.083	.108	.112	.112	.104	.098
All	1.000	.244	.219	.194	.188	.951	.260	.244	.213	.206

Table 8.12
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Spatial and Criterion = Core Technical Proficiency

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.591	.591	.421	.412	.287	.618	.619	.463	.455
13B	.076	-.552	-.607	-.478	-.641	.072	-.383	-.446	-.306	-.478
19K	.076	-.613	-.613	-.505	-.497	.072	-.449	-.449	-.389	-.381
31C	.016	-.058	-.303	-.389	-.258	.015	-.021	-.265	-.299	-.178
63B	.126	-.017	-.080	-.002	.075	.120	.036	-.030	.026	.093
71L	.046	.550	.466	.462	.431	.044	.594	.509	.506	.479
88M	.082	-.182	-.182	-.369	-.413	.078	-.090	-.090	-.310	-.360
91A	.164	.010	.010	-.040	-.358	.156	.065	.065	.014	-.298
95B	.114	.840	.773	1.106	1.123	.108	.852	.783	1.136	1.151
All	1.000	.194	.167	.156	.099	.951	.255	.226	.215	.155

Table 8.13
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Spatial and Criterion = Overall Performance

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.685	.656	.674	.666	.287	.704	.673	.686	.683
13B	.076	-.038	-.194	-.043	.089	.072	.007	-.190	-.032	.101
19K	.076	.119	.020	-.066	-.025	.072	.144	.038	-.046	-.003
31C	.016	.081	.081	-.023	-.096	.015	.111	.112	-.013	-.088
63B	.126	.073	.073	.049	.142	.120	.091	.091	.058	.153
71L	.046	.154	-.233	.007	-.059	.044	.181	-.181	.015	-.053
88M	.082	-.115	-.115	-.045	-.020	.078	-.068	-.068	-.035	-.010
91A	.164	.097	.097	-.050	-.211	.156	.118	.118	-.039	-.196
95B	.114	.169	-.021	.072	.036	.108	.187	-.005	.081	.045
All	1.000	.256	.188	.197	.187	.951	.280	.209	.208	.201

Table 8.14
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Computer-Administered Psychomotor and Criterion = Core Technical Proficiency

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.524	.418	.259	.160	.287	.545	.443	.287	.188
13B	.076	-.495	-.618	-.341	-.476	.072	-.296	-.448	-.235	-.377
19K	.076	-.244	-.410	-.568	-.713	.072	-.116	-.292	-.409	-.548
31C	.016	.279	-.133	-.426	-.165	.015	.320	-.085	-.325	-.038
63B	.126	.054	-.120	.154	.197	.120	.115	-.056	.196	.236
71L	.046	.540	.361	.531	.526	.044	.591	.399	.562	.556
88M	.082	.024	-.246	-.136	-.373	.078	.121	-.144	-.084	-.328
91A	.164	.160	.061	.118	-.072	.156	.200	.101	.158	-.039
95B	.114	.844	.725	1.141	1.149	.108	.856	.734	1.143	1.154
All	1.000	.262	.120	.184	.093	.951	.320	.176	.232	.139

Table 8.15
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + Computer-Administered Psychomotor and Criterion = Overall Performance

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.658	.592	.608	.592	.287	.671	.602	.632	.613
13B	.076	.030	-.210	-.054	.049	.072	.089	-.195	-.039	.067
19K	.076	.229	.044	-.039	-.040	.072	.272	.079	-.028	-.023
31C	.016	.167	-.632	-.046	-.058	.015	.240	-.682	-.025	-.039
63B	.126	.145	-.085	.065	.040	.120	.175	-.056	.077	.049
71L	.046	.351	-.029	.032	-.250	.044	.371	-.014	.048	-.240
88M	.082	.381	.081	-.070	-.239	.078	.418	.124	-.055	-.224
91A	.164	.243	.127	-.018	-.031	.156	.267	.149	-.001	-.016
95B	.114	.304	.052	.091	.031	.108	.337	.075	.108	.054
All	1.000	.361	.177	.187	.150	.951	.389	.197	.204	.168

Table 8.16
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + ABLE and Criterion = Core Technical Proficiency

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.610	.506	.619	.650	.287	.618	.511	.661	.691
13B	.076	-.340	-.463	-.515	-.676	.072	-.200	-.337	-.325	-.499
19K	.076	-.487	-.734	-.505	-.510	.072	-.322	-.596	-.402	-.406
31C	.016	.408	-.293	-.061	-.160	.015	.483	-.246	-.037	-.131
63B	.126	-.052	-.312	.015	.099	.120	.012	-.262	.068	.149
71L	.046	.547	.357	.575	.575	.044	.621	.427	.611	.603
88M	.082	-.107	-.572	-.391	-.475	.078	-.020	-.503	-.320	-.407
91A	.164	.041	-.079	-.049	-.198	.156	.068	-.053	.006	-.142
95B	.114	.718	.610	.553	.228	.108	.756	.650	.613	.289
All	1.000	.226	.044	.160	.097	.951	.280	.090	.225	.160

Table 8.17
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + ABLE and Criterion = Overall Performance

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.599	.544	.681	.671	.287	.637	.577	.697	.685
13B	.076	-.059	-.209	-.073	.012	.072	.024	-.174	-.053	.040
19K	.076	.043	-.160	-.077	-.102	.072	.093	-.125	-.055	-.083
31C	.016	.227	-.650	-.035	-.346	.015	.311	-.652	-.023	-.376
63B	.126	.230	.043	.041	-.089	.120	.256	.057	.054	-.071
71L	.046	.242	-.105	.014	-.193	.044	.298	-.059	.029	-.165
88M	.082	.437	.221	-.083	-.250	.078	.462	.243	-.065	-.227
91A	.164	.226	.093	-.046	-.101	.156	.256	.118	-.033	-.087
95B	.114	.327	.076	.171	.449	.108	.354	.116	.182	.456
All	1.000	.333	.168	.205	.184	.951	.372	.198	.220	.200

Table 8.18
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + AVOICE and Criterion = Core Technical Proficiency

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.613	.485	.567	.573	.287	.641	.506	.599	.605
13B	.076	-.397	-.526	-.428	-.606	.072	-.237	-.378	-.281	-.463
19K	.076	-.071	-.207	-.585	-.633	.072	.048	-.101	-.427	-.467
31C	.016	.684	.373	.134	.281	.015	.757	.440	.193	.312
63B	.126	-.072	-.347	-.008	.046	.120	-.012	-.291	.035	.081
71L	.046	.748	.622	.632	.728	.044	.818	.690	.692	.776
88M	.082	-.120	-.593	-.315	-.278	.078	-.028	-.497	-.256	-.224
91A	.164	.035	-.083	-.045	-.153	.156	.100	-.020	-.007	-.102
95B	.114	.747	.606	.708	.585	.108	.794	.657	.741	.632
All	1.000	.266	.088	.172	.141	.951	.331	.149	.230	.200

Table 8.19
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + AVOICE and Criterion = Overall Performance

	Assign All Applicants					Assign 95 Percent of Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.677	.603	.685	.673	.287	.700	.621	.697	.684
13B	.076	.058	-.203	-.064	.057	.072	.108	-.196	-.043	.082
19K	.076	.399	.233	-.060	-.075	.072	.428	.262	-.043	-.072
31C	.016	.346	-.518	-.047	.013	.015	.380	-.527	-.034	.016
63B	.126	.148	-.109	.045	.023	.120	.176	-.090	.059	.035
71L	.046	.333	-.124	.052	-.045	.044	.362	-.097	.071	-.017
88M	.082	.162	-.249	-.075	-.087	.078	.188	-.242	-.060	-.068
91A	.164	.083	-.139	-.049	-.120	.156	.115	-.107	-.034	-.105
95B	.114	.360	.099	.132	.222	.108	.378	.101	.150	.242
All	1.000	.346	.124	.205	.201	.951	.374	.142	.220	.216

Table 8.20
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + All Experimental Predictors and Criterion = Core Technical Proficiency

	Assign All Applicants					Assign 95 Percent Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.523	.313	.439	.389	.287	.561	.349	.467	.417
13B	.076	-.237	-.475	-.268	-.423	.072	-.066	-.330	-.180	-.333
19K	.076	.021	-.245	-.563	-.688	.072	.145	-.132	-.402	-.521
31C	.016	.898	.428	-.257	-.080	.015	.892	.417	-.193	.019
63B	.126	.159	-.167	.203	.206	.120	.217	-.107	.227	.228
71L	.046	.711	.441	.700	.745	.044	.810	.527	.733	.782
88M	.082	.211	-.347	-.104	-.300	.078	.317	-.275	-.046	-.238
91A	.164	.211	.029	.090	-.192	.156	.269	.084	.151	-.122
95B	.114	.808	.598	.717	.406	.108	.817	.606	.757	.464
All	1.000	.352	.090	.211	.082	.951	.417	.148	.262	.139

Table 8.21
Values of Two Classification Efficiency Indices for Assigning Army Applicants Under Two Conditions of Assignment Strategy and Two Predictor Composite Weighting Systems: Predictor Set = ASVAB + All Experimental Predictors and Criterion = Overall Performance

	Assign All Applicants					Assign 95 Percent Applicants				
		LS Weights		Synth Weights			LS Weights		Synth Weights	
MOS	Quota	eMPP	reMAP	eMPP	reMAP	Quota	eMPP	reMAP	eMPP	reMAP
11B	.301	.629	.487	.609	.588	.287	.656	.512	.633	.614
13B	.076	.083	-.328	-.048	-.077	.072	.179	-.315	-.033	.092
19K	.076	.382	.094	-.052	-.063	.072	.455	.141	-.036	-.036
31C	.016	.706	-.540	-.014	-.152	.015	.776	-.526	.000	-.135
63B	.126	.305	-.057	.062	-.035	.120	.350	-.019	.083	-.017
71L	.046	.577	-.122	.058	-.116	.044	.627	-.082	.068	-.108
88M	.082	.701	.300	-.064	-.261	.078	.752	.333	-.048	-.246
91A	.164	.349	.122	-.010	-.085	.156	.378	.153	.008	-.073
95B	.114	.613	.283	.167	.261	.108	.640	.306	.185	.277
All	1.000	.485	.184	.198	.161	.951	.527	.214	.218	.180

Table 8.22
Minimum and Maximum Mean Standard Score Values of reMAP Among Six Predictor Sets and Two Criteria of reMAP Estimators by MOS

	Core Technical Proficiency				Overall Proficiency			
	LS Weights		Synth Weights		LS Weights		Synth Weights	
MOS	Min	Max	Min	Max	Min	Max	Min	Max
11B	.31(f)	.65(a)	.16(c)	.65(d)	.49(f)	.67(a)	.59(f)	.68(a)
13B	-.62(c)	-.46(d)	-.68(a)	-.42(f)	-.33(f)	-.05(a)	.01(d)	.09(b)
19K	-.73(d)	-.21(e)	-.71(c)	-.50(b)	-.16(d)	.23(e)	-.10(d)	-.03(b)
31C	-.30(b)	.43(f)	-.26(b)	.28(e)	-.65(d)	.08(b)	-.35(d)	.05(a)
63B	-.35(e)	-.05(a)	.05(e)	.21(f)	-.11(e)	.09(a)	-.09(d)	.14(b)
71L	.36(d)	.62(e)	.39(a)	.75(f)	-.23(b)	-.03(a)	-.25(c)	-.05(e)
88M	-.59(e)	-.18(b)	-.50(a)	-.28(e)	-.25(e)	.30(f)	-.26(f)	-.02(b)
91A	-.08(e)	.06(c)	-.36(b)	-.07(c)	-.14(e)	.13(c)	-.21(b)	-.03(c)
95B	.60(f)	.77(b)	.23(d)	1.15(c)	-.02(b)	.28(f)	.03(c)	.45(d)

Note. Random selection would yield .00 for all entries.

Predictor Set Key: (a) = ASVAB Only
(b) = ASVAB + Spatial
(c) = ASVAB + Computer-Administered Psychomotor
(d) = ASVAB + ABLE
(e) = ASVAB + AVOICE
(f) = ASVAB + All Experimental Predictors

It is very important to note that these results are highly dependent on a very specific set of conditions of assignment. In this case, the conditions approximated are those where a linear programming algorithm is to have been applied to raw predicted performance scores with a specific, albeit "realistic," set of incumbency quotas. The results at the level of individual MOS could be dramatically altered by changing the distribution of incumbency quotas, altering the metric of the raw predicted performance scores, or using an altered assignment algorithm that included differential priority (utility) information, that is, directly valuing some MOS higher than others. It seems unlikely that the specific conditions which have been investigated here would be of interest for immediate operational application; they serve primarily to highlight the important point that the "mean" classification efficiency values, while very important, can disguise some deleterious effects at the level of the individual MOS.

An important issue is how well the results of the LVI data will generalize to other MOS that were not included in the study. In terms of classification efficiency, these results would indicate that adding groups to the assignment process beyond those already included in an investigation of classification efficiency could involve substantial risk. This is the case because the changes to the individual MOS in expected performance are largely unknown, and could result in a substantial undesired reduction in important MOS.

Estimates of Mean Actual Performance, reMAP, Averaged Across MOS

Table 8.23 shows the average estimates of reMAP, extracted from Tables 8.10 - 8.21, for easier reference. Also shown are the means and standard deviations of the indices across the six sets of predictor variables. Figures 8.2 through 8.5 provide graphic summaries of the results for reMAP.

Before returning to the five issues raised at the beginning of this section, we note first the general effect of classification over and above the effect of selection. The values in the "All Assigned" part of Table 8.23 show the effect of classification, since all applicants are assigned and no one is selected out. The values in the "95% Assigned" part of Table 8.23 show the combined effects of selecting out 5 percent of applicants and classifying the remainder. A working estimate of the gain due to selection can be obtained by subtracting the appropriate values in the "All Assigned" part from the "95% Assigned" part. (For example, .189 - .132 = .057 selection gain for ASVAB only, Core Technical Performance.) We say "working estimate" because the proportions assigned to each MOS differ across the "All Assigned" and "95% Assigned" conditions, as described above (see "Quota Conditions"), confounding such effects with selection effects.

In light of these results, we discuss below the remaining three issues raised at the beginning of this section (individual MOS results have already been discussed).

Least-Squares and Synthetic Weights. In general, weighting system made a small to very small difference, but not in a consistent direction (see Figures 8.2 - 8.5). For the Core Technical Proficiency criterion, least-squares weights produced slightly higher reMAP values for the ASVAB only, ASVAB plus Spatial, and ASVAB plus Psychomotor predictor sets, but synthetic weights produced slightly higher values for ASVAB plus ABLE and ASVAB plus AVOICE. The weighting systems showed virtually identical reMAP values when all predictors were used for CTP; for the Overall Performance criterion, the reMAP values were generally much closer between the two weighting systems. The ASVAB plus AVOICE predictor showed the largest difference, that being in favor of the synthetic weights.

As shown in Table 8.23, averaged across all the predictor sets, the reMAP differences between the two systems are in the third decimal. However, the synthetic weights consistently show less variation in reMAP values across predictor sets (about one-half the standard deviation value of least-squares weights) no matter which criterion or quota system is in effect.

Predictor Combinations. These results show that the ASVAB-only predictor set produces nearly as much, if not more, gain as the ASVAB combined with other predictors. For Core Technical Proficiency, the highest gain is produced by ASVAB plus Spatial using least-squares weights, but it is only .004 greater than ASVAB-only. Most of the other least-squares combinations are noticeably smaller. The synthetic weights show highest reMAP values that are about .02 less than the ASVAB-only, least-squares weights for CTP.

Table 8.23
Values of Two Classification Indices Averaged Across Nine Army Jobs for Two Criteria, Two Assignment Strategies, Two Weighting Systems, and Six Predictor Composites

	All Assigned				95 Percent Assigned			
	LS Weights		Synth Weights		LS Weights		Synth Weights	
	eMPP	reMAP	eMPP	reMAP	eMPP	reMAP	eMPP	reMAP
Core Technical Proficiency								
ASVAB only	.172	.163	.142	.132	.231	.222	.200	.189
ASVAB + spatial	.194	.167	.156	.099	.255	.226	.215	.155
ASVAB + psychomotor	.262	.120	.184	.093	.320	.176	.232	.139
ASVAB + ABLE	.226	.044	.160	.097	.280	.090	.225	.160
ASVAB + AVOICE	.266	.088	.172	.141	.331	.149	.230	.200
ASVAB + all predictors	.352	.090	.211	.082	.417	.148	.263	.139
AVG across sets	.245	.112	.171	.107	.306	.169	.228	.164
SD across sets	.058	.044	.022	.021	.061	.047	.019	.023
Overall Performance								
ASVAB only	.234	.219	.194	.188	.260	.244	.213	.206
ASVAB + spatial	.256	.188	.197	.187	.280	.209	.208	.201
ASVAB + psychomotor	.361	.177	.187	.150	.389	.197	.204	.168
ASVAB + ABLE	.333	.168	.205	.184	.372	.198	.220	.200
ASVAB + AVOICE	.346	.124	.205	.201	.374	.142	.220	.216
ASVAB + all predictors	.485	.184	.198	.161	.527	.214	.218	.180
AVG across sets	.336	.177	.198	.179	.367	.201	.214	.195
SD across sets	.081	.028	.006	.017	.087	.031	.006	.016

Note. Values in standard deviation units.

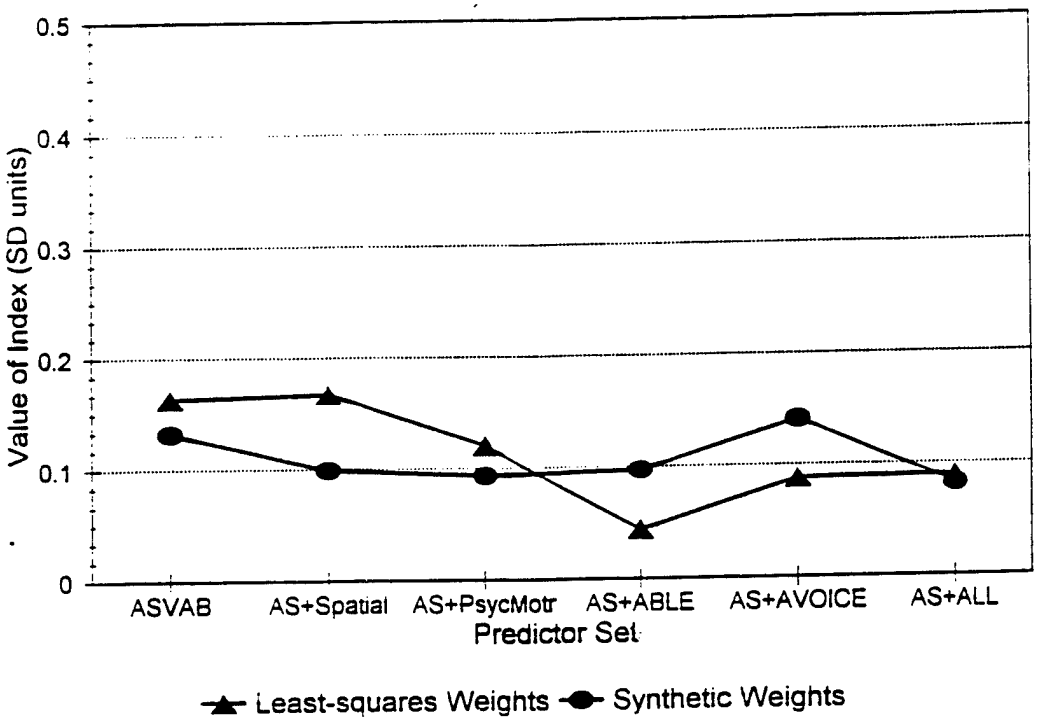


Figure 8.2. Classification efficiency, as measured by reMAP, comparing least-squares and synthetic weights using the Core Technical Proficiency criterion, selecting all applicants.

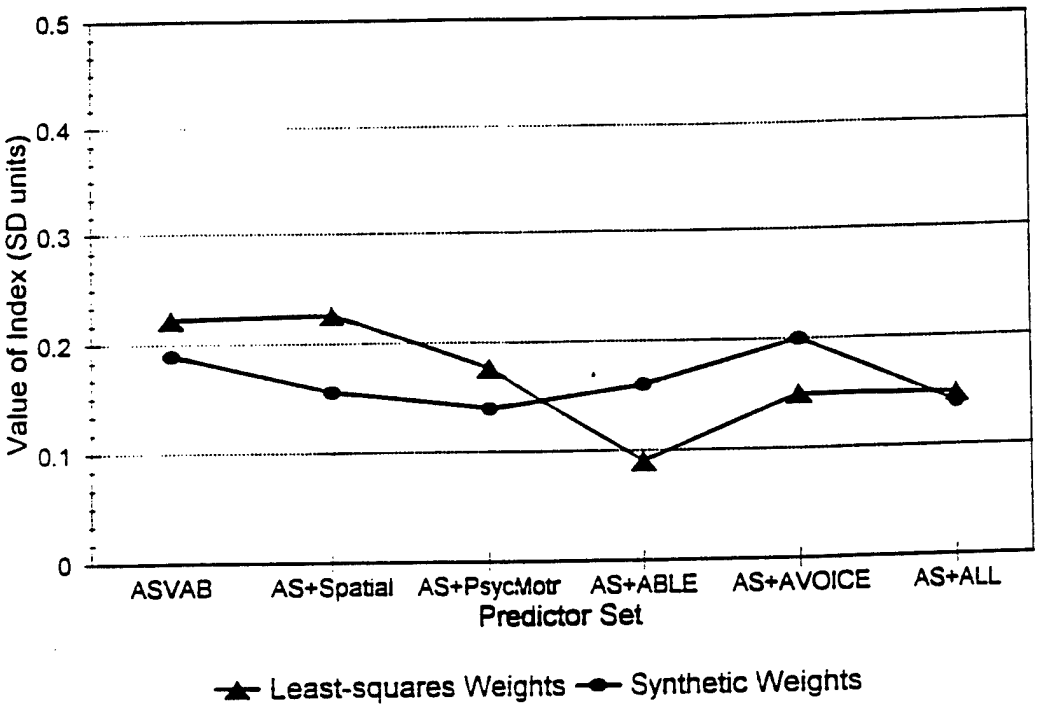


Figure 8.3. Classification efficiency, as measured by reMAP, comparing least-squares and synthetic weights using the Core Technical Proficiency criterion, selecting 95% of applicants.

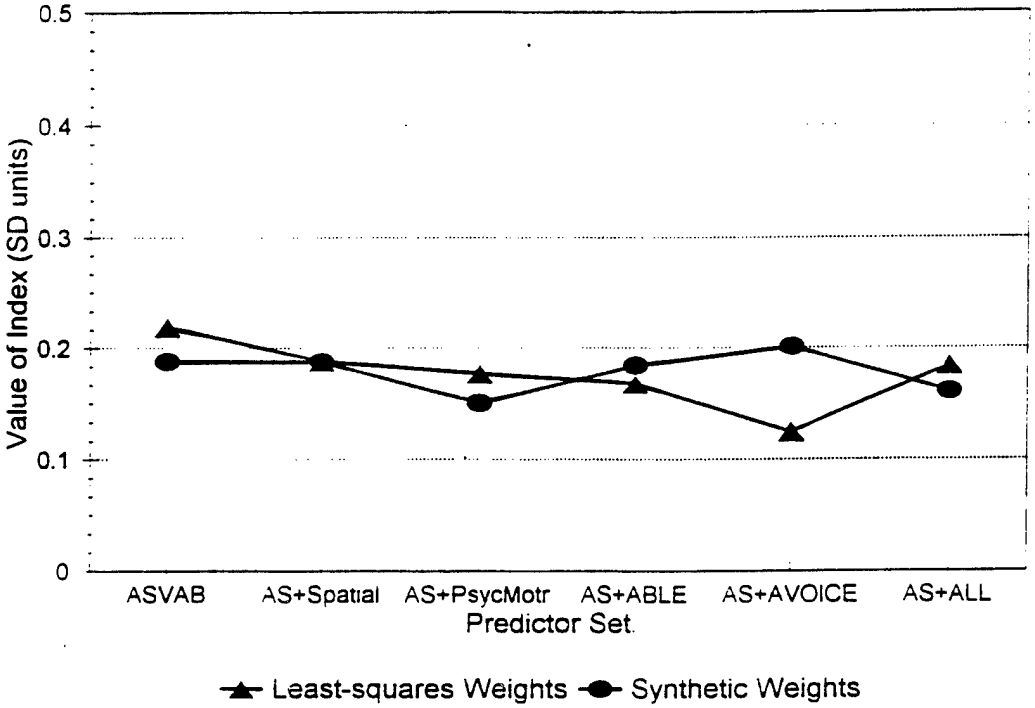


Figure 8.4. Classification efficiency, as measured by reMAP, comparing least-squares and synthetic weights using the Overall Performance criterion, selecting all applicants.

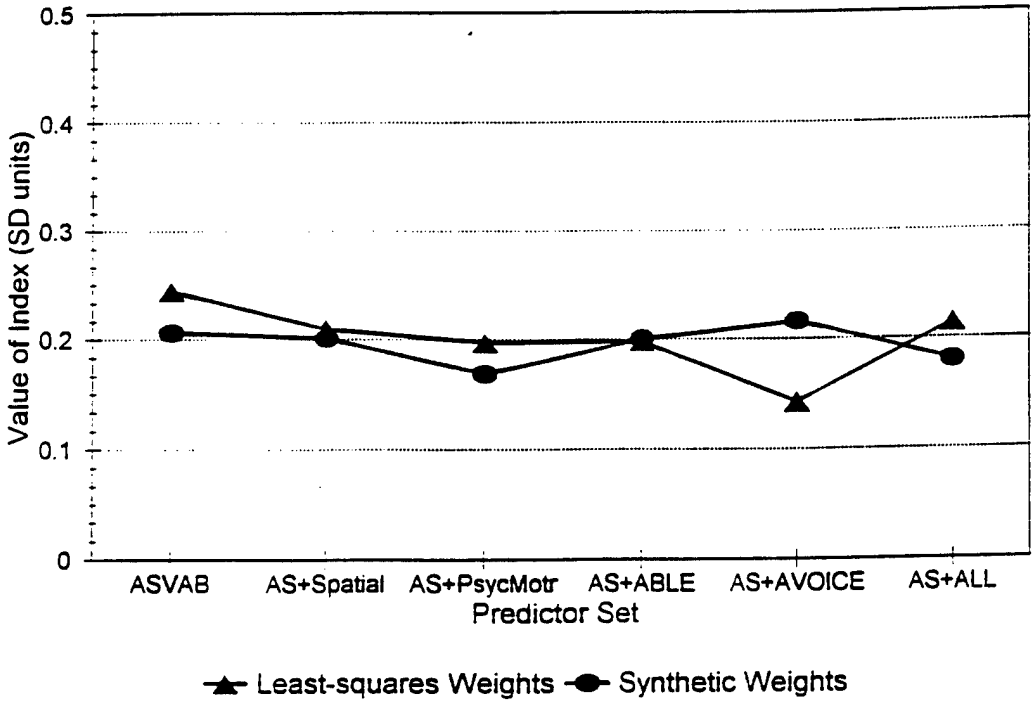


Figure 8.5. Classification efficiency, as measured by reMAP, comparing least-squares and synthetic weights using the Overall Performance criterion, selecting 95% of applicants.

For Overall Performance, the ASVAB-only combination (least-squares weights) also shows the highest gain. However, the differences in reMAP values are much smaller across the predictor combinations, no doubt due to the importance of the other predictors for predicting this more comprehensive criterion. Again, the highest synthetically weighted predictor combination is about .02-.03 points less than the ASVAB-only least-squares weighted combination.

Interestingly, the ASVAB plus AVOICE combination produces the highest reMAP value for the synthetic weights, in all cases. It significantly outperforms the least-squares weights for ASVAB plus AVOICE in all four conditions, and outperforms all least-squares combinations except ASVAB-only or ASVAB plus Spatial for CTP.

The Criterion. Table 8.23 shows that higher reMAP values are obtained using the Overall Performance criterion than using the Core Technical Proficiency criterion. This holds true for all four quota/weighting system conditions when values are averaged across all predictor combinations. These differences were about .03 points (for 95% Assigned) to .07 points (All Assigned). It also holds true for all but two of the 24 separate comparisons across predictor combination/quota/weighting system conditions—those two being ASVAB plus Spatial and ASVAB plus AVOICE combinations in the 95% Assigned, least-squares conditions. These findings in favor of Overall Performance are somewhat surprising. They may be due to the greater predictability of the more comprehensive criterion because of its higher reliability (but we note that the Core Technical Proficiency criterion is highly reliable itself), its inclusion of all five major components of job performance, or the differential weighting across MOS of the five major components.

Classification Efficiency of Full vs. Reduced Prediction Equations

In Chapter 4 of this report the population estimates of selection validity were computed for prediction equations using (a) all the Experimental Battery predictor composite scores plus the four ASVAB factor scores, and (b) two reduced equations that were limited to 10 or fewer predictors (including ASVAB). The two reduced equations were obtained via an expert judgment procedure that emphasized either maximizing selection validity or maximizing classification efficiency for the nine Batch A MOS when CTP was used as the maximizing function. These three equations were also compared in terms of their classification efficiency by using both eMPP and reMAP to evaluate the effects of three sets of job assignments based on linear programming.

Each set of job assignments made use of one of the three prediction equations. Consequently, it is possible to ask whether the expert judgments produced a difference in the two sets of reduced equations in terms of the estimates of classification efficiency yielded by each one. If the expert judgments achieved the intended objective, then the reduced equations developed to maximize classification should yield higher estimates of classification efficiency than the reduced equations developed to maximize selection validity.

The results, averaged across MOS, are shown in Table 8.24. The estimation procedure again used linear programming to make job assignments in proportion to FY 93 accessions under the two conditions of (a) no prior selection, and (b) elimination of 5 percent of the sample, using AFQT. As before, each set of job assignments was repeated 50 times on simulated samples of applicants.

The top row in Table 8.24 is analogous to the sixth row (ASVAB + all predictors) in Table 8.23, columns 1 and 2, and 5 and 6 (least-squares weights). The actual values in the two tables are slightly different because Table 8.24 is based on an independent set of simulated samples and, for the ASVAB, this table used the four factor scores instead of the nine subtest scores that were used in the computations for Table 8.23. The values in Table 8.23 are slightly higher, which may mean that using the four ASVAB factor scores loses a small amount of classification information, as compared to the full set of nine subtests.

Table 8.24
Values of Two Classification Indices Averaged Across Nine Army Jobs for the Core Technical Proficiency Criterion and Three Types of Prediction Equations

	Aggregate Classification Efficiency			
	eMPP		reMAP	
	All Assigned	95 Percent Assigned[a]	All Assigned	95 Percent Assigned[a]
Full Equation (All Preditors)	.325	.401	.081	.153
Reduced Equation: Selection	.248	.317	.113	.180
Reduced Equation: Classification	.278	.361	.136	.217

Note. Least-squares weighting system. Values in standard deviation units.

[a] Applicant pool reduced 4.9% by prior selection on AFQT.

A comparison of the full equation with the reduced equations (using reMAP) shows that the reduced equations produce greater classification efficiency than the full equation, presumably because they contain less "noise." The result is not in the same direction for eMPP because it does not correct for "shrinkage" at the job assignment stage. This shrinkage is greater for the longer equation. That is, eMPP is proportionally more inflated when more predictors are used.

Comparison of the classification gains produced by the two sets of expert-developed reduced equations shows a small advantage in favor of the equations that were in fact intended to emphasize classification. The classification equations produced a 15-20 percent greater gain than the selection equations.

These are tantalizing results and they invite a much more complete examination of how maximum potential classification efficiency can best be achieved, and of how such gains will be affected by variation in the critical constraints that are part of any operational system.

Concluding Comments

This section has used classification efficiency analyses to provide information relevant to a number of issues around the use of the ASVAB and Project A/Career Force experimental predictors. Each of these issues has been discussed above. Some more general conclusions are offered here.

First, these analyses highlight the intricacy of studying classification efficiency and the considerable amount of research remaining before we reach ground stable enough to enable us to form firm conclusions. We think the results presented in this chapter sound a cautionary note for the acceptance of mean predicted performance, as measured by Brogden's Allocation Average, as the appropriate index of classification efficiency. MPP is, after all, intended as an estimate of Mean Actual Performance (MAP). We have offered here an alternative estimate of MAP, called reMAP, that seems to more closely match the behavior of MAP in practical situations. We have striven to provide a comprehensive rationale and explanation of this estimate and welcome its scrutiny by other interested investigators.

Second, it seems clear that classification strategies can increase productivity, certainly over random allocation of individuals, and perhaps in fairly realistic scenarios such as the "95% assigned" condition investigated here. We say "perhaps" because these investigations point out the relative fragility of any classification strategy. That is, modifying the mix of jobs that are the targets for classification, changing the proportions assigned to each job, or causing increases or decreases in the underlying validity of predictor composites for the jobs can produce rather dramatic shifts in the expected increase in level of productivity, especially at the individual MOS level.

These two points seems to us to point to the need for relatively specific modeling exercises aimed at particular formulations of more general classification strategies when operational classification procedures are being considered. Although the Project A/Career Force database is not all-inclusive, it is the most comprehensive dataset available now or likely to be available in the near future. It should be fully utilized in this regard. Careful consideration must be given to the best way to supplement and take advantage of the database for these purposes.

Finally, a not unreasonably optimistic interpretation of these results appears to highlight some bright spots. It appears that the ASVAB alone can provide useful

classification efficiency. It also appears that judgmental methods, as operationalized in the Army's synthetic validation strategy and in the reduced equation analysis, can be used (a) to create predictor equations with relatively small losses in classification efficiency from that to be expected using relatively unobtainable least-squares equations, and (b) to select optimal sets of predictors that maximize classification efficiency with virtually no loss in selection validity.

Chapter 9
THE FINAL CHAPTER

John P. Campbell and James H. Harris

This chapter is intended to be a stand-alone presentation that can serve as an overall summary of the two research projects known as Project A and Career Force. It attempts to emphasize the most critical findings and to organize them around the original goals for the two projects. It is also intended as a statement of appreciation to both the U.S. Army Research Institute and the U.S. Army by the members of the research consortium for the opportunity to participate in these two landmark projects.

The research was performed in two phases, starting in 1982. The first phase was Project A; its overall goals were to validate the Armed Services Vocational Aptitude Battery (ASVAB), as well as a comprehensive battery of project-developed experimental tests, by collecting data from a representative sample of Military Occupational Specialties (MOS) using criterion measures that represent the entire domain of critical performance components. Phase two of the research program was Career Force, the overall goals of which were to determine the longitudinal relationship between the new predictors and first-tour performance, and to examine how selection and classification tests administered before a soldier's first enlistment, in conjunction with performance assessments obtained during that soldier's first enlistment, predict performance in a second enlistment.

THE SPECIFIC GOALS AND DESIGN OF PROJECT A AND CAREER FORCE

Project A and Career Force were designed to provide the Army with the greatest possible increase in overall performance and readiness that could be obtained from improved selection, classification, and allocation of enlisted personnel. These two research projects provided an integrated examination of performance measurement, selection/classification methods, supply and demand parameters, and allocation procedures such that the Army could attempt to optimize the achievement of multiple personnel management goals (e.g., increase performance and decrease attrition).

The impetus for the research program came from the practical, professional, and legal need to demonstrate the validity of the ASVAB and other selection variables for predicting job performance. Much of the existing validity data was based on using training measures as criteria. As ARI began reviewing the design needed to meet that requirement, the concept of a larger program began to emerge. With only a moderate amount of additional resources, new selection/classification measures in the perceptual, psychomotor, interest, temperament, and biodata domains could be evaluated as well. In addition, a longitudinal research database could be developed, linking soldiers' performance on a variety of variables from enlistment, through training, first-tour assignments, reenlistment decisions, and for some, to their second tour.

Specific Program Objectives

The specific objectives of Project A were to:

(1) Develop new measures of job performance that could be used as criteria against which to validate selection/classification measures.

(2) Develop a general model of performance for entry-level skilled jobs.

(3) Validate existing selection measures against both existing and project-developed criteria.

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

(5) Develop and validate new selection and classification measures.

(6) Develop a utility scale for different performance levels across MOS.

The specific objectives of Career Force were to:

(1) Develop a complete array of valid and reliable measures of second-tour performance as an Army NCO, using the Project A prototypes as a starting point.

(2) Develop a model of second-tour NCO performance that parallels the first-tour performance model from Project A and that identifies the major components of second-tour performance, provides information on their construct validity, and establishes how they should be combined for specific prediction or interpretation purposes.

(3) Carry out a complete incremental predictive validation of (a) the ASVAB and the Project A Experimental Battery of predictors, (b) measures of training success, and (c) the full array of first-tour performance criteria developed as part of Project A. The criteria against which these three sets of predictors were validated, both individually and incrementally for each major criterion component, are the second-tour job performance measures.

(4) Estimate the degree of differential prediction across (a) major domains of predictor information (e.g., abilities, personality, interests), (b) major factors of job performance, and (c) different types of 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 the analytic framework needed to evaluate the optimal equations for predicting (a) training performance; (b) first-tour performance; (c) first-tour attrition and the reenlistment decision; and (d) second-tour performance, under the conditions when testing time is limited to a specified amount and when there must be a tradeoff among alternative selection/ classification goals (e.g., maximizing aggregate performance vs. minimizing discipline and low-motivation problems vs. minimizing attrition).

(7) Design and develop a fully functional and user-friendly research database that includes all relevant personnel data on 1981/82, 1983/84, and 1986/87 accessions. All Project A and Career Force Project data and all relevant Enlisted Master File, Accession File, and Army Training Requirements and Resources System data are included.

The Research Samples

It is with some wonderment that we note that all the above objectives were in fact achieved, for both projects. The detailed record is contained in the Project A/Career Force annual reports (FY83 - FY94) and in the deliverable technical reports; in a few instances, for the convenience of readers needing specific details, we have included references to chapters within this report. Not all of the objectives are addressed in this summary, but information on the exceptions is treated elsewhere (scaling performance utility is referenced in this chapter, estimating differential prediction across subgroups is described in Chapter 5 of this report, and describing the database format will be reported separately).

In general, the combined design for Project A/Career Force encompasses two major cohorts of soldiers (new accessions for 1983/84 and for 1986/87), both of which were followed into their second tour of duty and which collectively have produced six major research samples. For each sample there is a battery of predictor measures and an array of performance measures. For each of the six samples the predictor battery is composed of the ASVAB and either the Trial Battery or the Experimental Battery version of the new tests developed in Project A (see Campbell & Zook, 1991). Three distinct arrays of performance measures corresponded to the need to assess (a) training performance, (b) first-tour job performance, and (c) second-tour job performance.

The MOS in the two groups were carefully sampled to represent the variation in job content in the Army occupational structure. In addition, they were selected so as to overrepresent both the combat specialties and those MOS with the larger proportions of women and minority groups. The MOS selection procedure has been described in detail in previous Project A reports (e.g., Campbell, 1987).

In each sample the individuals to be assessed were selected from two predetermined sets of MOS -- Batch A and Batch Z. They are listed in Figure 9.1. The two sets differed in that tests administered to soldiers in Batch A MOS included MOS-specific rating scales, job knowledge tests, and hands-on tests, whereas the only MOS-specific measure administered to the soldiers in Batch Z MOS was an end-of-training test.

Batch A		Batch Z	
MOS		MOS	
11B	Infantryman	12B	Combat Engineer
13B	Cannon Crewmember	16S	MANPADS Crewman
19E	M60 Armor Crewman	27E	Tow/Dragon Repairer
19K	M1 Armor Crewman[a]	29E	Comm-Electronics Radio Repairer[d]
31C	Single Channel Radio Operator	51B	Carpentry/Masonry Specialist
63B	Light-Wheel Vehicle Mechanic	54B	NBC Specialist[e]
71L	Administrative Specialist	55B	Ammunition Specialist
88M	Motor Transport Operator[b]	67N	Utility Helicopter Repairer
91A/B	Medical Specialist/Medical NCO[c]	76Y	Unit Supply Specialist
95B	Military Police	94B	Food Service Specialist
		96B	Intelligence Analyst[d]

[a] Except for the type of tank used, this MOS is equivalent to the 19E MOS originally selected for Project A testing.
[b] This MOS was formerly designated as 64C.
[c] Although 91A was the MOS originally selected for Project A testing, second-tour medical specialists are usually reclassified as 91B.
[d] This MOS was added after the Concurrent Validation
[e] This MOS was formerly designated as 54E.

Figure 9.1. Project A/Career Force Military Occupational Specialties (MOS).

A glossary of terms for the samples and for the different measurement batteries is given in Figure 9.2. The six major samples, their approximate size, and the predictor and/or performance batteries that were to be administered to each are shown in Figure 9.3.

Procedure and Design

The data collection procedures for each sample have been described in detail in previous reports (e.g., see Campbell & Zook, 1990). Each data collection involved on-site administration by a trained data collection team headed by a team leader from the contractor staff who worked closely with a designated Army point-of-contact at the site. Each of the six samples is briefly characterized below in terms of the timing, location, and duration (per soldier) of the data collection.

The Concurrent Validation (CVI) sample. The data were collected at 13 posts in the continental United States and at multiple locations in Germany. Each individual was assessed for 1 1/2 days on the project-developed first-tour job performance measures and for 1/2 day on the new predictor measures (the Trial Battery). Most of the individuals in the sample had been in the Army for 18-24 months.

The Longitudinal Validation (LV) Sample. All individuals were assessed on the 4-hour Experimental Predictor Battery within 2 days of first arriving at their assigned Reception Battalion where they would undergo basic/advanced individual training. Data were collected over a 14-month period at eight Reception Battalions by a permanent, on-site data collection team.

Glossary of Terms	
CVI Sample (CVI)	Soldiers who entered the Army between 1 Jul 83 - 30 Jun 84 and were in 1985 Project A Concurrent Validation. They were administered the Trial Predictor Battery and the first-tour job performance measures.
CVII Sample (CVII)	Soldiers who entered the Army between 1 Jul 83 - 30 Jun 84 and were in the 1985 Project A Concurrent Validation (CVI) and the 1988 Second-Tour Concurrent Validation (CVII). They were administered the second-tour job performance measures and were re-administered the ABLE.
LV Sample (LV)	Soldiers in the Longitudinal Validation sample who entered the Army between 20 Aug 86 - 30 Nov 87 and were administered the Experimental Predictor Battery and End-of-Training measures.
LV Training Sample (LVT)	Soldiers in the Longitudinal Validation sample who finished AIT and who were administered the End-of-Training measures.
LVI Sample (LVI)	Soldiers who entered the Army between 20 Aug 86 - 30 Nov 87 and were in the LV Sample and the 1988 First-Tour Longitudinal Validation Sample. They were administered the first-tour job performance measures.
LVII Sample (LVII)	Soldiers who entered the Army between 20 Aug 86 - 30 Nov 87 and were in the LVI Sample and the Longitudinal Validation (LVII) sample. They were administered the second-tour job performance measures in LVII.
Note. Glossary definitions reflect the original research plan. In actuality, some CVII soldiers did not have CVI data, some LVI soldiers did not have LV data, and some LVII soldiers did not have both LV and LVI data.	

Figure 9.2. Glossary of terms for Project A/Career Force research samples.

The Longitudinal Validation End-of-Training (LVT) Sample. The EOT performance measures were administered to those individuals in the LV sample who completed Advanced Individual Training (AIT) phase of Initial Entry Training, which could take from 2 to 6 months, depending on the MOS. The training performance measures consisted of an MOS-specific training achievement test and a series of rating scales completed by peers and drill instructors. Data collection took place during the last three days of AIT.

The Longitudinal Performance Measurement (LVI) Sample. The individuals in the 86/87 cohort who were measured with the Experimental Predictor Battery, completed AIT, and remained in the Army were assessed with the full array of first-tour job performance measures when most of them were between 18-24 months of service. Data were collected at 13 posts in the United States and multiple locations in Europe (primarily in Germany). The administration of the LVI first-tour criterion measures took one day per soldier.

The Concurrent Validation Second-Tour (CVII) Sample. The same teams that administered the first-tour performance measures to the LVI sample administered the second-tour performance measures at the same location and during the same time periods to a sample of junior NCOs from the 83/84 cohort in their second tour of duty (4-5 years

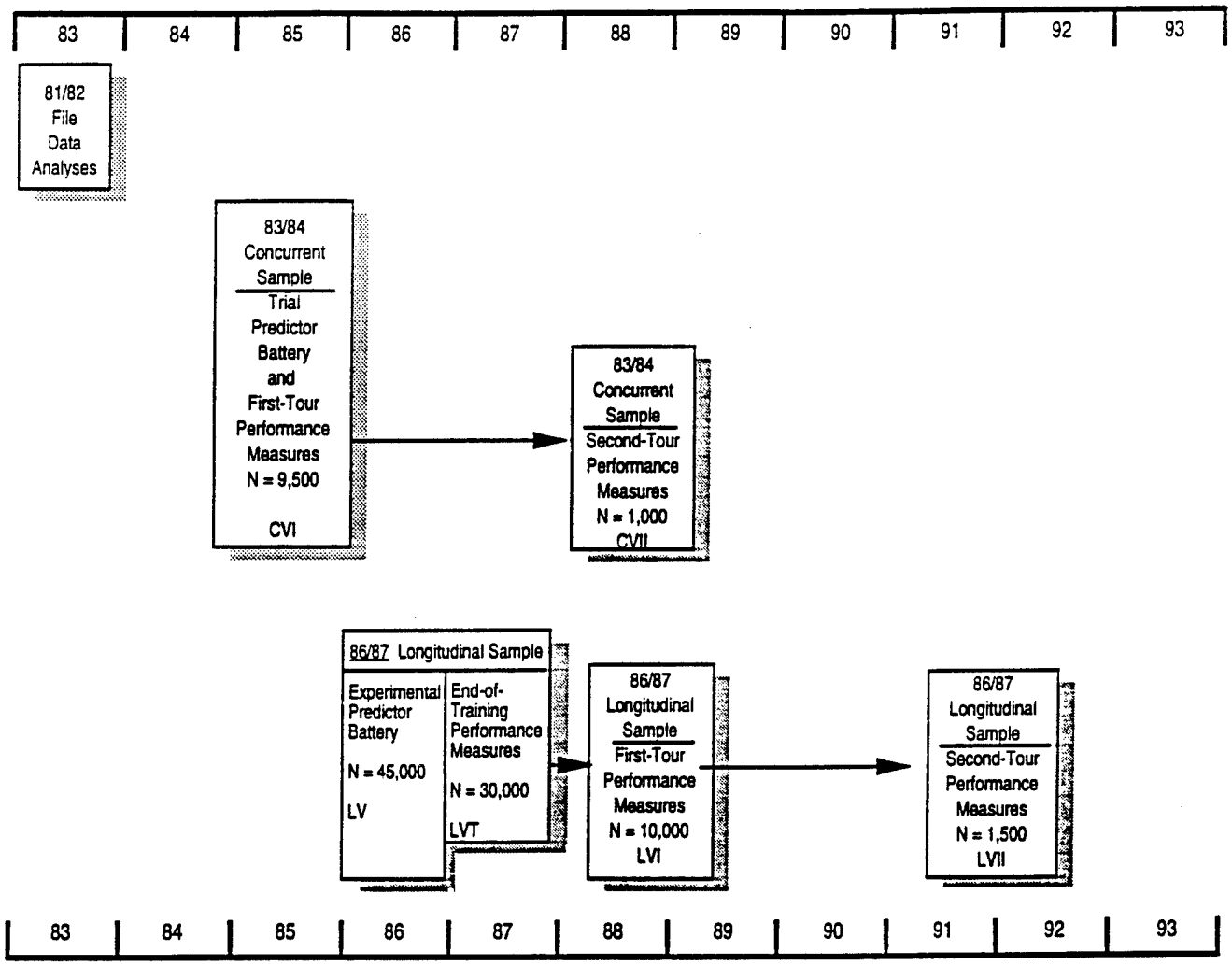


Figure 9.3. Project A/Career Force research flow and samples.

of service). Every attempt was made to include second-tour personnel from the designated MOS who had been part of CVI. The CVII data collection took one day per soldier.

The Longitudinal Validation Second-Tour (LVII) Sample. This sample includes members of the 86/87 cohort from the designated MOS who were part of the LVI (first-tour job performance measures) samples and who reenlisted for a second tour. The revised second-tour performance measures were administered at 15 U.S. posts, multiple locations in Germany, and two locations in Korea. The LVII performance assessment took one day per soldier.

PREDICTOR DEVELOPMENT IN PROJECT A

A major objective was to develop an experimental battery of new selection/classification tests that would be potentially valuable additions to ASVAB and would maximize the Army's capability to make accurate selection/classification decisions. Consequently, the overall Project A strategy was to identify a universe of potential predictor constructs appropriate for the population of enlisted MOS, sample representatively from it,

construct tests for each construct sampled, and refine and improve the measures through a series of pilot and field tests. The intent was to develop a predictor battery that was maximally useful for an entire population of jobs.

The long process of predictor development began with an in-depth search of the entire personnel selection literature. Literature review teams were created for cognitive abilities, perceptual and psychomotor abilities, and non-cognitive characteristics such as personality, interest, and biographical history. Every available automated and manual technique was used in the search and an initial list of several hundred variables was compiled. The list went through several waves of expert review and was eventually reduced to 53 potentially useful predictor variables. They are listed in Table 9.1, where they have been clustered as described below.

A sample of 35 personnel selection experts was then asked to estimate the correlation between each predictor construct and each criterion factor, when that correlation was corrected for restriction of range and criterion unreliability (criterion development is described in the following section). The resulting judgments could be analyzed for the interjudge agreement, rows and columns could be factor analyzed, and the results could be compared to analogous information from the empirical literature. Most importantly, the exercise provided another substantial set of expert judgments about which predictor constructs should be the most useful. A hierarchical analysis of the predictor validity profiles is also shown in Table 9.1.

All the available information was then used to arrive at a final set of variables for which new measures would be constructed. This represented months of effort by many people to select the variables that would best supplement the ASVAB in predicting job performance across all MOS. What followed were many months more of instrument construction, several waves of pilot tests, and a series of major field tests. Included in these efforts were the development of a computerized battery of perceptual/psychomotor tests, the creation of the software, the design and construction of a special response pedestal permitting a variety of responses (e.g., one-hand tracking, two-hand coordination), and the acquisition of portable computerized testing stations. Also developed were several paper-and-pencil cognitive tests and three inventories of temperament, interests, and experience (Assessment of Background and Life Experiences, ABLE; Army Vocational Interest Career Examination; AVOICE, and Job Orientation Blank, JOB). After each data collection, revisions were made on the basis of item statistics and expert review.

The set of new tests that constitute the final version of the Experimental Predictor Battery is shown in Figure 9.4. The title for each specific test, the construct it is intended to measure, and a brief description of the test content are included.

On the basis of further expert review, and an extensive analysis of the intercorrelations of the Experimental Battery scores using the total Longitudinal Validation (LV) sample (N = 30,000), the full array of test and scale scores produced by the ASVAB and the Experimental Battery were grouped into 28 basic predictor composite scores. These scores were the prediction variables used in all subsequent validation analyses. They are shown

Table 9.1
Hierarchical Map of Predictor Space

Constructs	Clusters	Factors
Verbal Comprehension Reading Comprehension Ideational Fluency Analogical Reasoning Omnibus Intelligence/Aptitude Word Fluency	A. Verbal Ability/ General Intelligence	Cognitive Abilities
Word Problems Inductive Reasoning Concept Formation Deductive Logic	B. Reasoning	
Numerical Computation Use of Formula/Number Problems	C. Number Ability	
Perceptual Speed and Accuracy	N. Perceptual Speed and Accuracy	
Investigative Interests	U. Investigative Interests	
Rote Memory Follow Directions	J. Memory	
Figural Reasoning Verbal and Figural Closure	F. Closure	
Two-dimensional Mental Rotation Three-dimensional Mental Rotation Spatial Visualization Field Dependence (Negative) Place Memory (Visual Memory) Spatial Scanning	E. Visualization/Spatial	Visualization/ Spatial
Processing Efficiency Selective Attention Time Sharing	G. Mental Information Processing	Information Processing
13. Mechanical Comprehension	L. Mechanical Comprehension	Mechanical
48. Realistic Interests 51. Artistic Interests (Negative	N. Realistic vs. Artistic Interests	
Control Precision Rate Control Arm-hand Steadiness Aiming	I. Steadiness/Precision	Psychomotor
Multilimb Coordination Speed of Arm Movement	D. Coordination	
Manual Dexterity Finger Dexterity Wrist-finger Speed	K. Dexterity	
Sociability Social Interests	Q. Sociability	Social Skills
Enterprising Interests	R. Enterprising Interests	
Involvement in Athletics and Physical Conditioning Energy Level	T. Athletic Abilities/Energy	Vigor
Dominance Self-esteem	S. Dominance/Self-esteem	
Traditional Values Conscientiousness Non-delinquency Conventional Interests	N. Traditional Values/Conventionality/ Non-delinquency	Motivation/ Stability
Locus of Control Work Orientation	O. Work Orientation/Locus of Control	
Cooperativeness Emotional Stability	P. Cooperation/Emotional Stability	

as Figure 9.5. Because the JOB inventory did not show any significant correlations with measures of performance in CVI, it was omitted from the LVI and LVII validation analysis.

COGNITIVE PAPER-AND-PENCIL MEASURES

Construct/Measure	Content/Description
SPATIAL VISUALIZATION - ROTATION	
Assembling Objects	The test contains 36 items with an 18-minute time limit. The task involves figuring out how a disassembled object will look when its parts are put back together. The parts fit like the pieces of a puzzle. For each item, four alternative assemblies are provided and the subject must pick the correct one.
Object Rotation	The test contains 90 items with a 7 1/2-minute time limit. The task involves examining a test object and determining whether the figure represented in each item is the same as the object, only rotated. Each test object has five test items, each requiring a response of "same" or "not same."
SPATIAL VISUALIZATION - SCANNING	
Maze Test	The test contains 24 items with a 5 1/2-minute time limit. Each item is a rectangular maze with four labeled entrance points and four exit points. The task is to determine which entrance leads to a pathway to one of the exit points.
SPATIAL ORIENTATION	
Orientation Test	The test contains 24 items with a 10-minute time limit. Each item is a picture or scene within a circular or rectangular frame. The bottom of the frame has a circle with a dot inside it. The picture is not in an upright position. The task is to mentally rotate the frame around the stationary picture so that the bottom of the frame is at the bottom of the picture, and then to decide where the dot will appear in the circle.
Map Test	The test contains 20 items with a 12-minute time limit. Subjects are presented with a map including landmarks such as a forest, a lake, etc. In each item, subjects are given compass directions by indicating the direction of one landmark to another. The subject must then determine which compass direction to go to reach yet another landmark.
INDUCTION	
Reasoning Test	The test contains 30 items with a 12-minute time limit. Subjects are presented with a series of four figures. The task is to identify the pattern or sequential relationship among the figures and then to identify from among five possible answers the one figure that appears next in the series.

COMPUTERIZED COGNITIVE/PERCEPTUAL/PSYCHOMOTOR MEASURES
(Examinees use a pedestal apparatus with two joy sticks, two sliding resistors, and colored buttons. For timed tests, the stimulus appears only while the subject's hands hold down four Home Keys.)

REACTION TIME	
Simple Reaction Time	When the subject's hands are in the Ready position, a small box appears on the computer screen. After a delay of 1.5 to 3.0 seconds, the word YELLOW appears in the box. The subject must press the Yellow button on the testing panel, then return his/her hands to the Ready position to receive the next item. The test contains 14 items.
Choice Reaction Time	This measure is similar to the simple reaction time, but rather than seeing the same stimulus (YELLOW) on each trial, the subjects may see either BLUE or WHITE on the screen. When the stimulus appears, the subject is to press the botton that corresponds with the term (BLUE or WHITE) on the screen.

Figure 9.4. The Project A Experimental Predictor Battery (page 1 of 5 pages).

COMPUTERIZED COGNITIVE/PERCEPTUAL/PSYCHOMOTOR MEASURES (Continued)

Construct/Measure	Content/Description
PERCEPTUAL SPEED AND ACCURACY	
Perceptual Speed and Accuracy	This test is designed to measure the ability to rapidly compare two visual stimuli presented simultaneously and determine whether they are the same or different. At the beginning of each trial (with the subject's hands in the Ready position), after a brief delay, the stimuli are presented. The subject must then press a white button if the stimuli are the same or a blue button if the stimuli are different. Three different "types" of stimuli are used: alpha, numeric, and symbolic. Within each type, the length of the stimulus is varied. Three different levels of length are presented: two-character, five-character, and nine-character. The test consists of 36 trials; the primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response.
Target Identification	The subject is presented with a target object and three stimulus objects. The objects are pictures of military vehicles or aircraft (e.g., tanks, planes, helicopters). The target object is the same as one of the stimulus objects. The target may be rotated relative to its stimulus counterpart. The subject must determine which of the three stimulus objects is the same as the target object and then press a button corresponding to that choice. The test consists of 36 trials; the primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response.
MEMORY	
Short-Term Memory	A stimulus contains one, three, or five objects (letters or symbols). Following a delay, the stimulus set disappears. When the probe appears, the subject must decide whether it was part of the stimulus set and press the white button if it was, and the blue button if it wasn't. The test consists of 36 trials.
Number Memory	The subject is first presented with a single number on the computer screen and instructed to press 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 appears along with an operation term (e.g., "Add 9" or "Subtract 6"). Once the subject has combined the first number with the second, he/she must press a button to receive a new number and operation term. This procedure continues until a solution to the problem is presented. The subject must then indicate whether the solution presented is correct or incorrect. In total, the test consists of 28 such items.
PRECISION/STEADINESS	
Target Shoot	At the beginning of a trial, a crosshair appears in the center of the screen and a target box appears at some other location. The target then begins to move about the screen, frequently changing speed and direction. The subject can control movement of the crosshair using a joystick. The task is to move the crosshair into the center of the target and press a red button on the response pedestal to "fire" at the target before the time limit on each trial is reached. The subject receives three scores. The first is the percentage of "hits." The second is the average time elapsed from the beginning of the trial until the subject fires at the target. The third score is the average distance from the center of the crosshair to the center of the target at the time the subject fires at the target. The test consists of 30 trials.
Target Tracking 1	This is a pursuit tracking test. On each trial of the test, subjects are shown a path consisting entirely of vertical and horizontal line segments. At the beginning of the path is a target box. Centered in the box is a crosshair. 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 crosshair centered within the target at all times. The subject uses a joystick to control movement of the crosshair. The subject's score is the average distance from the center of the crosshair to the center of the target across all 18 test trials.

Figure 9.4. The Project A Experimental Predictor Battery (page 2 of 5 pages).

COMPUTERIZED COGNITIVE/PERCEPTUAL/PSYCHOMOTOR MEASURES (Continued)

Construct/Measure	Content/Description
MULTILIMB COORDINATION	
Target Tracking 2	This is a test of multilimb coordination. The test is virtually identical to Target Tracking #1. The only difference is that the subject must use two sliding resistors (instead of a joystick) to control the movement of the crosshair. One controls movement in the vertical plane while the second controls movement in the horizontal plane. As with Target Tracking #1, the subject's score is the average distance from the center of the crosshair to the center of the target across all 18 test trials.
MOVEMENT JUDGMENT	
Cannon Shoot	At the beginning of each trail, a stationary cannon appears on the computer screen. The starting position of this cannon varies. 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 such that the shell intersects the target when the target crosses the shell's line of fire. The test includes 36 items. The primary dependent variable is a deviation score indicating the difference between time of fire and optimal fire time (e.g., direct hits yield a deviation score of zero).

ASSESSMENT OF BACKGROUND AND LIFE EXPERIENCES (ABLE)
(A paper-and-pencil biographical/temperament inventory)

Scale	Content Definition
Emotional Stability	Assesses the amount of emotional stability and tolerance for stress a person 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 time of stress.
Self-Esteem	Is defined as the degree of confidence a person has in his or her abilities. A person with high self-esteem feels largely successful in past undertakings and expects to succeed in future undertakings. A person with low self-esteem feels incapable and is self-doubting.
Cooperativeness	Assesses the degree of pleasantness versus unpleasantness a person exhibits in interpersonal relations. The agreeable and likable person is pleasant, tolerant, tactful, helpful, not defensive and is generally easy to get along with. His or her participation in a group adds cohesiveness rather than friction. A disagreeable and unlikable person is critical, fault-finding, touchy, defensive, alienated, and generally contrary.
Conscientiousness	Assesses a person's tendency to be reliable. The person who scores high on this scale is well organized, planful, prefers order, thinks before acting, and holds him- or herself accountable. The person who scores low tends to be careless and disorganized, and acts on the spur of the moment.
Nondelinquency	Assesses a person's acceptance of laws and regulations. The person who scores high on this scale is rule abiding, avoids trouble, and is trustworthy and wholesome. The person who scores low on this scale is rebellious, contemptuous of laws and regulations, and neglectful of duty or obligation.

Figure 9.4. The Project A Experimental Predictor Battery (page 3 of 5 pages).

ASSESSMENT OF BACKGROUND AND LIFE EXPERIENCES (ABLE) (Continued)

Scale	Content Definition
Traditional Values	Assesses a person's acceptance of societal values. The person who scores high on this scale accepts and respects authority and the value of discipline. The person who scores low on this scale is unconventional or radical and questions authority and other established norms, beliefs, and values.
Work Orientation	Assesses the tendency to strive for competence in one's work. The work-oriented person works hard, sets high standards, tries to do a good job, endorses the work ethic, and concentrates on and persists in the completion of the task at hand. The less achievement-oriented person has little ego involvement in his or her work, does not expend much effort, and does not feel that hard work is desirable.
Internal Control	Assesses a person's belief in the amount of control people have over rewards and punishments. The person with an internal locus of control believes that there are consequences associated with behavior and that people control what happens to them by what they do. The person with an external locus of control believes that what happens to people is beyond their personal control.
Energy Level	Assesses the amount of energy and enthusiasm a person has. The person high in energy is enthusiastic, active, vital, optimistic, cheerful, zesty, and has the energy to get things done. The person low in energy is lethargic, pessimistic, and tired.
Dominance	Is defined as the tendency to seek and enjoy positions of leadership and influence over others. The highly dominant person is forceful and persuasive when adopting such appropriate behavior. The relatively non-dominant person is less inclined to seek leadership positions and is timid about offering opinions, advice, or direction.
Physical Condition	Measures the frequency and degree of participation in sports, exercise, and physical activity. Individuals high on this scale actively participate in individual and team sports and/or exercise vigorously several times per week. Those low on this scale have participated only minimally in athletics and exercise infrequently.
Unlikely Virtues	Is designed to detect intentional distortion of one's self-description in a favorable direction. High scorers evade answering the ABLE questions frankly and honestly.
Self-Knowledge	Consists of items designed to elicit information about how self-aware and introspective the individual is.
Non-Random Response	Consists of items that have obvious correct and incorrect response options. The correct options are so obvious that a person responding incorrectly is either inattentive to item content or unable to read or understand the items.
Poor Impressions	Measures a variety of negative characteristics. It was developed because of concern that, if the military were to return to a draft, some respondents might distort their self-descriptions in a negative direction to avoid mandatory military service.

ARMY VOCATIONAL INTEREST CAREER EXAMINATION (AVOICE)
(A paper-and-pencil inventory of personal interests)

Construct	Construct Definition
REALISTIC INTERESTS	Is defined as a preference for concrete and tangible activities, characteristics, and tasks. Persons with realistic interests enjoy and are skilled in 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; negatively associated with such occupations as social work and artist. Scales in the AVOICE that measure realistic interests are: Mechanics, Heavy Construction, Electronics, Electronic Communication, Drafting, Law Enforcement, Fire Protection, Audiographics, Rugged Individualism, Firearms Enthusiast, Combat, and Vehicle Operator.

Figure 9.4. The Project A Experimental Predictor Battery (page 4 of 5 pages).

ARMY VOCATIONAL INTEREST CAREER EXAMINATION (AVOICE) (Continued)

Construct	Construct Definition
CONVENTIONAL INTERESTS	Refers to one's degree of preference for well-ordered, systematic, and practical activities and tasks. Persons with conventional interests may be characterized as conforming, unimaginative, efficient, and calm. Conventional interests are associated with occupations such as accountant, clerk, and statistician; negatively associated with occupations such as artist or author. AVOICE scales that measure Conventional interests are: Clerical/Administration, Warehousing/ Shipping, Food Service--Professional, and Food Service--Employee.
SOCIAL AND ENTERPRISING INTERESTS	Are defined as the amount of liking one has for social, helping, and teaching activities as well as persuasive and leadership activities and tasks. The one AVOICE scale that measures both Social and Enterprising interests is Leadership/Guidance.
INVESTIGATIVE INTERESTS	Refers to one's preference for scholarly, intellectual, and scientific activities and tasks. Persons with investigative interest enjoy analytical, ambiguous, and independent tasks, but dislike leadership and persuasive activities. Investigative interests are associated with such occupations as astronomer, biologist, and mathematician; negatively associated with occupations such as salesman or politician. AVOICE scales that measure Investigative interests are Medical Services, Mathematics, Science/Chemical, and Computers.

JOB ORIENTATION BLANK (JOB)
(A paper-and-pencil inventory of job reward preferences)

Construct	Construct Definition
JOB PRIDE	Includes preferences for work environments that are characterized by such positive characteristics as friendly coworkers, fair treatment, and comparable pay. Persons who score high on this scale like the work environment to allow them to feel a sense of accomplishment and to receive recognition for accomplishment.
JOB SECURITY/ COMFORT	Includes preferences for work environments that provide secure and steady employment, where persons receive good training and can utilize their abilities.
SERVING OTHERS	Includes preferences for work environments where persons are reinforced for doing things for other people and for serving others through the work performed.
JOB AUTONOMY	Includes preferences for work environments that reinforce independence and responsibility. Persons who score high on this construct prefer to work alone, try out their own ideas, and decide for themselves how to get the work done.
JOB ROUTINE	Includes preferences for work environments that lack variety, where people do the same or similar things every day, have about the same level of responsibility for quite a while, and follow others' directions.
AMBITION	Measures preferences for work environments that have prestige and status. Persons who score high on this scale prefer work environments that have opportunities for promotion and for supervising or directing others' activities.

Figure 9.4. The Project A Experimental Predictor Battery (page 5 of 5 pages).

ASVAB Factor Composites

1) Quantitative
   - Mathematics Knowledge
   - Arithmetic Reasoning

2) Speed
   - Coding Speed
   - Number Operations

3) Technical
   - Auto/Shop Information
   - Mechanical Comprehension
   - Electronics Information

4) Verbal
   - Word Knowledge
   - Paragraph Comprehension
   - General Science

Paper-and-Pencil Test Composite

1) Spatial
   - Assembling Objects Test
   - Object Rotation Test
   - Maze Test
   - Orientation Test
   - Map Test
   - Reasoning Test

Computer-Administered Test Composites*

1) Movement Time
   - Pooled Movement Time

2) Number Speed and Accuracy
   - Number Memory (Operation DT)
   - Number Memory (PC)

3) Perceptual Accuracy
   - Perceptual Speed & Accuracy (PC)
   - Target Identification (PC)

4) Perceptual Speed
   - Perceptual Speed & Accuracy (DT)
   - Target Identification (DT)

5) Psychomotor
   - Target Tracking 1 Distance
   - Target Tracking 2 Distance
   - Cannon Shoot Time Score
   - Target Shoot Distance

6) Short-Term Memory
   - Short-Term Memory (PC)
   - Short-Term Memory (DT)

7) Basic Speed
   - Simple Reaction Time (DT)
   - Choice Reaction Time (DT)

8) Basic Accuracy
   - Simple Reaction Time (PC)
   - Choice Reaction Time (PC)

ABLE Composites

1) Achievement Orientation
   - Self-Esteem
   - Work Orientation
   - Energy Level

2) Adjustment
   - Emotional Stability

3) Physical Condition
   - Physical Condition

4) Internal Control
   - Internal Control

5) Cooperativeness
   - Cooperativeness

6) Dependability
   - Traditional Values
   - Conscientiousness
   - Nondelinquency

7) Leadership Potential
   - Dominance

AVOICE Composites

1) Administrative
   - Clerical/Administrative
   - Warehousing/Shipping

2) Audiovisual Arts
   - Drafting
   - Audiographics
   - Aesthetics

3) Food Service
   - Food Service - Professional
   - Food Service - Employee

4) Structural/Machines
   - Mechanics
   - Heavy Construction
   - Electronics
   - Vehicle Operator

5) Protective Services
   - Fire Protection
   - Law Enforcement

6) Rugged/Outdoors
   - Combat
   - Rugged Individuals
   - Firearms Enthusiast

7) Interpersonal
   - Medical Services
   - Leadership/Guidance

8) Skilled/Technical
   - Science/Chemical
   - Computers
   - Mathematics
   - Electronic Communication

Note. Scores based on JOB have been omitted.

* DT = Decision Time and PC = Proportion Correct

Figure 9.5. Longitudinal Validation Experimental Battery: 28 composite prediction scores and their constituent basic subtest scores.

PERFORMANCE CRITERION MEASUREMENT

The goals of training performance and job performance measurement in Project A and Career Force were to define the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor. The general procedure for criterion development followed a basic cycle of a comprehensive literature review, comprehensive job analyses using several methods, initial scale/exercise construction, pilot testing, scale/exercise revision, field testing, and proponent (management) review. Some additional specific goals were to:

(1) Make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency.

(2) Compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multitrait, multimethod approach).

(3) Develop rating scale measures of performance factors that are common to all first-tour enlisted MOS (Army-wide measures), as well as for factors that are specific to each MOS.

(4) Develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance.

(5) Evaluate existing archival and administrative records as possible indicators of job performance.

Criterion Development: First Tour

Given these intentions, the criterion development effort for first-tour job incumbents employed three major measurement 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.

The Initial Theory

The development efforts were guided by a model that viewed performance as truly multidimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization.

For the population of entry-level enlisted positions, two major types of job performance components were postulated. The first is composed of components that are specific to a particular job and that would reflect specific technical competence or specific job behaviors which are not required for other jobs. It was anticipated that there

would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills and would be reflected by different task content.

The second type includes components that are defined and measured in the same way for every job. These are referred to as Army-wide performance components and incorporate the basic notion that total performance is much more than task or technical proficiency. In addition to specific skills and abilities that are needed in all jobs, it might include such things as contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity.

In sum, the working model of total performance with which Project A began viewed performance as multidimensional within the two broad categories of factors or constructs. The job analysis and criterion construction methods were designed to explicate the content of these factors via an exhaustive description of the total performance domain, several iterations of data collection, and the use of multiple methods for identifying basic performance factors. As a final analysis step, alternative models of the latent structure were subjected to confirmatory tests.

Saying that performance is multidimensional does not preclude using just one index to make a specific personnel decision (e.g., select/not select, promote/not promote). 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. 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 was to be a matter for research.

Job Analyses

Virtually all subsequent criterion development in Project A/Career Force was based on extensive job analyses of the first-tour Batch A MOS. Task descriptions, critical incident analysis, and interviews with subject matter experts (SMEs) were used extensively. The Soldier Manuals, the Manual of Common Tasks, and the Army Occupational Survey results were used to enumerate the complete population of major tasks for each MOS (approximately 75-150 tasks per MOS). The total array of tasks for each MOS was then grouped into clusters and rated for criticality and difficulty by panels of SMEs.

Additional panels of SMEs were used in a workshop format to generate some 700-800 critical incidents of effective and ineffective performance that were specific to each MOS, and approximately 1,100 critical incidents that could apply to any MOS. For both the MOS-specific and Army-wide critical incidents, a full retranslation procedure was carried out to establish categories, or dimensions, of performance.

Together, the task descriptions and critical incident analysis of MOS-specific and Army-wide performance produced a detailed content description of all the major components of performance in each MOS. These are the job analyses results that were used to begin development of the actual criterion measures.

Criterion Measures for First-Tour Performance

To assess performance on the major components of performance identified by the job analyses, job sample (hands-on) measures and job knowledge tests were developed for a representative sample (as judged by SMEs) of critical job tasks; rating scales were developed for both MOS-specific and Army-wide performance dimensions; and individual personnel records were analyzed extensively to evaluate the usefulness of various archival (administrative) records for performance assessment purposes. The resulting array of Batch A first-tour performance measures is shown as Figure 9.6.

This array of criterion measures produced more than 150 separate scores per individual (including both peer and supervisor ratings), which was too many with which to deal. Consequently, the first criterion analysis objective was to reduce this large number to a smaller and more useful set of what was then termed "basic criterion scores." The analyses were a combination of expert judgment and factor and cluster analyses and they produced the scores shown as Figure 9.7. No scores were discarded. This step focused on combining individual scores into composites such that the loss of specific variance was held to a minimum. It is this set of basic performance criterion scores for each MOS that was used in subsequent modeling analyses.

Development of the LVI Performance Model

A latent factor model (i.e., a particular specification of the number of factors and their substantive content) of first-tour performance, developed using data from the Project A Concurrent Validation (CVI) sample, has been described by Campbell, McHenry, and Wise (1990). This model included the now familiar five performance factors--Core Technical Proficiency (CTP), General Soldiering Proficiency (GSP), Effort and Leadership (ELS), Maintaining Personal Discipline (MPD), and Physical Fitness and Military Bearing (PFB)--and two measurement method factors, a Ratings method factor and a Paper-and-Pencil Test method factor. During Career Force, the CVI model was subjected to a confirmatory analysis, using first-tour performance data collected from the Longitudinal Validation (LVI) sample. Additionally, comparative analyses aimed at evaluating more parsimonious models of first-tour performance were carried out.

To test the fit of the different models to the LVI data, confirmatory factor-analytic techniques were applied to each MOS individually, using LISREL 7 (Jöreskog & Sörbom, 1989). The first alternative tested was the five-factor model developed using CVI data. After the fit of the five-factor model was assessed in each MOS, four reduced models (all nested within the five-factor model) were examined. Finally, as had been done in the original CVI analyses, the five-factor model was applied to the Batch A MOS simultaneously (using LISREL's multigroups option). The fit statistics (e.g., root mean-square residuals [RMSRs]) of the five-factor model for each MOS in the LVI and

1. Ten behaviorally anchored rating scales designed to measure factors of non-job-specific performance.

   Technical Knowledge/Skill
   Peer Leadership and Support
   Demonstrating Effort
   Self-Development
   Maintaining Equipment

   Following Regulations
   Self-Control

   Integrity
   Military Bearing
   Physical Fitness

2. Single scale rating of overall job performance.

3. Single scale rating of NCO (noncommissioned officer) potential.

4. A 19-scale assessment instrument for rating an individual's expected performance in combat.

5. MOS-specific behaviorally anchored rating scales. From 6 to 11 BARS were developed for each MOS to represent the major factors that constitute job-specific technical and task proficiency.

6. Five performance indicators from administrative records. The first three were obtained via self-report and the last two from computerized records.

   - Total number of awards and letters of recommendation.
   - Physical fitness qualification.
   - Number of disciplinary infractions.
   - Rifle marksmanship qualification score.
   - Promotion rate (in deviation units).

7. Job-sample (hands-on) test of MOS-specific task proficiency. The soldier was tested on each of 14-17 major job tasks.

8. Paper-and-pencil job knowledge tests designed to measure both task-specific and common job knowledge. The individual was scored on 150-200 multiple-choice items representing 30 major job tasks sampled representatively from the enumerated population of major job tasks in each MOS. A subsample of 14-17 of the tasks was also measured hands-on.

Figure 9.6. Criterion measures used to assess first-tour performance (Batch A MOS).

Hands-On Performance Test

1. Safety-survival performance score
2. General (common) task performance score
3. Communication performance score
4. Vehicles performance score
5. MOS-specific task performance score

Job Knowledge Test

6. Safety/survival knowledge score
7. General (common) task knowledge score
8. Communication knowledge score
9. Identify targets knowledge score
10. Vehicles knowledge score
11. MOS-specific task knowledge score

Army-Wide Rating Scales

12. Overall effectiveness rating
13. Technical skill and effort factor
14. Personal discipline factor
15. Physical fitness/military bearing factor

MOS-Specific Rating Scales

16. Mean of all MOS-specific scales

Combat Performance Prediction Scales

17. Overall Combat Prediction scale composite (available for males only)

Personnel File Form

18. Awards and Certificates
19. Disciplinary Actions (Articles 15 and Flag Actions)
20. Physical Readiness
21. M16 Qualification
22. Promotion Rate

Figure 9.7. Basic criterion scores derived from first-tour performance measures.

CVI samples were very similar. In fact, for three of the MOS (11B, 13B, and 71L), the RMSRs for the LVI data were smaller than those for the CVI data. These results indicate that the model developed using the CVI data does fit the LVI data quite well.

The four reduced models were also examined using the LVI data. For the four-factor model, the Core Technical Proficiency and General Soldiering Proficiency performance factors were collapsed into a single "can do" performance factor. The three-factor model retained the "can do" performance factor of the four-factor model, but also collapsed the Effort and Leadership and Maintaining Personal Discipline performance factors into a "will do" performance factor. For the two-factor model, the "can do" performance factor was retained; however, the Physical Fitness and Military Bearing performance factor became part of the "will do" performance factor. Finally, for the one-factor model, the "can do" and "will do" performance factors, or equivalently, the five original performance factors, were collapsed into a single performance factor.

The chi-square statistics and RMSRs, respectively, for the four reduced models, as well as for the five-factor model, indicate that the four- and five-factor models fit the LVI data well, while the one-, two-, and three-factor models fit less well. The results also indicated that the parameter estimates for the five-factor model were generally similar across the 10 Batch A MOS.

The final step was to determine whether the variation in some of these parameters could be attributed to sampling variation. To do this, the following were specified to be invariant across jobs: (a) the correlations among performance factors, (b) the loadings of all the Army-wide measures on the performance factors and on the rating method factor, (c) the loadings of the MOS-specific score on the rating method factor, and (d) the uniqueness coefficients for the Army-wide measures.

The results indicated that the fit of the five-factor model is not as good when the parameters listed above are constrained to be equal across the 10 jobs. Still, the root mean-square residuals associated with the across-MOS model were not substantially greater than those for the within-job analyses. (The average RMSR for the across-MOS model is .0676; the average for the within-MOS models is .0585.)

To create criterion construct scores for use in validation analyses, the scoring procedures were based on the five-factor model. Although the four-factor model has the advantage of greater parsimony, the five-factor model offered the advantage of corresponding to the criterion constructs generated in the CVI validation analyses. It is shown as Figure 9.8. The actual factor scores were obtained simply by standardizing the components and taking the simple sum of the components.

Criterion Development: Second Tour

The basic conceptual foundation for criterion development in the second tour was the same as it was for the first tour. As would be expected, the second-tour job analyses, measurement development, and criterion data analyses were informed by the first-tour criterion analysis.

1) Core Technical Proficiency (CTP)
   - Hands-on Test Score - MOS-Specific Tasks
   - Job Knowledge Test Score - MOS-Specific Tasks

2) General Soldiering Proficiency (GSP)
   - Hands-on Test Score - Common Tasks
   - Job Knowledge Test Score - Common Tasks

3) Effort and Leadership (ELS)
   - Administrative Index - Number of Awards and Certificates
   - Army-Wide BARS Overall Effectiveness Rating Scale
   - Army-Wide BARS Effort/Leadership Ratings Factor
   - Average of MOS BARS Ratings Scales

4) Maintaining Personal Discipline (MPD)
   - Administrative Index - Number of Articles 15 and Flag Actions
   - Administrative Index - Promotion Grade Deviation Score
   - Army-Wide BARS Personal Discipline Ratings Factor

5) Physical Fitness and Military Bearing (PFB)
   - Administrative Index - Physical Readiness Score
   - Army-Wide BARS Fitness/Bearing Ratings Factor

Figure 9.8. Five LVI first-tour performance factor scores and the basic criterion scores that define them as obtained from the first-tour performance measures.

Job Analyses

The job analysis goals for the second tour included the description of the major differences in technical task content between first and second tour and the description of the leadership/supervision component of the junior NCO position. The task analysis and critical incident analysis steps used for first tour were also used for second tour. In addition, a special 46-item job analysis instrument, the Supervisory Description Questionnaire, was constructed and used to collect similarity and criticality judgments from SMEs. Consequently, the supervisory/leadership tasks judged to be critical for an MOS became part of the population of tasks for that MOS.

Rather than using the critical incident and task description data to start from scratch, the project staff used the second-tour job analyses information to modify and expand the first-tour job descriptions. In general, the technical components of performance retained the same type of content but were at a higher level of complexity and required greater expertise. Also, both the task descriptions and the critical incident descriptions yielded additional components of leadership and supervisory performance.

Criterion Measures for Second-Tour Performance

First-tour measures were revised for use with second-tour personnel and new measures reflecting the unique components of second-tour jobs were added. A summary description of the specific measures follows.

Rating Scales. On the basis of second-tour critical incident analyses, the Army-wide BARS and MOS-specific BARS were revised and scales having to do with leadership and supervision were added. Further, based on job analysis data, seven new scales pertaining to supervision and leadership responsibilities were also added. A full list of the Army-wide rating scales is shown below. Not shown are the MOS BARS for each MOS, which were revised to reflect second-tour performance demands, and the Combat Performance Prediction Scales, which were the same as those used in LVI, and which were not administered to female NCOs.

Army-Wide Behavior Scales:

1. Demonstrating Technical Knowledge and Skill
2. Demonstrating Effort
3. Supervising Subordinates
4. Following Regulations and Orders
5. Demonstrating Integrity
6. Training and Development of Subordinates
7. Maintaining Equipment
8. Physical Fitness
9. Self-Development
10. Showing Consideration for Subordinates
11. Demonstrating Appropriate Military Bearing
12. Demonstrating Appropriate Self-Control

Additional Leadership Scales:

13. Serving as a Role Model
14. Communication With Subordinates
15. Personal Counseling
16. Monitoring Subordinate Performance
17. Organizing Missions/Operations
18. Personnel Administration
19. Performance Counseling

General Scales:

20. Overall Effectiveness
21. Senior NCO Potential

Situational Judgment Test (SJT). A new paper-and-pencil measure of supervisory judgment was developed by describing prototypical judgment situations and asking the respondent to select the most appropriate and the least appropriate courses of action. The situation descriptions and the scoring keys were refined through extensive SME judgments.

Supervisory Simulation Exercises. These measures were developed to assess NCO performance in job areas that were judged to be best assessed through the use of interactive exercises. The simulations were designed to evaluate performance in counseling and training subordinates. A trained evaluator (role player) played the part of a subordinate to be counseled or trained and the examinee assumed the role of a first-line supervisor who was to conduct the counseling or training. Evaluators also scored the examinee's performance, using a standard set of rating scales.

- Personal Counseling Simulation: A private first class (PFC) is exhibiting declining job performance and personal appearance. Recently, the PFC's wall locker was left unsecured. The supervisor has decided to counsel the PFC about these matters.

- Disciplinary Counseling Simulation: There is convincing evidence that the PFC lied to get out of coming to work today. The PFC has arrived late to work on several occasions and has been counseled for lying in the past. The PFC has been instructed to report to the supervisor's office immediately.

- Training Simulation: The commander will be observing the unit practice formation in 30 minutes. The private, although highly motivated, is experiencing problems with the hand salute and about face.

For each exercise, examinee performance was evaluated on rating scales reflecting specific behaviors tapped by the exercises. Factor analyses of the ratings data were used to develop basic scores for these measures.

Administrative Measures. The self-report Personnel File Form (PFF) used in LVI was modified for use with second tour and six administrative indices of performance were obtained.

Job Knowledge and Hands-On Measures. The content of each of these measures was revised on the basis of the second-tour job analyses and the revised instruments were subjected to extensive SME review. Analyses of alternative aggregations of item and scale scores from both of these measures resulted in the adoption of a general (Army-wide) and an MOS-specific score for each of them.

Basic (LVII) Criterion Scores. The problem of reducing the total number of second-tour criterion scores available for each individual to a more manageable number was the same as it was for first tour, and was addressed in much the same way. The array of basic second-tour criterion scores that were used to develop a model of performance factors for the second-tour junior NCO is shown in Figure 9.9.

Hands-On Performance Test

1. MOS-specific task performance score
2. General (common) task performance score

Job Knowledge Test

3. MOS-specific task knowledge score
4. General (common) task knowledge score

Army-Wide Rating Scales

5. Overall effectiveness rating
6. Leadership/supervision factor
7. Technical skill and effort factor
8. Personal discipline factor
9. Physical fitness/military bearing factor

MOS-Specific Rating Scales

10. Overall MOS composite

Combat Performance Prediction Scales

11. Overall Combat Prediction scale composite

Personnel File Form

12. Awards and certificates
13. Disciplinary actions (Articles 15 and Flag actions)
14. Physical readiness
15. Promotion rate

Situational Judgment Test

16. Total test score or, alternatively, the following subscores:
    a. Discipline soldiers when necessary
    b. Focus on the positive
    c. Search for underlying causes
    d. Work within chain of command
    e. Show support/concern for subordinates
    f. Take immediate/direct action

Supervisory Simulation Exercises: Subscores via Factor Analysis

17. Personal counseling - Communication/Interpersonal skill
18. Personal counseling - Diagnosis/Prescription
19. Disciplinary counseling - Structure
20. Disciplinary counseling - Interpersonal skill
21. Disciplinary counseling - Communication
22. Training - Structure
23. Training - Motivation maintenance

Figure 9.9. Basic criterion scores generated from second-tour performance measures (LVII).

Development of the LVII Performance Model

The specific objective was to determine which model, from among several proposed alternative models of the latent structure of basic criterion score intercorrelations, best fit the observed data for LVII. Analyses were guided by the same general framework that was used in modeling the covariation among performance measures for first-tour performance.

One alternative was the model developed based on data from the Project A Concurrent Validation second-tour (CVII) sample. This model, referred to as the Training and Counseling model, is described in detail in Campbell and Oppler (1990).

Several additional alternative models of second-tour soldier performance were hypothesized by the project staff on strictly conceptual grounds. The fit of these alternative models was then assessed using the LVII data and compared with the fit of the CVII Training and Counseling model. Second, because the Combat Performance Prediction Scales were not included in this initial modeling, key analyses were rerun with these scales included to confirm that the Combat scales fit the models as expected and to determine whether including them would affect the degree of fit. Once a best fitting model was identified, subsequent analyses were conducted to determine whether the model fit equally well across MOS and across demographic subgroups. Finally, based on the results of these analyses, a set of criterion construct scores to be used in the LVII validation analyses was specified.

To generate alternative hypotheses for the latent structure, definitions of the LVII basic criterion scores used in the modeling exercise were circulated to the project staff, and a variety of hypotheses concerning the nature of the underlying structure of second-tour soldier performance were obtained. These hypotheses were consolidated into one principal central alternative model, plus several variations on this model, and a series of more parsimonious models that involved collapsing two or more of the substantive factors.

The central alternative, the Consideration/Initiating Structure model, differed from the CVII Training and Counseling model primarily in that it includes two leadership factors. Based on staff judgment, the leadership rating scales and each of the Supervisory Judgment Test (SJT) and supervisory simulation scores were assigned to one of these two factors.

Because the within-MOS sample sizes in the LVII sample were relatively small (ranging from 69 to 281), initial tests of the models were conducted using the entire LVII sample. Criterion scores were first standardized within each MOS, then the intercorrelations among these standardized basic scores were computed across all MOS. The total sample matrix was used as input for the first set of analyses.

The analysis plan was to first compare the fit of the Consideration/Initiating Structure model with the variations of this model and with the Training and Counseling model to identify the alternatives that best fit the LVII covariance structure. The next

set of analyses involved comparing a series of nested models to determine the extent to which the observed correlations could be accounted for by fewer underlying factors. LISREL 7 was used to estimate the parameters and evaluate the fit of each of the alternative models.

The fit of the CVII Training and Counseling model in the LVII sample was remarkably similar to the fit of this same model in the original CVII sample, especially considering that the performance data were collected several years apart using somewhat different measures. Tests of the newly proposed Consideration/Initiating Structure model and its variations resulted in a relatively poor fit to the data.

To determine whether there were other reasonable alternative models of second-tour soldier performance, the LVII total sample was randomly divided into two subsamples: 60 percent for model development and 40 percent for cross-validation/ confirmation.

The matrix of intercorrelations among the basic criterion scores for the developmental subsample was examined by project staff and several alternative models were suggested. A number of alternatives tried different arrangements of the supervisory simulation, SJT, and rating scale basic scores, while still preserving two leadership factors. None of the these alternatives resulted in a good fit. However, a model that collapsed the Consideration and Initiating Structure factors into a single Leadership factor, included a single Simulation Exercise method factor, and moved the promotion rate variable to the new Leadership factor did result in a considerably better fit to the data in both the developmental and the holdout samples.

The "Leadership Factor" model that was developed based on these exploratory analyses is shown in Figure 9.10. The fit of the new Leadership Factor model to the LVII data is, for all practical purposes, identical to the fit of the Training and Counseling model to these same data. The 90 percent confidence intervals for the RMSEAs overlap almost completely.

Because these models have an equally good fit to the data and because the Leadership Factor model does not confound method variance with substantive variance, the Leadership Factor model was chosen as the best representation of the latent structure of second-tour performance.

The Leadership Factor model was tested again with the Combat Performance Prediction Scales included. For one comparison, the Combat Prediction Score was constrained to load only on the Leadership factor and the Rating Method factor. For the second, the Combat Prediction score was constrained to load on the Achievement and Effort and the Rating Method factors only. The assignment of the Combat Prediction Score to the Achievement and Effort factor produced a much better fit.

Latent Variable	Scores Loading on Latent Variables
Core Technical Proficiency (CT)	MOS-Specific Hands-On MOS-Specific Job Knowledge
General Soldiering Proficiency (GP)	General Hands-On General Job Knowledge
Achievement and Effort (AE)	Awards and Certificates Army-Wide Ratings: Technical Skill/Effort Composite Overall Effectiveness Rating MOS Ratings: Overall Composite Combat Prediction: Overall Composite
Personal Discipline (PD)	Disciplinary Actions (reversed) Army-Wide Ratings: Personal Discipline Composite
Physical Fitness/Military Bearing (PF)	Physical Readiness Score Army-Wide Ratings: Physical Fitness/Bearing Composite
Leadership (LD)	Promotion Rate Army-Wide Ratings: Leading/Supervising Composite SE - Disciplinary Structure SE - Disciplinary Communication SE - Disciplinary Interpersonal Skill SE - Counseling Diagnosis/Prescription SE - Counseling Communication/Interpersonal Skills SE - Training Structure SE - Training Motivation Maintenance SJT - Total Score
Written Method	Job-Specific Knowledge General Job Knowledge SJT - Total Score
Ratings Method	Four Army-Wide Ratings Composites Overall Effectiveness Rating MOS Ratings: Total Composite Combat Prediction: Overall Composite
Simulation Exercise Method	All Seven Simulation Exercise Scores

Figure 9.10. Leadership Factor Model.

Nested Models. The Leadership Factor model was used as the starting point to develop a series of more parsimonious nested models, similar to those tested in the LVI sample by Oppler, Childs, and Peterson (1994). The first was identical to the full Leadership Factor model except that the Achievement and Effort factor was collapsed with the Leadership factor.

Similarly, the second nested model was identical to the model just described except that, in addition, the Core Technical and General Soldiering Proficiency factors were replaced with a single Can-Do factor. Third, the Personal Discipline factor and the new Achievement/Leadership factor were also collapsed. The fourth model involved adding the variables from the Physical Fitness factor to this Achievement/Leadership/Personal Discipline factor, resulting in a single Will-Do factor. The final model collapsed all of the substantive factors into a single overall performance factor.

Because these more parsimonious models are nested within each other, the significance of the loss of fit could be tested by comparing the chi-square values for the various models. In the first nested model, which involved collapsing the Leadership factor with the Achievement and Effort factor, the resulting decrement in fit was very small. Similarly, collapsing the two can do factors resulted in a very small reduction in model fit. Based on these results, a model with only four substantive factors (and three method factors) can account for the data almost as well as the full Leadership Factor model. Collapsing additional factors beyond this level resulted in larger decrements in model fit.

Training Performance Measurement

The performance measures that were collected at the end of the basic/technical training period consisted of the revised CVI-developed training knowledge test and seven Army-wide behavioral summary rating scales that were developed to parallel the analogous rating scales used to assess first-tour performance. The Training knowledge test contains items that were MOS content specific and items that did not have a specific MOS referrent.

The training performance criterion factor scores that were derived from these measures are shown in Figure 9.11. In the case of the rating scales, the initial factors were specified a priori on the basis of their correspondence with the CVI/LVI factors. No other arrangement yielded a better fit to the LVT intercorrelation matrix.

The Correlation of Performance With Performance

The Project A/Career Force database presents what may be the only existing opportunity in the history of personnel research to estimate the accuracy of the prediction of future performance from current performance across three organizational levels when a consistent approach to job analysis and performance measurement was used at each level and the objective was to assess performance comprehensively and model its latent structure. It invites an analysis of the factor by factor cross-correlations for the EOT, LVI, and LVII performance factor scores. The fact that the design is truly longitudinal and multiple factors of performance are assessed for a representative sample of jobs makes it a very powerful analysis.

Analytic Approach

Within this context, the analysis of the performance x performance correlation addressed three specific questions:

(1) To what degree does performance in training predict subsequent job performance?

(2) To what degree does an individual's level of performance in a first-tour enlisted position predict performance as a junior NCO during his or her second tour?

SCORES BASED ON TRAINING ACHIEVEMENT TESTS

1) Basic Knowledge Score

- Items measuring knowledge requirements common to all MOS.

2) Technical Knowledge Score

- Items measuring technical knowledge requirements specific to each MOS.

SCORES BASED ON RATING SCALES (PEER RATINGS)*

3) Effort and Technical Skill (ETS) Score

- Degree of effective acquisition of technical knowledge/skill
- Degree to which individual demonstrates extra effort

4) Leadership Potential (LEAD)

- Degree of expected leadership effectiveness

5) Maintaining Personal Discipline (MPD) Score

- Degree to which individual adheres to regulations and orders
- Degree to which the individual practices effective self-control

6) Physical Fitness and Military Bearing (PFB)

- Degree to which individual maintains proper military appearance
- Degree to which individual maintains military standards of physical fitness

* Each of the bullets describes a behavioral summary rating scale (i.e., seven individual scales) constructed to parallel the analogous scale developed to assess first-tour performance (i.e., parallel to CVI and LVI).

Figure 9.11. Six EOT performance factor scores based on measures of training performance at the end of basic and technical training.

(3) Given that performance is not unidimensional, do the separately measured components of performance exhibit the appropriate patterns of convergent and divergent validity when current performance is used to predict future performance?

The total sample for each MOS at each measurement point is shown in Table 9.2. These are the maximum possible sample sizes. Because of missing data considerations, the numbers for any specific analysis were smaller.

Table 9.2
Sample Sizes From Each Batch A MOS When Performance was Assessed at Three Points in Time for the Project A/Career Force LV Sample

MOS		Sample Size Before Data Editing		
		End-of-Training (EOT) Performance	First-Tour (LVI) Job Performance	Second-Tour (LVII) Job Performance
11B	Infantryman	8,117	909	47
13B	Cannon Crewmember	4,712	916	80
19E/K	Armor Crewman	2,048	1,073	168
31C	Single Channel Radio Operator	667	529	60
63B	Light-Wheel Vehicle Mechanic	1,215	752	194
71L	Administrative Specialist	1,414	678	157
88M	Motor Transport Operator	1,354	682	89
91A/B	Medical Specialist	3,218	824	222
95B	Military Police	3,639	452	168
	Total	26,384	6,815	1,595

If the latent structure of performance is consistent across replications within organizational level and differs as expected (given the limitations of measurement) across levels (i.e., first tour vs. second tour), then it is reasonable to expect that the observed correlations of performance with performance, or performance factors with performance factors, would show the expected convergent and divergent relationships. That is, a particular performance factor measured at time one should have a higher correlation with itself at time two than it does with other performance factors at time two.

To examine these patterns of correlations for the Project A/Career Force Project database, a number of basic intercorrelation matrices were computed. Two are summarized here:

(1) End-of-Training Performance vs. First-Tour Performance (a 6 x 5 matrix).

(2) First-Tour Performance vs. Second-Tour Performance (a 5 x 6 matrix).

Each matrix was calculated by computing the intercorrelations within each MOS and then averaging over MOS. All correlations are corrected for restriction of range by using a multivariate correction that treated the six EOT performance factors as the "implicit" selection variables. It was felt that, in comparison to other incidental selection variables, these factors would have the most to do with whether an individual advanced in the organization.

Making the corrections in this way means that the referent population consists of all the soldiers in the LV sample who had completed their training course. This is the population for which we would like to estimate the prediction of performance with performance. It is also the population for which the comparison of the validities of the experimental predictor tests and training criteria as predictors of future performance is the most meaningful. So long as the implicit selection variables are the best available approximation to the explicit selection variables, the corrected coefficients will be a better estimate of the population values than the uncorrected coefficients, but they will still be an underestimate (Linn, 1968). Since the degree of range restriction from EOT to LVI is not very great, the effects of the corrections were not very large.

Results

The correlations of training performance with first-tour performance are shown in Table 9.3; the correlations of first-tour performance with second-tour performance are shown in Table 9.4.

In general, correlations of performance with performance are substantial. Performance in training does predict performance as a first-tour job incumbent, and performance in the first tour of duty does predict performance in the second-tour, after reenlistment.

There is also a reasonable pattern of convergent and divergent validity across performance factors, even without correcting these coefficients for attenuation and thereby controlling for the effects of differential reliability. The greatest departure from the expected pattern is found in the differential correlations of the two Can-Do test-based factors (i.e., CTP and GSP). The correlation patterns for the Will-Do factors, which are based largely on ratings, virtually never violate the expected pattern, even when peer ratings during training are being correlated with supervisory ratings obtained during the second tour.

The one possible exception to this pattern is the predictability of the leadership performance factor for second-tour personnel. This component of NCO performance is predicted by almost all components of past performance. However, this result is quite consistent with saying that effective leadership is a function of multiple determinants. A similar result occurs in the validation analysis when the question is what trait constructs predict leadership performance. Consequently, the selection and training of effective leaders must take the full set of determinants into account. Neglecting one or more critical sources of variance would be counterproductive.

In sum, the success of the Project A/Career Force efforts to conceptualize and measure soldiers went far beyond expectations. We learned much more about the measurement and prediction of performance than anyone ever anticipated.

Table 9.3
Zero-Order Correlations of Training Performance (EOT) Variables With First-Tour Job Performance (LVI) Variables: Weighted Average Across MOS

	EOT Variables									
LVI Variables	EOT:TECH	EOT:BASE	EOT:ETS	EOT:MPD	EOT:PFB	EOT:LEAD	EOT:ELS	EOT:CAN	EOT:WILL	EOT:TOT
Core Technical Proficiency (CTP)	**.482** 3857	.380 3582	.217 3843	.153 3843	.049 3843	.180 3843	.208 3843	.475 3582	.180 3843	.485 3535
General Soldiering Proficiency (GSP)	.493 3857	**.452** 3582	.230 3843	.171 3843	.043 3843	.162 3843	.203 3843	.526 3582	.181 3843	.534 3535
Effort and Leadership (ELS)	.209 3795	.167 3525	.354 3783	.250 3783	.277 3783	.353 3783	**.376** 3783	.208 3525	.365 3783	.251 3479
Maintain Personal Discipline (MPD)	.174 3908	.136 3633	.310 3894	**.355** 3894	.214 3894	.272 3894	.307 3894	.170 3633	.340 3894	.211 3586
Physical Fitness and Bearing (PFB)	-.011 3908	-.016 3633	.262 3894	.127 3894	**.444** 3894	.308 3894	.307 3894	-.015 3633	.330 3894	.031 3586
"Can Do" Performance Composite (CAN)	.530 3857	.451 3582	.245 3843	.177 3843	.050 3843	.187 3843	.226 3843	**.545** 3582	.197 3843	.555 3535
"Will Do" Performance Composite (WILL)	.167 3795	.128 3525	.386 3783	.302 3783	.373 3783	.389 3783	.413 3783	.163 3525	**.427** 3783	.216 3479
Total Performance Composite (TOT)	.388 3741	.322 3471	.407 3729	.314 3729	.298 3729	.379 3729	.416 3729	.394 3471	.413 3729	**.438** 3425
NCO Potential Rating <Supv> (NCO)	.165 3458	.140 3458	.309 3444	.224 3444	.259 3444	.306 3444	.327 3444	.169 3458	.324 3444	.208 3397

Note. Corrected for range restriction. Pairwise Ns are printed below each correlation. Correlations between matching variables are in bold.

VALIDATION RESULTS

At this point we have summarized the Project A/Career Force research that led to the development of the Experimental Predictor Battery and to the development of criterion measures for training performance, first-tour job performance, and performance as a junior NCO during the second tour of duty. At each of these three stages in a soldier's career, performance is represented by the scores on the set of performance factors that were judged to "fit the data best "in both a theoretical and a confirmatory sense. The latent structure of performance across these three organizational levels was consistent where it should be and different where it should be, given the results of the job analyses. The empirical correlations of the performance factors with later performance also exhibit a surprising degree of the appropriate convergent and divergent patterns.

The next major step was to assess the predictive validities of the ASVAB and the Experimental Battery for predicting performance during training, during the first tour, and during the second tour, after reenlistment. As stipulated by the original objectives for the projects, this was an attempt to validate a comprehensive battery of new predictor

Table 9.4
Zero-Order Correlations of First-Tour Job Performance (LVI) Variables With Second-Tour Job Performance (LVII) Variables: Weighted Average Across MOS

	LVI Variables								
LVII Variables	LVI:CTP	LVI:GSP	LVI:ELS	LVI:MPD	LVI:PFB	LVI:CAN	LVI:WILL	LVI:TOT	LVI:NCO
Core Technical Proficiency (CTP)	**.440** 412	.413 412	.249 400	.078 413	.015 412	.449 412	.181 400	.375 397	.230 379
General Soldiering Proficiency (GSP)	.511 412	**.569** 412	.219 400	.085 413	-.008 412	.578 412	.157 400	.440 397	.220 379
Achievement and Effort (AE)	.103 390	.167 390	**.450** 377	.280 390	.319 390	.150 390	.464 377	.470 374	.412 353
Leadership (LEAD)	.359 344	.411 344	**.379** 333	.272 343	.169 342	.421 344	.365 333	.517 332	.378 319
Leadership Minus SJT Score	.264 348	.310 348	**.372** 337	.249 347	.233 346	.321 348	.380 337	.467 336	.398 322
Achievement, Effort and Leadership	.275 333	.335 333	**.471** 322	.292 332	.264 331	.336 333	.459 322	.620 321	.444 307
Maintain Personal Discipline (MPD)	-.044 406	.038 406	.116 393	**.257** 406	.166 406	-.002 406	.211 393	.164 390	.114 370
Physical Fitness and Bearing (PFB)	-.026 392	-.013 392	.220 379	.135 392	**.460** 392	-.022 392	.333 379	.250 376	.265 356
"Can Do" Performance Composite (CAN)	.520 412	.533 412	.259 400	.097 413	.010 412	**.562** 412	.193 400	.452 397	.260 379
"Will Do" Performance Composite (WILL)	.141 321	.190 321	.370 310	.295 320	.347 319	.182 321	**.433** 310	.445 309	.404 296
Total Performance Composite (TOT)	.336 313	.357 313	.381 302	.240 312	.252 311	.375 313	.394 302	**.521** 301	.423 289

Note. Corrected for range restriction. Pairwise Ns are printed below each correlation. Correlations between matching variables are in bold.

information on a representative sample of jobs from the entire personnel system, using incumbents sampled from three different career stages.

As a general summary of the full Project A/Career Force validation results, four types of validation data are presented below.

(1) Previous project publications have used the term "basic validation analysis" to describe the correlations of each predictor domain (i.e., ASVAB, Spatial, Computerized Perceptual and Psychomotor, ABLE, AVOICE, and JOB) with each performance criterion factor. In these "basic" analyses the individual predictors within each domain are regression weights, except for the tests that comprise the paper-and-pencil spatial battery, which have always been unit weighted. The basic validity data for EOT, LVI, and LVII were compared.

(2) Incremental validities were examined by comparing the multiple correlations of the four ASVAB factors plus a particular criterion with the multiple correlation for ASVAB plus each of the other predictor domains in turn. The incremental validities also were compared for EOT, LVI, and LVII.

(3) For the longitudinal prediction of first-tour performance (LVI), three kinds of "optimal" prediction equations were compared in terms of the maximum level of predictive validity that could be achieved.

(4) For the prediction of performance in the second tour (LVII), the validities of alternative prediction equations using different combinations of test data and previous performance data were compared.

The specific MOS x MOS sample sizes for each specific prior validation will not be given here. They are contained in the earlier chapters of this report and in the previous four annual reports that constitute the primary documentation for the Career Force Project. The EOT, LVI, and LVII sample information that is appropriate will be summarized for each of the tables of results given below.

Basic Validation and Incremental Validity Comparisons for the Prediction of Training Performance (EOT), First-Tour Job Performance (LVI), and Second-Tour Job Performance (LVII)

The comparisons of basic validities and incremental validities for EOT, LVI, and LVII used modified versions of setwise deletion to constitute the samples for each estimate. That is, to be included in a particular sample, an individual was required to have complete data on the criterion variable being predicted, the ASVAB, and the Experimental Battery predictor set being evaluated. The general nature of each sample is described in Figure 9.12.

Analytic Approach

For both the basic validities and the incremental validities, the general analytic steps for computing the estimates for each criterion were the following:

- Within each MOS, correct the covariance matrix for restriction of range, using the multivariate correction described by Lord and Novick (1968).
- Compute the designated multiple correlations within each MOS.
- Adjust each multiple R for shrinkage (Rozeboom, 1978, formula 8)
- Compute the unweighted mean across MOS.

- End-of Training (EOT)
  - 11 "Batch A" MOS included in validation
  - Single sample used for all analyses. Trainees were required to have complete predictor and criterion data (N = 4,039)
- Longitudinal Validation First-Tour (LVI)
  - 9 "Batch A" MOS included in validation
  - Separate analysis sample used for each predictor set. Soldiers in each analysis sample were required to have complete criterion data, complete ASVAB data, and complete data for the EB predictor set being evaluated (Ns range from 3,797 to 4,450)
- Longitudinal Validation Second-Tour (LVII)
  - 7 "Batch A" MOS included in validation
  - Separate analysis sample used for each predictor set/criterion combination. Soldiers in each analysis sample were required only to have complete ASVAB data and complete data for the EB predictor set being evaluated (Ns range from 810 to 1,224)

Figure 9.12. Description of samples for EOT, LVI, and LVII comparisons of basic validities and incremental validities.

- Correct the mean adjusted R for attenuation due to criterion unreliability by using the median reliability across MOS for each criterion.

Results

The results for the basic validity estimates and the incremental validity estimates are shown in Tables 9.5 and 9.6, respectively. These tables compress a great deal of validation analysis into a small space. In an attempt to be equally succinct, the following summary statements are offered.

- With only a few exceptions, the validities of the different predictor domains for particular performance factors are quite consistent from EOT to LVI to LVII. ASVAB predicts the Can-Do performance factors (CTP and GSP) in LVII and LVI just as accurately as it does at

Table 9.5
Average Multiple Correlations Computed Within Job for EOT, LVI, and LVII Validation Samples for ASVAB Factors, Spatial, Computer, JOB, ABLE, and AVOICE

Sample/ Criterion	No. of MOS[a]	ASVAB Factors [4]	Spatial [1]	Computer [8]	JOB [3]	ABLE [7]	AVOICE [8]
EOT							
SK-Tech	11	76 (80)	63 (66)	61 (64)	41 (43)	33 (35)	44 (46)
SK-Basic	9	68 (76)	57 (64)	57 (64)	38 (42)	30 (34)	37 (41)
ETS	11	41 (50)	35 (43)	36 (44)	24 (30)	19 (23)	22 (27)
LEAD	11	30 (38)	24 (30)	28 (35)	18 (23)	22 (28)	17 (21)
MPD	11	25 (30)	22 (26)	21 (25)	09 (11)	19 (23)	11 (13)
PFB	11	14 (17)	05 (06)	11 (13)	05 (06)	29 (35)	07 (08)
LVI							
CTP	9	63 (70)	58 (64)	49 (54)	31 (34)	21 (23)	39 (43)
GSP	8	66 (75)	65 (74)	55 (63)	32 (36)	24 (27)	38 (43)
ELS	9	37 (40)	33 (36)	30 (33)	12 (21)	13 (14)	20 (22)
MPD	9	16 (18)	14 (16)	10 (11)	06 (07)	15 (17)	05 (06)
PFB	9	16 (18)	08 (09)	13 (14)	07 (08)	28 (31)	09 (10)
LVII							
CTP	7	64 (75)	57 (67)	53 (62)	33 (39)	24 (28)	41 (48)
GSP	6	63 (74)	58 (68)	48 (56)	28 (33)	19 (22)	29 (34)
EA	7	29 (31)	27 (29)	09 (10)	07 (08)	13 (14)	09 (10)
LEAD	7	63 (68)	55 (59)	49 (53)	34 (37)	34 (37)	35 (38)
MPD	7	15 (17)	15 (17)	12 (13)	03 (03)	06 (07)	06 (07)
PFB	7	16 (18)	13 (14)	03 (03)	07 (08)	17 (19)	09 (10)

Note: All results corrected for multivariate range restriction (Lord & Novick, 1968) and adjusted for shrinkage (Rozeboom formula 8, 1978). Results in parentheses disattenuated for criterion unreliability. Numbers in brackets are the numbers of predictor scores entering prediction equations. Decimals omitted.

[a] Number of MOS for which validities were computed.

Table 9.6
Average Increments in Multiple Correlations Over ASVAB, Computed Within Job for EOT, LVI, and LVII Validation Samples for Spatial, Computer, JOB, ABLE, and AVOICE

Sample/ Criterion	No. of MOS[a]	Spatial [4 + 1]	Computer [4 + 8]	JOB [4 + 3]	ABLE [4 + 7]	AVOICE [4 + 8]
EOT						
SK-Tech	11	.01 (.01)	.01 (.01)	-- (--)	-- (--)	-- (--)
SK-Basic	9	.01 (.01)	-- (--)	-- (--)	-- (--)	-- (--)
ETS	11	.01 (.02)	.01 (.02)	-- (--)	.03 (.04)	-- (--)
LEAD	11	-- (--)	.01 (.01)	-- (--)	.05 (.06)	-- (--)
MPD	11	-- (--)	-- (--)	-- (--)	.09 (.11)	-- (--)
PFB	11	-- (--)	.03 (.03)	.01 (.01)	.17 (.20)	.01 (.01)
LVI						
CTP	9	.01 (.01)	-- (--)	-- (--)	-- (--)	-- (--)
GSP	8	.03 (.04)	.01 (.01)	-- (--)	-- (--)	-- (--)
ELS	9	-- (--)	-- (--)	-- (--)	-- (--)	-- (--)
MPD	9	-- (--)	-- (--)	-- (--)	.08 (.09)	-- (--)
PFB	9	-- (--)	-- (--)	.01 (.01)	.17 (.18)	-- (--)
LVII						
CTP	7	-- (--)	-- (--)	-- (--)	-- (--)	-- (--)
GSP	6	.01 (.02)	-- (--)	-- (--)	-- (--)	-- (--)
EA	7	.02 (.03)	-- (--)	-- (--)	-- (--)	-- (--)
LEAD	7	-- (--)	-- (--)	-- (--)	.01 (.01)	-- (--)
MPD	7	-- (--)	-- (--)	-- (--)	-- (--)	-- (--)
PFB	7	-- (--)	-- (--)	-- (--)	.05 (.06)	-- (--)

Note: All results corrected for multivariate range restriction (Lord & Novick, 1968) and adjusted for shrinkage (Rozeboom formula 8, 1978). Results in parentheses disattenuated for criterion unreliability. Numbers in brackets are the numbers of predictor scores entering prediction equations.

[a] Number of MOS for which validities were computed.

the end of training. Their consistency is especially noteworthy given that the LVI and LVII factors use both the knowledge test and job sample measures to assess the constructs while the training score is based only on the knowledge tests.

- ASVAB also predicts the leadership component of performance with considerable validity. As should be the case, the relationship is higher for the prediction of realized leadership performance (LVII) than it is for leadership potential, which was assessed in EOT and LVI. The LVI leadership factor is predicted by every predictor domain. It was noted earlier that this factor was also predicted by prior performance on virtually all components of first-tour performance. It is tempting to infer that good leadership is a function of a number of distinct determinants, to a greater extent than are other components of performance.

- Vocational interests predict technical performance, and to the extent that technical performance contributes to leadership performance, interests predict leadership as well.

- The validity of ABLE goes up between LVI and LVII for predicting the leadership-related factors but goes down for MPD and PFB. The latter result seems to have occurred because the variance in the two factors is reduced in LVII. Also, the meaning of MPD for second-tour personnel may be different than for first-tour personnel.

- ABLE provided the greatest incremental validity for predicting Will-Do performance; however, the degree of increment fell sharply between LVI and LVII.

- The spatial composite generally provided a small amount of incremental validity (approximately one point) for predicting Can-Do performance.

Validity Estimates for Optimal Equations

The objectives for this part of the analysis were to estimate the predictive validity, in the LVI sample, (a) of the full ASVAB + Experimental Battery predictor set (k = 28) for 12 unique equations, and (b) of the same set of 12 unique equations after they were reduced in length with the goal of preserving either maximum selection efficiency or maximum classification efficiency. The 12 equations corresponded to a unique equation in each of the nine Batch A MOS for predicting CTP and a pooled (across MOS) equation for predicting each of the three Will-Do factors (i.e., ELS, MPD, and PFB).

Analytic Approach

The Full Prediction Equations. For each of the 12 unique prediction situations identified in the previous analysis (i.e., one equation for CTP in each MOS and one equation for all MOS for ELS, MPD, and PFB), the appropriate covariance matrices,

corrected for range restriction, were used to compute full least-squares estimates of the multiple correlation between the full predictor battery (ASVAB plus EB) and the relevant criterion score. This estimate was adjusted for shrinkage using Rozeboom's formula 8 (1978). The correlations of the weighted composite of all predictors with the criterion were also computed in two other ways: with equal weights for all predictors and with zero order validities used as weights.

The validities for predicting CTP were averaged across MOS. The validities for predicting ELS, MPD, and PFB were computed on a pooled sample. All validity estimates were corrected for attenuation in the criterion measure (using the reliability estimates reported in Chapter 2).

The Reduced Prediction Equations. The reduced equations were obtained via expert judgment, using a panel of three experts whose task was to identify independently what they considered to be the optimal equations for maximizing selection validities, and the optimal equations for maximizing classification efficiency. The judgment task was constrained by stipulating in each case that the prediction equations (i.e., for selection and for classification) could contain no more than ten predictors. The judges were free to use fewer variables if they thought that a smaller number would reduce error while preserving relevant variance, or would improve classification efficiency without significantly reducing the overall level of selection validity.

The equations independently identified by each judge were very similar. Differences were resolved by consensus. The zero order validities, regression weights, and predictor battery validities were then recomputed for the two sets (selection vs. classification) of reduced equations.

Results

The results for the full and reduced equations for predicting ELS, MPD, and PFB are shown in Table 9.7. The mean results across MOS for predicting CTP are shown in Table 9.8. In general, differential predictor weights do provide some incremental validity over unit weights. However, zero-order validity coefficients as weights are virtually as good as regression weights and the reduced equations yield about the same level of predictive accuracy as the full equations. In fact, the reduced equations do slightly better.

Perhaps the most striking feature in Table 9.8 is the overall level of the correlations. The validities are very high. The best available estimate of the validity of the Project A/Career Force predictor battery for predicting Core Technical Performance is contained in the last column of the table, which is the adjusted multiple R corrected for unreliability in the criterion (i.e., CTP in LVI). The estimated validities (averaged over MOS) in this column are .78 ± .01. The reduced equations produce this level of accuracy, which does break the so-called validity ceiling, just as readily as the full equation, with perhaps more potential for producing classification efficiency.

Table 9.7
Validity Estimates for Full and SME-Reduced (Optimal) Equations for Maximizing Selection Efficiency for Predicting Will-Do Criterion Factors

	ELS	MPD	PFB
N	3,086	3,086	3,086
Full Equations			
Foldback R	.372	.277	.346
Adjusted R	.349	.243	.321
Corrected R[a]	(.381)	(.270)	(.354)
Reduced Equations			
Foldback R	.354	.232	.310
Adjusted R	.347	.224	.304
Validity Weights	.346	.202	.297
Unit Weights	.341	.187	.252

[a] Corrected for criterion unreliability.

Table 9.8
Estimates of Maximizing Selection Efficiency Aggregated Over MOS: Predicting Core Technical Proficiency

	Mean Selection Validity				
	Unit Weights	Validity Weights	Foldback R	Adjusted R	Corrected R*
Full Equation (All Predictors)	.576	.697	.762	.701	(.772)
Reduced Equation: Selection	.668	.720	.739	.716	(.789)
Reduced Equation: Classification	.651	.716	.734	.714	(.786)

* Corrected for criterion unreliability.

The Reenlistment Decision: Optimal Prediction of Second-Tour Performance

For a small sample of individuals for whom second-tour performance measures are available, the Project A/Career Force database can also provide ASVAB scores, Experimental Predictor Battery scores, and first-tour performance measures. Consequently, for such a sample it was possible to estimate the validity with which the components of second-tour performance could be predicted by alternative combinations of ASVAB factor scores, Experimental Battery predictor composites, and LVI performance factor assessments.

Analytic Approach

For the LVII analyses, soldiers were required to have complete data on (a) the four ASVAB composites, (b) nine of the Experimental Battery predictors, (c) four of the LVI criterion factors, and (d) five of the LVII criterion factors. Complete data were favored instead of pairwise deletion (i.e., complete data required only for the specified variables being correlated) for all analyses because of the possibility of ill-conditioned covariance matrices (e.g., not positive definite). In light of the multivariate adjustment applied to the primary covariance matrix (i.e., correction for range restriction), the loss of sample size was considered less detrimental to the analyses than the possibility of a poor covariance structure. Only a subset of Experimental Battery predictors was included so that the predictor/sample size ratio remained reasonable for the LVII analyses. The following variables were used in the analyses summarized here:

PREDICTORS

ASVAB:	Quantitative Speed Technical Verbal
Experimental Battery:	Spatial Rugged/Outdoors Interests (AVOICE) Achievement Orientation (ABLE) Adjustment (ABLE) Physical Condition (ABLE) Cooperativeness (ABLE) Internal Control (ABLE) Dependability (ABLE) Leadership (ABLE)
First-Tour Performance (LVI):	Can-Do (CTP + GSP) Will-Do (ELS + MPD + PFB)

LVII CRITERIA

Core Technical Proficiency (CTP)
Leadership (LDR)
Effort/Achievement (EA)
Will-Do (LDR + EA + MPD + PFB)

Two MOS did not appear in the LVII analyses--19E (too few soldiers) and 31C (no LVII criterion scores). Sample sizes for each of the MOS constituting the LVII analytic samples ranged from 10 - 31 and the total number of usable cases was 130.

The basic analytic approach was to calculate the multiple correlations for a selected set of hierarchical regression models. The correlations reflect (a) correction for range restriction using the procedure developed by Lawley (1943) and described by Lord and Novick (1968, p. 147), and (b) adjustment for shrinkage using Rozeboom's formula 8 (1978).

For the prediction of LVII performance, the procedure that was used to correct for range restriction is a function of the specific prediction, or personnel decision, that is being made, which in turn governs how the population parameter to be estimated is defined. There are two principal possiblities. First, we could be interested in predicting second-tour performance at the time the individual first applies to the Army. In this case the referent population would be the applicant sample. Second, we could be interested in the reenlistment decision and in predicting second-tour performance from information available during an individual's first tour. In this case, the referent population is all first-tour job incumbents; it is the reenlistment decision that is being modeled.

Consequently, for the LVII "rollup" analyses, a covariance matrix containing the ASVAB composites, the selected Experimental Battery predictors, LVI Can-Do and Will-Do composites, and the five basic LVI criterion composites served as the target matrix. The EOT measures were treated as incidental, rather than explicit, selection variables. The matrix was calculated using scores from all LVI soldiers having complete data on the specified measures (N = 3,702).

Results

A summary of the results is shown in Table 9.9 (Chapter 6 provides more detail). A hierarchical set of three equations for predicting four LVII criteria was developed. The four second-tour performance criterion scores were (a) the CTP factor, (b) the Leadership (LDR) factor, (c) the Achievement and Effort (EA) factor, and (d) the Will-Do composite of EA, LDR, MPD, and PFB. The three alternative predictor batteries are:

(1) ASVAB alone (4 scores).

(2) ASVAB + the Experimental Battery (4 + 9 scores).

(3) ASVAB + the Experimental Battery + LVI Performance (4 + 9 + 1 scores). The LVI performance score is either the Can-Do composite of CTP + GSP or the Will-Do composite of ELS + MPD + PFB, depending on the criterion score to be predicted. The LVI Can-Do composite was used in the equation when CTP was being predicted and the LVI Will-Do composite was used in the equation when the LVII LDR, EA, and Will-Do criterion scores were being predicted.

Three estimates of each validity are shown in Table 9.9: (a) the unadjusted, or foldback multiple correlation coefficient, (b) the multiple R corrected for shrinkage, and (c) the zero order correlation of the unit-weighted predictor composite with the criterion.

Table 9.9
Multiple Correlations for Predicting Second-Tour Job Performance (LVII) Criteria From ASVAB and Various Combinations of ASVAB, Selected Experimental Battery Predictors, and First-Tour (LVI) Performance Measures

		Predictor Composite			
LVII Criterion	Type	A	A+X	A+X+1	1
Core Technical Proficiency (CTP)	Unadj	.69	.80	.80	.35
	Adj	.64	.69	.68	.35
	Unit	.52	.39	.42	.35
Can-Do	Unadj	.72	.83	.86	.54
	Adj	.68	.74	.76	.54
	Unit	.60	.50	.54	.54
Leadership (LDR)	Unadj	.43	.61	.76	.53
	Adj	.36	.43	.65	.53
	Unit	.40	.43	.50	.53
Effort and Achievement (EA)	Unadj	.23	.30	.58	.54
	Adj	.00	.00	.38	.54
	Unit	.16	.13	.21	.54
Will-Do	Unadj	.18	.33	.62	.57
	Adj	.00	.00	.47	.57
	Unit	.14	.15	.23	.57

Note: All values in the table are corrected for restriction of range and criterion unreliability. Unadj and Adj reflect raw and shrunken (by Rozeboom formula 8, 1978) multiple correlations, respectively. Unit indicates unit weighted.

Key A = ASVAB factors (Quantitative, Speed, Technical, Verbal).
X = Experimental Battery (Spatial, Rugged/Outdoors Interests, Achievement Orientation, Adjustment, Physical Condition, Internal Control, Cooperativeness, Dependability, Leadership).
1 = The LVI Can-Do or Will-Do composite.

All three estimates have been corrected for unreliability in the criterion measures (using the reliability estimates contained in Chapter 2).

When predicting Core Technical Proficiency, the fully corrected estimate of the population validity using ASVAB alone is .64. Adding the EB raises it to .69 but LVI performance information does not add incrementally. The corresponding increments for the Can-Do composite criterion are .68 to .74 to .76. First-tour performance information adds considerably more to the prediction of LDR, EA, and the Will-Do composite. For example, for LDR, the validities go from .36 (ASVAB alone) to .43 (ASVAB + EB) to .65 (ASVAB + EB + LVI).

In general, the prediction of the major factors of second-tour performance can be made with considerable validity; however, the relative contribution of prior performance information depends on the specific components of performance that are being predicted. The best single predictor of Can-Do is ASVAB. The best single predictor of Will-Do is first-tour performance.

Selection Validity and Classification Efficiency

The cumulative validation analyses of Project A/Career Force show that, when compared to the full history of personnel selection research, performance as a first-tour soldier and performance as a second-tour soldier can be predicted with very high accuracy. The validities may be at the limit of what is possible, especially for technical proficiency, when appropriate criterion measures are used. They are considerably above the population values identified in the most recent meta analyses (e.g., Schmidt, Ones, & Hunter, 1992), and the difference cannot be attributed to artifacts. The samples in Project A/Career Force were too large, there was too much replication over jobs and over cohorts, and both predictors and criteria have too much construct validity. The most fundamental objectives of these two projects were met and surpassed.

Comparison of different types of equations has provided a glimpse of what the potential added benefits of classification might be (Table 8-24 in Chapter 8 of this report). By using a reduced prediction equation of less than 10 variables, making job assignments to only nine job families, eliminating only 5 percent of applicants with the AFQT, and using an operationally realistic set of quotas, a conservative estimate of the potential gain from classification is an average of about .20 of a standard deviation in technical task performance per person. The system-wide implications of this kind of average gain are enormous. Given a higher selection ratio, more job families, and more expertly identified classification equations, the potential gains could be even greater.

Performance Component Criticality and the Utility of Performance

During Project A, two scaling procedures were carried out. The first one used officers and senior NCOs as SMEs to scale the relative importance of each of five first-tour performance factors within each MOS (Sadacca, Campbell, White, & DiFazio, 1988). A modified form of conjoint measurement was used to obtain the scale values, which exhibited very high reliabilities. Across all five factors the range of importance values within MOS varied between multiples of 1.5 to 1.0 and 3.0 to 1.0, and the profiles of scale values were not the same across MOS. The largest differences among MOS were due to the importance values assigned to CTP vs. ELS. For some MOS, CTP was judged as more critical and the differences were often large.

The second scaling procedure involved using NCO and officer SMEs to scale the utility of five levels of performance (10th, 30th, 50th, 70th, and 90th percentiles) with each MOS (Sadacca, DiFazio, & Schultz, 1988). Magnitude estimation techniques were used to scale each MOS x performance level combination against a "standard" performer. There were 276 MOS and five performance levels, which yielded 1,380 (276 x 5) unique combinations to be scaled. The procedure yielded a ratio scale, and the utilities of

specific MOS x performance level combinations can be directly compared. The scale values are very reliable and the ratios of high to low values across MOS or across performance levels range from approximately 3 to 1 and 6 to 1.

These two sets of scale values are important for the estimation of selection validity and classification efficiency for different combinations of performance components and for the evaluation of job assignment strategies. The utility values are in terms of multiples of the utility of the 90th percentile infantryman (MOS 11B). Consequently, Mean Predictor Performance or Mean Actual Performance can be transformed into a utility value that represents gain in terms of the number of additional soldiers equivalent to the 90th percentile infantryman, or some other "standard" performer. Since the regression of utility (i.e., the value of performance) on performance does not have the same slope and intercept across MOS, and is not always linear, the gain from classification expressed in the performance metric is not necessarily isomorphic with gains expressed in utility terms. Because the scaling procedure was relatively efficient and produced highly reliable scale values, the scale values could be recalibrated as significant changes occur in the Army occupational structure.

A FINAL SUMMARY

This is the last chapter of the final report for a remarkable series of personnel research and development projects. Taken together, these projects represented a paradigm shift in the way personnel research is conducted. Virtually the entire personnel system for entry-level and first-level supervisory positions in a large and complex organization was made the focus for a long-term, systematic research effort. In one sense it was an attempt to apply the entire textbook at once. The measurement of individual differences in abilities, personality, personal history, interests, and outcome preferences was meant to be as comprehensive as the science would allow. Jobs were sampled representatively from the entire personnel system and each job in the sample was analyzed as thoroughly as current job analysis technology would allow. Every critical component of performance in each job in the sample was measured with multiple measures. As a result of all this, the two projects have produced a database beyond anything that has been seen before.

The design called for both concurrent and longitudinal estimates of selection efficiency for training performance, entry-level job performance, and supervisory (NCO) performance. To make such estimation possible, thousands of individuals in the target jobs had to be assessed for 1 to 2 days on each of three specified occasions spanning a time interval of approximately 6 years. This was a data collection and data management task that was orders of magnitude beyond what had ever been attempted before.

Perhaps the most remarkable aspect of this entire effort was that the original design, complex and ambitious as it was, was executed in its entirety according to its original basic schedule. The cohorts that were in the original research plan were, indeed, the ones that were sampled, and within the appropriate time frames. Such continuity, cooperation, and goal directedness over such a long time is a credit to all the participants.

The Basic Objectives

As noted at the beginning of this chapter, and as we have tried to summarize in subsequent sections, all the major objectives for Project A/Career Force were met. That is:

The latent structure of performance at each of the three career stages was modeled and represented by observable measures with a reliability, validity, and consistency that was beyond anyone's most optimistic expectations. We view the way in which performance was conceptualized and assessed in the project as a model for how personnel selection and classification research data should be accumulated.

An experimental predictor battery was produced that was comprehensive, versatile, and efficient; and that yielded basic predictor scores which are highly reliable and construct valid.

The ASVAB was shown to be an excellent predictor of future performance. The two components of performance that it predicts best are technical task performance and leadership. The validities are significantly above the generalized parameter estimates produced by the existing meta-analyses of the extant personnel selection research data and the difference cannot be explained by sampling error.

The Experimental Battery was fully analyzed in terms of its absolute validity, discriminant validity, and incremental validity for different performance factors within jobs (MOS) and for prediction of specific performance factors across jobs. The role of prior performance for the prediction of future performance was also examined. These are the most comprehensive and systematic estimates of these parameters that have ever been obtained. When the full set of predictors is used, it can be seen that while cognitive abilities are the primary predictor of technical task performance, certain aspects of personality and interests also contribute in such a way that the mean selection validity across jobs (corrected for artifacts) is .78 ± .01. This value is significantly greater than with ASVAB alone. Leadership can also be predicted with considerable accuracy from the full predictor battery. If prior performance data are available, the prediction of leadership performance becomes even more accurate. That leadership effectiveness is multiply determined in very significant ways is a crucial finding.

The differential prediction/classification efficiency glass is half full or half empty, depending on how one chooses to look at it. Virtually any prediction equation devised from the ASVAB plus the Experimental Battery will show substantial correlations with all performance components in all MOS, with the possible exception of personnel discipline (MPD) and physical fitness and bearing (PFB). At the same time, there is a significant distinction between the variables that predict Can-Do best and those that predict Will-Do best. Also, under a reasonable set of conditions, the predictor battery will produce significant gains from classification over selection. The systemwide average gain per individual of .20 standard deviation on technical task performance that was obtained is enormous. The question of how such a potential gain would be preserved by various kinds of operational job assignment systems is a matter for future research.

Also yet to be exploited fully are the project-generated estimates for the relative importance of specific performance components across jobs (MOS) and for the utility of performance at different levels in different jobs. These two scaling efforts yielded criticality and utility estimates that are very reliable and that differ across MOS. Such information will greatly enhance future evaluation of job assignment stategies.

Taken as a whole, the methodology used to analyze absolute versus discriminant validity for the ECAT battery, the classification model used to develop and evaluate the new estimate for Mean Actual Performance (reMAP), and the criticality/utility scales used to estimate the performance component constitute an analytic framework that can be used to evaluate selection validity versus classification. Together with the Project A/Career Force database itself, they constitute the foundation for a very useful selection/classification system "test bed" for a variety of personnel selection and job assignment procedures.

As with a number of other aspects of these projects, the breadth and quality of the database itself has probably exceeded everyone's initial expectations. It is extensive and well documented, and should prove very useful for many years to come.

Beyond the Objectives

Beyond the basic objectives themselves, the Project A/Career Force projects have produced a long list of additional findings and products that are discussed throughout the full set of contract technical reports. For illustrative purposes, a partial list follows. The list is divided into products for the "science" (personnel research) and products for the organization (the Army). The list is intended to move from the scientific to the applied. However, the distinction is not always easy to make since many products are useful for both.

(1) There exist, in technical report form, comprehensive reviews of all validity evidence pertaining to selection and classification for skilled jobs.

(2) As a by-product of the analyses involving ASVAB, there exists a much clearer idea of its factor structure, of what the factors are measuring, and of what its strengths and limitations are.

(3) The scope of the project made it possible to examine virtually the entire domain of selection information, sample from it, and investigate the basic validities and psychometric characteristics produced by each major piece of information. These data can be used by future investigators for a wide range of research questions pertaining to specific variables.

(4) The results of an expert judgment study of expected correlations between predictor constructs and performance factors are available. In brief, a large sample of personnel experts considered the population of predictor and criterion variables appropriate for entry-level jobs and forecasted what the validity coefficients would be.

The consistency in the judgments and their correspondence with known data points make these a potentially valuable tool for future test selection and synthetic validation work.

(5) Much has been learned about the nature of performance in entry-level skilled jobs (e.g., first-tour MOS). We now have a much clearer idea of what major factors constitute performance and how they can be measured. The "criterion problem" is better understood. This knowledge should better inform future enlistment and promotion policy, as well as future personnel research.

(6) The performance measurement and validation data support the assertion that supervisory ratings of performance have considerable construct validity if procedures for developing measurement and collecting data are carefully followed. Peer ratings can also be very useful. Both supervisor and peer ratings of first-tour performance produced the same three-factor structure of the Army-wide ratings, and in spite of the general factor; the three-factor structure replicated across cohorts virtually to the second decimal place.

(7) The potential of the AVOICE for differentially predicting "can do" performance in combat vs. technical vs. administrative support MOS has been established. Empirical scoring keys have been developed for this purpose.

(8) The Project A job/task analysis procedures worked well and can be used by the Army in the future to develop training curricula, performance measures, tests for requalifying on the job, and field exercises. The job analysis summaries for each MOS serve as a model for future job analysis work in the Army as well as in the public and private sector.

(9) Advanced Individual Training achievement measures have been developed for 21 MOS. The training measures showed that training performance predicts job performance.

(10) The package of procedures for administering rating scales can be used in future personnel research in the Army. A major effort in the Project A research was to develop an effective and very efficient set of procedures for administering performance rating scales to large numbers of people. These procedures and the package of materials can be adapted for use in other Army personnel research where ratings of many persons are required.

(11) The Supervisory Description Questionnaire (which came out of second-tour job analyses work) is a very useful instrument for future work in designing leadership training or evaluating leadership/supervisor performance. The questionnaire is based on a clear rationale and is straightforward to use.

(12) The Project A performance measures, against which new selection/ classification decision procedures were calibrated, have been demonstrated not to be discriminatory. The Project A samples that examine the interactions of rater and ratee race exceed the magnitude of the combined samples from all previous research on this issue.

(13) The MOS-specific rating scales developed using the BARS procedure and the attendant critical incident pool can easily be used for coaching/training purposes.

(14) The first-tour performance rating scales can also be used as predictors, as for selection into the Special Forces or selection for reenlistment.

(15) New statistical procedures for the estimation of discriminant validity and the estimation of classification efficiency have been developed.

This list does not count numerous other projects conducted or sponsored by the Army Research Institute that have been stimulated or made possible by Project A/Career Force. It also does not attempt to enumerate all the ways in which the data based on results are used by both ARI and the consortium members on an almost daily basis.

A NOTE OF THANKS

Participating in Project A and Career Force has been the career opportunity of a lifetime for many of us. We are extremely grateful to the Army Research Institute for its absolute unwavering support of our efforts and for the expert and competent "bosses" that were provided for us. To say that this has been a collegial effort on the part of the consortium and ARI is to understate the relationship exponentially. ARI staff participated with us in all phases of the work and the worry--conceptualizing, information gathering, instrument developing, theory building, data collecting, data analyzing, report writing, and briefing. They felt the same anxieties and pressures we felt as "can't miss" deadlines approached and we struggled to meet them.

We are also deeply indebted to the U.S. Army for providing so many soldiers and so much support. We are exhilarated by and grateful for their competence, enthusiasm, esprit, and candor.

The field of personnel psychology also owes a debt of gratitude to ARI and the U.S. Army for making these projects possible. As described above, the products of these efforts have considerable value in terms of their contribution to the general understanding of selection, classification, and performance assessment issues and in terms of their potential applications in other organizations.

Again, we offer our heartfelt appreciation and thanks for all that was made possible.

References

Abbe, C. N. (1968). *Statistical properties of allocation averages* (Research Memorandum 68-13). Washington, DC: U.S. Army Behavioral Science Research Laboratory.

Abrahams, N. M., Alf, E. F., Kieckhaefer, W. F., Pass, J. J., Cole, D. R., & Walton-Paxton, E. (1994). *Classification utility of test composites from the ASVAB, CAT-ASVAB, and ECAT batteries* (Contract N66001-90-D-9502, Delivery Order 7J16). San Diego: Navy Personnel Research and Development Center.

Abrahams, N. M., Pass, J. J., Kusulas, J. W., Cole, D. R., & Kieckhaefer, W. F. (1993). *Incremental validity of experimental computerized tests for predicting training criteria in military technical schools* (Contract N66001-90-D-9502, Delivery Order 7J13). San Diego: Navy Personnel Research and Development Center.

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

Alley, W. E., Wilbourn, J. M., & Berberich, G. L. (1976). *Relationship between performance on the Vocational Interest-Career Examination and reported job satisfaction* (AFHRL-TR-76-89). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data* (Sage University paper series on quantitative applications in the social sciences, No. 07-046). Beverly Hills, CA: Sage.

Austin, J. T., & Hanisch, K. A. (1990). Occupational attainment as a function of abilities and interests: A longitudinal analysis using project TALENT data. *Journal of Applied Psychology, 75,* 77-89.

Barge, B. N., & Hough, L. M. (1988). Utility of biographical assessment: A review and integration of the literature. In L.M. Hough (Ed.), *Utility of temperament, biodata, and interest assessment for predicting job performance: A review of the literature* (ARI Research Note 88-02, pp. 91-130). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A192 109)

Bartling, H. C., & Hood, A. B. (1980). *Validity of measured interest for decided and undecided students.* Paper presented at the annual convention of the American Psychological Association.

Brogden, H. (1946a). An approach to the problem of differential prediction. *Psychometrika, 11,* 139-154.

Brogden, H. (1946b). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37,* 65-76.

Brogden, H. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educational and Psychological Measurement, 11,* 173-195.

Brogden, H. (1954). A simple proof of a personnel classification theorem. *Psychometrika, 19,*(3), 205-208.

Brogden, H. (1955). Least squares estimates and optimal classification. *Psychometrika, 20*(3), 249-252.

Brogden, H. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement, 19(2),* 181-190.

Brogden, H. E., & Taylor, E. K. (1950). The dollar criterion--applying the cost accounting concept to criterion construction. *Personnel Psychology, 3*, 133-154.

Brush, D. H., & Owens, W. A. (1979). Implementation and evaluation of an assessment classification model for manpower utilization. *Personnel Psychology, 32*, 369-383.

Bryk, A. S., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Camara, W. J., & Laurence, J. H. (1987). *Military classification of high aptitude recruits* (FR-PRD-87-21). Alexandria, VA: Human Resources Research Organization.

Campbell, C. H., Ford, P., Rumsey, M. G., Pulakos, E. D., Borman, W. C., Felker, D. B., de Vera, M. V., & Riegelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. *Personnel Psychology, 43*, 277-300.

Campbell, D. P. (1966). *Manual for Strong Vocational Interest Blanks.* Stanford, CA: Stanford University Press.

Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB-SCII* (3rd ed.). Stanford, CA: Stanford University Press.

Campbell, J. P. (Ed.) (1987a). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1985 fiscal year* (ARI Technical Report 746). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A193 343)

Campbell, J. P. (Ed.) (1987b). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1986 fiscal year* (ARI Technical Report 813101). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. P. (Ed.) (1988). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1987 fiscal year* (ARI Technical Report 862). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 219 046)

Campbell, J. P. (Ed.) (1991). *Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1988 fiscal year* (ARI Research Note 91-34). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social & Sciences. (ADA 233 750)

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43*, 313-333.

Campbell, J. P., & Oppler, S. (1990). Modeling of second-tour performance. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1990 fiscal year* (ARI Technical Report 952). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social & Sciences. (ADA 252 675)

Campbell, J. P., Peterson, N. G., & Johnson, J. (in press). The prediction of future performance from current performance and from training performance. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1993 fiscal year* (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. P., & Zook, L. M. (Eds.) (1990). *Building and retaining the Career Force: New procedures for assessing and assigning Army enlisted personnel - Annual Report, 1990 fiscal year* (ARI Technical

Report 952). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 252 675)

Campbell, J. P., & Zook, L. M. (Eds.) (1991). *Improving the selection, classification, and utilization of Army enlisted personnel: Final Report an Project A* (ARI Research Report 1597). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 242 921)

Campbell, J. P., & Zook, L. M. (Eds.) (1994a). *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1991 fiscal year* (ARI Research Note 94-10). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 278 726)

Campbell, J. P., & Zook, L. M. (Eds.) (1994b). *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1992 fiscal year* (ARI Research Note 94-27). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. P., & Zook, L. M. (Eds.) (in press). *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1993 fiscal year* (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, R. (1985). *Scorer training materials* (ARI RS-WP-85). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

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

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement, 2*, 295-601.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, M. E., & Hudson, D. L. (1992). Medical decision making using pattern classification techniques for establishment of knowledge bases. In A. Kandel (Ed.), *Fuzzy Expert Systems.*

Collins, J. M. & Clark, M. R. (1993). An application of the theory of neural computation to the prediction of workplace behavior: An illustration and assessment of network analysis. *Personnel Psychology, 46*, 503-524.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, 34*, 187-202.

Cronbach, L., & Gleser, G. (1957). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Devlin, S. E., Abrahams, N. M., & Edwards, J. E. (1992). Empirical keying of biographical data: Cross-validity as a function of scaling procedure and sample size. *Military Psychology, 4*. 119-136.

DuBois, P. H. (1964). A test-dominated society: China, 1115 B.C. - 1950 A.D. In *Proceedings, ETS Invitational Conference on Testing.*

Dunbar, S. B., & Novick, M. R. (1988). On predicting success in training for men and women: Examples from Marine Corps clerical specialties. *Journal of Applied Psychology, 73*, 545-550.

Equal Employment Opportunity Commission, U.S. Civil Service Commission, Department of Labor, & Department of Justice (1978). *Uniform guidelines on employee selection procedures.* 43 Fed. Reg. 166, 38290-38309.

Flanagan, J. C. (1948). The Aviation Psychology Program in the Army Air Forces. *AAF Aviation Psychology Program Research Reports, 1*, U.S. Government Printing Office.

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

Gamache, L. M., & Novick, M. R. (1985). Choice of variables and gender differentiated prediction within selected academic programs. *Journal of Educational Measurement, 22*, 53-70.

Gellatly, I. R., Paunonen, S. V., Meyer, J. P., Jackson, D. N., & Goffin, R. D. (1991). Personality, vocational interest, and cognitive predictors of managerial job performance and satisfaction. *Personality and Individual Differences, 12,* 221-231.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (vol. 2, pp. 141-165). Beverly Hills, CA: Sage.

Gottfredson, G. D., & Holland, J. L. (1975). Vocational choices of men and women: A comparison of predictors from the Self-Directed Search. *Journal of Counseling Psychology, 22,* 28-34.

Greener, J. M., & Osburn, H. G. (1980). Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and nonlinear distributions. *Educational and Psychological Measurement, 40*, 337-346.

Guilford, J. P. (1957). *A revised structure of intelligence* (Report No. 19). University of Southern California Psychological Laboratory.

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

Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB-SCII* (4th ed.). Stanford, CA: Stanford University Press.

Hansen, J. C., Collins, R. C., Swanson, J. L., & Fouad, N. A. (1993). Gender differences in the structure of interests. *Journal of Vocational Behavior, 42*, 200-211.

Hansen, J. C., & Tan, R. N. (1992). Concurrent validity of the 1985 Strong Interest Inventory for College Major Selection. *Measurement and Evaluation in Counseling and Development, 25*, 53-57.

Hanson, M. A., & Borman, W. B. (in press). *Development and construct validation of the Situational Judgment Test (SJT)* (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hanson, M. A., Campbell, J. P., & McKee, A. S. (1994). Development of the longitudinal validation sample second-tour performance model. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New Procedures for accessing and assigning Army enlisted personnel - Annual Report. 1992 fiscal year* (ARI Research Note 94-27). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hatch, R. S., Pierce, M. B., & Fisher, A. H. (1968). *Development of a computer-assisted recruit assignment system (COMPASS II)*. Rockville, MD: Decision Systems.

Hays, W. L. (1963). *Statistics* (1st ed.). New York: Holt Rinehart & Winston.

Holland, J. L. (1966). *The psychology of vocational choice: A theory of personality types and model environments.* Walthem, MA: Blaisdell.

Horst, P. (1954). A technique for the development of a differential prediction battery. *Psychological Monographs: General and Applied, 68*(5, Whole No. 380).

Horst, P. (1955). A technique for the development of a multiple absolute prediction battery. *Psychological Monographs,* No. 390.

Horst, P. (1956). Multiple classification by the method of least squares. *Journal of Clinical Psychology, 12,* 3-16.

Horst, P., & MacEwan, C. (1957). Optimal test length for multiple prediction: The general case. *Psychometrika, 22*(4), 311-324.

Hough, L. M. (1992). The "Big Five" personality variables - construct confusion: Description versus prediction. *Human Performance, 5,* 139-155.

Hough, L. M., Barge, B. N., & Kamp, J. D. (1987). Non-cognitive measures: Pilot testing. In N. G. Peterson (Ed.), *Development and field test of the Trial Battery for Project A* (ARI Technical Report 739, pp. 7-1 through 7-48). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 183 575)

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those variables. *Journal of Applied Psychology, Monograph, 75,* 581-595.

Hough, L. M., McCloy, R. A., Ashworth, S. D., & Hough, M. M. (1987). Analysis of temperament/biodata, vocational interest, and work environment preference measures: Concurrent validity sample. Working paper.

Hough, L. M., McGue, M. K., Houston, J. S., & Pulakos, E. D. (1987). Non-cognitive measures: Field tests. In N. G. Peterson (Ed.), *Development and field test of the Trial Battery for Project A* (ARI Technical Report 739, pp. 8-1 through 8-39). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 184 575).

Hough, L. M., & Paullin, C. (in press). Construct-oriented scale construction: The rational approach. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *The biodata handbook: Theory, research, and application.*

Houston, W. M., & Novick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement, 24,* 309-320.

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist or artist. *Journal of Applied Psychology, 78*, 250-261.

Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E. A. Fleishman & M. D. Dunnette (Eds.), *Human performance and productivity, Vol. 1: Human capability assessment*. Hillsdale, NJ: Erlbaum.

Johnson, C. D., & Zeidner, J. (1990). *Classification utility: Measuring and improving benefits in matching personnel to jobs* (IDA Paper P-2240). Alexandria, VA: Institute for Defense Analysis.

Johnson, C. D., & Zeidner, J. (1991). *The economic benefits of predicting job performance: Vol. II Classification efficiency*. New York: Praeger.

Johnson, C. D., Zeidner, J., & Scholarios, D. (1990). *Improving the classification efficiency of the Armed Services Vocational Aptitude Battery through the use of alternative test selection indices* (IDA Paper P-2427). Alexandria, VA: Institute for Defense Analysis.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL VI: Analysis of linear squares methods*. Uppsala, Sweden: University of Uppsala.

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structure relationship by the method of maximum likelihood*. Morresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Kluger, A. N., Reilly, R. R., & Russell, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology, 76*, 889-896.

Knapp, D. J. (1993, August). Alternative conceptualizations of turnover. In J. P. Campbell (Chair), *Prediction of turnover in a longitudinal sample using event history analysis*. Symposium conducted at the convention of the American Psychological Association, Toronto.

Knapp, D. J., Carter, G. W., McCloy, R. A., & DiFazio, A. S. (in press). The role of job satisfaction in performance, attrition, and reenlistment. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1993 fiscal year* (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Konieczny, F. B., Brown, G. N., Hutton, J., & Stewart, J. E. (1990). *Enlisted personnel allocation system: Final Report* (ARI Technical Report 902). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kroeker, L. P. (1988). Extending the Navy classification model. In B. F. Green, H. Wing, & A. K. Wigdor (Eds.), *Linking military enlistment standards to job performance: Report of a workshop*. Committee on the Performance of Military Personnel, National Research Council. Washington, DC: National Academy Press.

Kroeker, L. P. (1989). Personnel classification/assignment models. In M. F. Wiskoff & G. M. Rampton (Eds.), *Military personnel measurement: Testing, assignment, evaluation*. New York: Praeger.

Kroeker, L. P., & Folchi, J. (1984). *Classification and assignment within PRIDE (CLASP) system: Development and evaluation of an attrition component* (NPRDC TR 84-40). San Diego: Navy Personnel Research and Development Center.

Kroeker, L. P., & Rafacz, B. A. (1983). *CLASP: A recruit assignment model* (NPRDC TR 84-9). San Diego: Navy Personnel Research and Development Center.

Kuder, G. F., & Diamond, E. E. (1979). *Kuder DD Occupational Interest Survey General Manual.* Chicago: Science Research Associates.

Lawley, D. (1943). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh, Proceedings* (Section A), *62*, 28-30.

Lecznar, W. B. (1951). *Evaluation of a new technique for keying biographical inventories empirically* (Research Bulletin 51-2). San Antonio, TX: Lackland AFB, Human Resources Research Center.

Linn, R. L. (1968). Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin, 69*, 69-73.

Linn, R. L. (1994). Fair test use: Research and policy. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McBride, J. R. (1993). Compensatory screening model development. In T. Trent & J. H. Laurence (Eds.), *Adaptability screening for the Army Forces* (pp. 163-214). Washington, DC: Office of the Assistant Secretary of Defense.

McCloy, R. A. (1993). *An overview of survival analysis.* Paper presented at the conference of the American Psychological Association, Toronto.

McCloy, R. A., & DiFazio, A. S. (in press). Prediction of first-term military attrition using pre-enlistment predictors. In J. P. Campbell & L. M. Zook, (Eds.), *Building and retaining the career force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1993 fiscal year* (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McCloy, R. A., Harris, D. A., Barnes, J. D., Hogan, P. F., Smith, D. A., Clifton, D., & Sola, M. (1992). *Accession quality, job performance, and cost: A cost-performance tradeoff model* (FR-PRD-92-11). Alexandria, VA: Human Resources Research Organization.

McCloy, R. A., Hedges, L. V., & Harris, D. A. (1991). *Development of a methodology to link recruit quality requirements to job performance: Estimation of model parameters* (HumRRO Interim Report IR-PRD-91-07). Alexandria, VA: Human Resources Research Organization.

McCormick, E. J., DeNisi, A., & Shaw, B. (1979). Use of the Position Analysis Questionnaire for establishing the job component validity of tests. *Journal of Applied Psychology, 64*, 51-56.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 56*, 347-368.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43*, 335-354.

Mitchell, K. J., & Hanser, L. M. (1984). *1980 Youth Population Norms: Enlisted and Occupational Classification Standards in the Army.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Morrison, R. F. (1977). A multivariate model for the occupational placement decision. *Journal of Applied Psychology, 62*, 271-277.

Mosier, C. I. (1951). Problems in design of cross-validation. *Educational and Psychological Measurement, 11*, 1-11.

Mossholder, K. W., & Arvey, R. D. (1984). Synthetic validity: A conceptual and comparative review. *Journal of Applied Psychology, 69*, 322-333.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 1-31.

Nord, R., & Schmitz, E. (1991). Estimating performance and utility effects of alternative selection and classification policies. In J. Zeidner & C. D. Johnson, *The economic benefits of predicting job performance: Vol. 3. The gains of alternative policies.* New York: Praeger.

Nord, R. D., & White, L. A. (1988). The measurement and application of performance utility. In B. Green, H. Wing, & A. Wigdor (Eds.), *Linking military enlistment standards to job performance,* pp. 215-243. Washington, DC: National Academy Press.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*, 574-583.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Oppler, S. H., Childs, R. A., & Peterson, N. G. (1994). Development of the longitudinal validation sample first-tour performance model. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report. 1991 fiscal year* (ARI Research Note 94-10). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 278 726)

Oppler, S. H., Peterson, N. G., & Rose, A. M. (in press). Basic validation results for the LVII sample. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1993 fiscal year* (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Oppler, S. H., Peterson, N. G., & Rosse, R. R. (1993). *Examination of differential prediction across LVI criterion constructs and across Batch A MOS.* Presented at Scientific Advisory Group Meeting. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Oppler, S. H., Peterson, N. G., & Russell, T. (1994). Basic validation results for the LVI sample. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1991 fiscal year* (ARI Research Note 94-10). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 278 726)

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 609-644). Chicago: Rand-McNally.

Owens, W. A. (1978). Moderators and subgroups. *Personnel Psychology, 31*, 243-248.

Owens, W. A., & Schoenfeldt, L. F. (1979). Towards a classification of persons [Monograph]. *Journal of Applied Psychology, 64*, 569-607.

Peterson, N. G., Oppler, S. H., Sager, C. E., & Rosse, R. L. (1993). *Analysis of the Enhanced Computer Administered Test Battery: An evaluation of potential revisions and additions to the Armed Services Vocational Aptitude Battery* (Draft report prepared for the Selection and Classification for Critical MOS Project). Washington, DC: American Institutes for Research.

Peterson, N. G., Rosse, R. L., & Owens-Kurtz, C. K. (1990). Formation of job performance prediction equations and evaluation of their validity. In L. L. Wise, N. G. Peterson, R. G. Hoffman, J. P. Campbell, & J. M. Arabian (Eds.), *Army Synthetic Validation Project: Report of Phase III Results*. Alexandria, VA: American Institutes for Research.

Peterson, N., Russell, T., Hallam, G., Hough, L., Owens-Kurtz, C., Gialluca, K., & Kerwin, K. (1990). Analysis of the experimental predictor battery: LV sample. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for assessing and assigning Army enlisted personnel - Annual Report, 1990 fiscal year* (ARI Technical Report 952, pp. 73-199). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 252 675)

Peterson, N. G., Sager, C. E., Oppler, S. H., Rosse, R. L., & Crafts, J. L. (in press). Identification of optimal predictor batteries using a subset of the Project A/Career Force Experimental Predictor Battery. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the career force: New procedures for accessing and assigning army enlisted personnel, Annual Report for 1993 fiscal year* (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Pina, M. (1974). *The assignment of airmen by solving the transportation problem* (AFHRL-TP-87-41). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Pina, M. (1988). Air Force person-job match: Non-prior service enlisted classification. In B. F. Green, H. Wing, & A. K. Wigdor (Eds.), *Linking military enlistment standards to job performance: Report of a workshop*, Committee on the Performance of Military Personnel, National Research Council. Washington, DC: National Academy Press.

Pina, M., Jr., Emerson, M. S., Leighton, D. L., & Cummings, W. (1988). *Processing and Classification of Enlistees (PACE) system payoff algorithm development* (AFHRL-TP-87-41). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Prediger, D. J., Roth, J. D., & Noeth, R. J. (1974). Career development of youth: A nationwide study. *Personnel and Guidance Journal, 53*, 97-104.

Research Triangle Institute (1988). "Family research program and the Project A data base" (RTI/3795-44 WP). Draft working paper. Research Triangle Park, NC: Author.

Rosse, R. L., Whetzel, D. L., & Peterson, N. G. (1993). *Assessment of the classification efficiency of selection/assignment systems.* Washington, DC: American Institutes for Research.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A Clarification. *Psychological Bulletin, 85,* 1348-1351.

Rudnik, R. A., & Greenston, P. M. (in press). *Development of an Army prototype PC-based enlisted personnel allocation system (EPAS)* (ARI Study Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rulon, P. J., Tiedeman, D. V., Tatsuoka, M. M., & Langmuir, C. R. (1967). *Multivariate statistics for personnel classification.* New York: Wiley.

Sadacca, R., Campbell, J. P., DiFazio, A. S., Schultz, S. R., & White, L. A. (1990). Scaling performance utility to enhance selection/classification decisions. *Personnel Psychology, 43,* 367-378.

Sadacca, R., Campbell, J. P., White, L. A., & DiFazio, A. S. (1988). *Weighting criterion components to develop composite measures of job performance* (ARI Technical Report 838). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A210 357)

Sadacca, R., White, L. A., Campbell, J. P., DiFazio, A. S., & Schultz, S. R. (1988). *Assessing the utility of MOS performance levels in Army enlisted occupations* (ARI Technical Report 939). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Schmidt, F. L., Ones, D. S., & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627-670.

Schmitz, E. J. (1988). Improving personnel performance through assignment policy. In B. F. Green, H. Wing, & A. K. Wigdor (Eds.), *Linking military enlistment standards to job performance: Report of a workshop,* Committee on the Performance of Military Personnel, National Research Council. Washington, DC: National Academy Press.

Schoenfeldt, L. F. (1974). Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs. *Journal of Applied Psychology, 59,* 583-595.

Scholarios, T. M., Johnson, C. D., & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology, 79,* 412-424.

Skinner, H. A., & Jackson, D. N. (1977). The missing person in personnel classification: A tale of two models. *Canadian Journal of Behavioral Science, 9,* 147-160.

Society for Industrial and Organizational Psychology (1987). *Principles for the Validation and use of personnel selection procedures* (3rd ed.) College Park, MD: Author.

Sorenson, R. C. (1965). *Optimal allocation of enlisted men - Full regression equations versus aptitude area scores* (Technical Research Note 163). Washington, DC: U.S. Army Personnel Research Office.

Staff, AGO, Personnel Research Branch (1943a). Personnel research in the Army. I, Background and organization. *Psychological Bulletin, 40,* 129-135.

Staff, AGO, Personnel Research Branch (1943b). Personnel research in the Army. II, The classification system and the place of testing. *Psychological Bulletin, 40,* 205-211.

Staff, AGO, Personnel Research Branch (1943c). Personnel research in the Army. III, Some factors affecting research in the Army. *Psychological Bulletin, 40,* 237-278.

Staff, AGO, Personnel Research Branch (1943d). Personnel research in the Army. IV, The selection of radiotelegraph operators. *Psychological Bulletin, 40,* 357-371.

Staff, AGO, Personnel Research Branch (1943e). Personnel research in the Army. V, The Army specialized training program. *Psychological Bulletin, 40,* 429-435.

Staff, AGO, Personnel Research Branch (1943f). Personnel research in the Army. VI, The selection of truck drivers. *Psychological Bulletin, 40,* 499-508.

Statman, M. A. (1992, August). *Developing optimal predictor equations for differential job assignment and vocational counseling.* Paper presented at the Annual Convention of the American Psychological Association, Washington. DC.

Stead, W. H., & Shartle, C. L. (1940). *Occupational counseling techniques.* New York: American Book Company.

Stein, C. (1960). Multiple regression. In I. Olkin (Ed.), *Contributions to probability statistics.* Stanford, CA: Stanford University Press.

Strong, E. K., Jr. (1931). *Vocational interest blank for men.* Stanford, CA: Stanford University Press.

Strong, E. K. (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.

Stuit, D. B. (1947). *Personnel research and test development in the Bureau of Naval Personnel.* Princeton, NJ: Princeton University Press.

Thorndike, R. L. (1949). *Personnel selection.* New York: Wiley.

Tippett, L. H. C. (1925). On the extreme individuals and the range of samples taken from a normal population. *Biometrika, 17,* 364-387.

Valentine, L. D. (1977). *Prediction of Air Force technical training success from ASVAB and educational background* (AFHRL-TR-77-18). Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Ward, J. H., Jr. (1977, August). *Creating mathematical models of judgment processes: From policy-capturing to policy-specifying* (AFHRL-TR-77-47). Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory.

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). Manual for the Minnesota Satisfaction Questionnaire. *Minnesota Studies in Vocational Rehabilitation, 22.*

White, L. A. (1994). Development of composite scores for Assessment of the Background and Life Experiences (ABLE) instrument. In J. P. Campbell & L. M. Zook (Eds.), *Building and retaining the Career Force: New procedures for accessing and assigning Army enlisted personnel - Annual Report, 1992 fiscal year* (ARI Research Note 94-27), (pp. 25-32). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

White, L. A., & Moss, M. C. (1995). Factors influencing concurrent versus predictive validities of personality constructs: Impact of response distortion and item job content. In F. L. Schmidt (Chr.), Symposium: *Response distortion and social desirability in personality testing for personnel selection,* Society of Industrial and Organizational Psychology, Orlando.

Wilk, S. L., Desmarais, L. B., & Sackett, P. R. (1993). *Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests.* Paper presented at the Society of Industrial and Organizational Psychology, San Francisco.

Wise, L. L. (1994). Goals of the selection and classification decision. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification.* Hillsdale, NJ: Erlbaum.

Wise, L. L., McHenry, J. J., & Campbell, J. P. (1990). Identifying optimal predictor composites and testing for generalizability across jobs and performance factors. *Personnel Psychology, 43,* 355-366.

Wise, L. L., McHenry, J. J., & Young, W. Y. (1986). *Project A Concurrent Validation: Treatment of Missing Data.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wise, L. L., & McLaughlin, D. H. (1980). *Guidebook for the imputation of missing data.* Palo Alto, CA: American Institutes for Research.

Wise, L. L., Peterson, N. G., Hoffman, R. G., Campbell, J. P., Arabian, J. M. (1991). *The Army Synthetic Validity Project: Report of Phase III results. Vol. I* (ARI Technical Report 922). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Zadeh, L. A. (1972). Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics, 3*, 28-44.

Zeidner, J., & Johnson, C. D. (1989). *The utility of selection for military and civilian jobs* (IDA P-2239). Alexandria, VA: Institute for Defense Analysis.

Appendix A

Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency

Table A-1
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 11B

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat	2 (+)	7 (+)	+	+	3 (+)	7 (+)	+	
Rugged Indiv.	2 (+)	3 (+)	+		4 (+)	12 (+)	+	+
Firearms	1 (+)	5 (+)	+		2 (+)	6 (+)	+	
Drafting		1 (+)				5 (+)	+	+
Audiographics	1 (-)	1 (-)						
Aesthetics		1 (-)						
Medical Serv.		2 (-)				2 (-)		
Lead/Guidance						7 (+)	+	
Science/Chem.		2 (+)				3 (+)		
Computers	1 (-)	2 (-)	-					
Math						1 (+)		+
Elect. Commun.						1 (+)		
Cler./Admin.	3 (-)	10 (-)	-		3 (-)	5 (-)	-	-
Warehouse/Ship		1 (-)				3 (-)	-	
Food-Employee	1 (-)	3 (-)	-	-		6 (-)	-	
Food-Prof.	1 (-)	4 (-)	-			5 (-)	-	
Fire Prot.						1 (-)		
Law Enf.						1 (-)		
Mechanics						1 (+)		
Construction						2 (-)		
Electronics						2 (+)		
Vehicle Op.						4 (-)	-	
Unlikely Resp.								-

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-2
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 13B

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat	1 (+)	4 (+)	+		2 (+)	4 (+)	+	+
Rugged Indiv.	6 (+)	10 (+)	+	+	3 (+)	9 (+)	+	
Firearms	3 (+)	4 (+)	+			1 (+)		
Drafting					1 (+)	2 (+)		
Audiographics								
Aesthetics								
Medical Serv.								
Lead/Guidance		1 (-)						
Science/Chem.								
Computers								
Math								
Elect. Commun.								
Cler./Admin.	1 (-)	4 (-)	-		1 (-)	4 (-)	-	
Warehouse/Ship								
Food-Employee		1 (-)			3 (-)	3 (-)	-	-
Food-Prof.	1 (-)				1 (-)	1 (-)		
Fire Prot.					1 (+)	1		
Law Enf.								
Mechanics		4 (+)	+			1 (+)		
Construction		3 (+)						
Electronics		1 (+)				1 (+)		
Vehicle Op.						1 (-)		
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-3
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 63B Male

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat								
Rugged Indiv.	1 (+)	10 (+)	+		1 (+)	3 (+)	+	
Firearms		15 (+)						
Drafting								
Audiographics	1 (-)	2 (-)	-					
Aesthetics						1 (-)		
Medical Serv.		3 (-)				2 (-)		
Lead/Guidance						2 (-)		
Science/Chem.						1 (+)		
Computers		2 (-)						
Math								
Elect. Commun.					1 (-)	1 (-)		
Cler./Admin.		5 (-)	-		1 (-)	4 (-)	-	
Warehouse/Ship						1 (-)		
Food-Employee		2 (-)						
Food-Prof.		3 (-)				1 (-)		
Fire Prot.		1 (-)	-					
Law Enf.								
Mechanics	6 (+)	10 (+)	+	+	6 (+)	8 (+)	+	+
Construction	2 (+)	11 (+)	+		2 (+)	5 (+)	+	
Electronics	2 (+)	7 (+)	+		1 (+)	1 (+)		
Vehicle Op.		1 (+)	+					
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-4
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 63 Combined-Sex

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat		2 (+)				1 (+)		
Rugged Indiv.		12 (+)	+		1 (+)	8 (+)	+	
Firearms		4 (+)	+			4 (+)	+	
Drafting		1 (+) 1 (-)						
Audiographics		2 (-)	-					
Aesthetics		2 (-)				1 (-)		
Medical Serv.		5 (-)	-			2 (-)		
Lead/Guidance						2 (-)		
Science/Chem.						1 (+)		
Computers		2 (-)						
Math								
Elect. Commun.					1 (-)	1 (-)		
Cler./Admin.		7 (-)	-		1 (-)	10 (-)	-	
Warehouse/Ship						1 (-)		
Food-Employee						1 (-)		
Food-Prof.		3 (-)	-			1 (-)		
Fire Prot.		1 (-)	-					
Law Enf.								
Mechanics	9 (+)	10 (+)	+	+	7 (+)	10 (+)	+	+
Construction	2 (+)	13 (+)	+		1 (+)	8 (+)	+	
Electronics	1 (+)	9 (+)	+		1 (+)	2 (+)		
Vehicle Op.		3 (+)	+					
Unlikely Resp.							-	

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-5
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 71L Male

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat					1 (+)	2 (+)		
Rugged Indiv.	1 (+)				1 (+)	1 (+)		
Firearms								
Drafting								
Audiographics								
Aesthetics	1 (+)	1 (+)			1 (+)	1 (+)		
Medical Serv.	1 (+)							
Lead/Guidance					1 (+)	1 (+)		
Science/Chem.	1 (+)	1 (+)			1 (+)	1 (+)		
Computers					1 (+)	1 (+)		
Math	3 (+)	3 (+)	+	+	1 (+)	1 (+)		
Elect. Commun.					1 (+)	1 (+)		
Cler./Admin.	1 (+)	1 (+)			1 (+)	1 (+)		
Warehouse/Ship					1 (+)	1 (+)		
Food-Employee								
Food-Prof.								
Fire Prot.								
Law Enf.					1 (-)	1 (-)		
Mechanics	3 (+)	2 (+)						
Construction						1 (+)		
Electronics	1 (+)	1 (+)				1 (+)		
Vehicle Op.					1 (-)	1 (+) 1 (-)		
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-6
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 71L Female

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat								
Rugged Indiv.	4 (+)	6 (+)	+		1 (-)	1 (-)		
Firearms								
Drafting		1 (+)						
Audiographics					1 (+)	1 (+)		
Aesthetics					2 (+)	2 (+)		
Medical Serv.	2 (-)	3 (-)			2 (+)	2 (+)	+	+
Lead/Guidance	1 (+)	1 (+)				1 (+)		
Science/Chem.	1 (+)	1 (+)						
Computers								
Math					1 (+)	1 (+)		
Elect. Commun.								
Cler./Admin.	1 (-)	1 (-)			1 (+)	1 (+)		
Warehouse/Ship								
Food-Employee	2 (-)	3 (-)	-	-				
Food-Prof.					1 (+)	1 (+)		
Fire Prot.	1 (+)	1 (+)		+	1 (-)	1 (-)		
Law Enf.					2 (-)	2 (-)		
Mechanics						1 (-)		
Construction								
Electronics						1 (-)		
Vehicle Op.								
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-7
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 71L Combine-Sex

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat								
Rugged Indiv.	2 (+)	4 (+)			2 (+)	4 (+)	+	
Firearms								
Drafting					1 (+)	2 (+)	+	
Audiographics	1 (+)	1 (+)			3 (+)	3 (+)		
Aesthetics	1 (+)	1 (+)	+		2 (+)	2 (+)	+	
Medical Serv.		1 (+)			1 (-)	1 (-)		
Lead/Guidance	1 (+)	3 (+)				1 (+)		
Science/Chem.								
Computers								
Math	2 (+)	3 (+)	+		1 (+)	1 (+)		
Elect. Commun.				+				
Cler./Admin.	1 (+)	2 (+)			1 (+)	1 (+)		
Warehouse/Ship								
Food-Employee	2 (-)	2 (-)	-	-				
Food-Prof.					1 (+)	1 (+)		
Fire Prot.								
Law Enf.						2 (-)		
Mechanics						1 (-)		
Construction	2 (-)	4 (-)						
Electronics								
Vehicle Op.								
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-8
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 91A Male

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat	3 (+)	6 (+)	+					
Rugged Indiv.	2 (+)	5 (+)	+	+	3 (+)	4 (+)	+	
Firearms						1 (+)		
Drafting								
Audiographics	1 (-)	4 (-)	-	-				
Aesthetics								
Medical Serv.	3 (+)	7 (+)	+	+	3 (+)	4 (+)	+	
Lead/Guidance		1 (+)						
Science/Chem.	2 (+)	2 (+)	+		1 (+)	2 (+)		
Computers		1 (-)						
Math								
Elect. Commun.	1 (-)	1 (-)						
Cler./Admin.		2 (-)	-		1 (-)	1 (-)		
Warehouse/Ship								
Food-Employee								
Food-Prof.								
Fire Prot.								
Law Enf.					3 (+)	3 (+)	+	+
Mechanics					1 (+)	1 (+)		
Construction								
Electronics								
Vehicle Op.								
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-9
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 91A Female

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat								
Rugged Indiv.	5 (+)	5 (+)	+	+	3 (+)	2 (+)		
Firearms								
Drafting					1 (+)	1 (+)		
Audiographics					1 (-)	1 (-)		
Aesthetics					1 (+)	1 (+)		
Medical Serv.	1 (+)	1 (+)						
Lead/Guidance					1 (+)	1 (+)		
Science/Chem.								+
Computers	2 (-)	2 (-)	-	-	3 (-)	3 (-)	-	-
Math								
Elect. Commun.								
Cler./Admin.	1 (-)	1 (-)						
Warehouse/Ship								
Food-Employee								
Food-Prof.								
Fire Prot.								
Law Enf.					1 (+)			
Mechanics								
Construction	1 (+)	2 (+)						
Electronics	2 (+)	2 (+)			1 (-)	1 (-)		
Vehicle Op.								
Unlikely Resp.							-	

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Table A-10
Content of Empirical Scales and Composites Developed to Predict Core Technical Proficiency[a] (CTP) in the CV and LV Cohorts: 91A Combined-Sex

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat	3 (+)	6 (+)	+					
Rugged Indiv.	2 (+)	8 (+)	+	+	3 (+)	3 (+)		
Firearms								
Drafting		1 (-)						
Audiographics		2 (-)	-	-	1 (-)	1 (-)		
Aesthetics								
Medical Serv.	4 (+)	8 (+)	+	+	1 (+)	1 (+)		
Lead/Guidance								
Science/Chem.	1 (+)	3 (+)	+		1 (+)	2 (+)		+
Computers	1 (-)	2 (-)						
Math								
Elect. Commun.	1 (-)	1 (-)			1 (-)	1 (-)		
Cler./Admin.		6 (-)	-		1 (-)	1 (-)		-
Warehouse/Ship								
Food-Employee								
Food-Prof.								
Fire Prot.					2 (+)	2 (+)	+	+
Law Enf.					1 (+)	1 (+)		
Mechanics					1 (-)	1 (-)		
Construction		1 (+)						
Electronics								
Vehicle Op.								
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher CTP score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher CTP score.

Appendix B

Content of Empirical Scales and Composites Developed to Predict Leadership

Table B-1
Content of Empirical Scales and Composites Developed to Predict Leadership[a] in the LVII Cohorts: All MOS

	LVII Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat				
Rugged Indiv.	3 (+)	7 (+)	+	
Firearms				
Drafting				
Audiographics				
Aesthetics		2 (-)		
Medical Serv.	1 (-)	2 (-)		
Lead/Guidance		3 (+)		
Science/Chem.				
Computers				
Math				
Elect. Commun.				
Cler./Admin.	2 (-)	6 (-)	-	
Warehouse/Ship	1 (-)	3 (-)	-	
Food-Employee	3 (-)	5 (-)	-	-
Food-Prof.	1 (-)	3 (-)	-	
Fire Prot.				
Law Enf.				
Mechanics				
Construction	2 (-)	1 (+)		
Electronics				
Vehicle Op.	1 (-)	3 (-)	-	
Unlikely Resp.			-	-

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to exhibit a higher Leadership score. A negative weight indicates that soldiers who expressed lack of interest or dislike tended to exhibit a higher Leadership score.

Appendix C

Content of Empirical Scales and Composites Developed to Predict Attrition

Table C-1
Content of Empirical Scales and Composites Developed to Predict Attrition[a] in the LV Cohort: 13B and 91A Males

	CV Cohort				LV Cohort			
AVOICE Scale	12-Item Scale	All-Significant Scale	Hand-Picked Composite[b]	Regression-Picked Composite[b]	12-Item Scale	All-Significant Scale	Hand-Picked Composite	Regression-Picked Composite
Combat								
Rugged Indiv.	3 (-)	3 (-)			3 (+)	6 (+)	+	+
Firearms					1 (+)	2 (+)		
Drafting								
Audiographics		1 (-)			1 (+)	3 (+)		
Aesthetics	1 (+)	1 (-)			1 (+)	1 (+)		
Medical Serv.	1 (-)	1 (+)				1 (+)		
Lead/Guidance	1 (-)	1 (+)			1 (+)	2 (+)		
Science/Chem.	1 (+)	1 (-)			1 (+)	5 (+)	+	
Computers						2 (+)		
Math								
Elect. Commun.								
Cler./Admin.					1 (-)	2 (-)		
Warehouse/Ship								
Food-Employee						1 (-)		
Food-Prof.								
Fire Prot.					1 (+)	2 (+)		
Law Enf.	1 (+)	1 (-)						
Mechanics	1 (+)	1 (-)				1 (+)		
Construction	2 (+)	2 (-)			1 (+)	1 (+)		
Electronics					1 (+)	2 (+)	+	
Vehicle Op.	1 (+)	1 (-)				1 (-)		
Unlikely Resp.								

[a] A positive weight indicates that soldiers who expressed stronger interest or liking tended to stay in the Army. A negative weight indicates that soldiers who expressed stronger interest or liking tended to attrite.
[b] No composite could be formed.

Appendix D

Cross-Sex Transportability for Core Technical Proficiency

Table D-1
Cross-Sex, Within-Cohort Transportability of Single-Sex Scales/Composites Developed to Predict Core Technical Proficiency

	Correlation Between CTP and Single-Sex Scales/Composites			
	Applied in Same-Sex Samples		Transported to Opposite-Sex Samples	
	CV Scales/ Composites in CV Cohort (3 samples)	LV Scales/ Composites in LV Cohort (2 samples)	CV Scales/ Composites in CV Cohort (4 samples)	LV Scales/ Composites in LV Cohort (3-4 samples)
Scales				
12-Item Scale				
Median	.15	.20	.26	.16
Range	.14-.22	.17-.23	.18-.28	.12-.37
All-Significant Scale				
Median	.18	.19	.25	.15
Range	.16-.20	.17-.20	.15-.27	.11-.26
Composites				
Hand-Picked	.24	.13	.24	.09
Median	.22-.25	.10-.16	-.03-.30	-.06-.19
Range				
Regression-Picked				
Median	.24	.08	.25	.09
Range	.07-.25	.01-.16	-.02-.30	.04-.19

Table D-2
Cross-Sex Transportability for Combined-Sex Scales/Composites Developed to Predict Core Technical Proficiency

	Correlation Between CTP and Combined-Sex Keys Transported to Combined-Sex Samples			
	Same Cohort		Opposite Cohort	
	CV Scales/ Composites in CV Cohort (3 samples)	LV Scales/ Composites in LV Cohort (2-3 samples)	CV Scales/ Composites in LV Cohort (3 samples)	LV Scales/ Composites in CV Cohort (2-3 samples)
Scales				
12-Item Scale				
Median	.27	.18	.14	.25
Range	.16-.47	.10-.38	.11-.36	.05-.49
All-Significant Scale				
Median	.26	.18	.17	.22
Range	.22-.51	.17-.36	.16-.37	.12-.49
Composites				
Hand-Picked				
Median	.19	.04	.14	.09
Range	.19-.47	.03-.35	.08-.37	.08-.47
Regression-Picked				
Median	.21	.20	.15	.28
Range	.21-.47	.10-.30	-.05-.35	.11-.45

Appendix E

Content of Occupational Scales Developed to Predict MOS Membership

Table E-1
Content of 12-Item Occupational Scales Developed in the CV Sample

AVOICE Scale	11B Male	13B Male	63B Male	63B Combined-Sex	71L Male (Full)	71L Female	71L Combined-Sex	91A Male	91A Female (Full)	91A Combined-Sex
Combat	2 (+)	2 (+)							1 (-)	
Rugged Indiv.	4 (+)		1 (+)	1 (+)			1 (-)		1 (+)	
Firearms	2 (+)				3 (-)	1 (-)	3 (-)			
Drafting										
Audiographics						1 (-)				
Aesthetics										
Medical Serv.						1 (-)		12 (+)	8 (+)	12 (+)
Lead/Guidance										
Science/Chem.		1 (+)								
Computers		1 (+)			1 (+)					
Math										
Elect. Commun.									1 (-)	
Cler./Admin.		4 (+)			5 (+)	1 (+)	2 (+)		1 (+)	
Warehouse/Ship		1 (+)				1 (-)				
Food-Employee										
Food-Prof.		2 (+)								
Fire Prot.										
Law Enf.	4 (+)	1 (+)				1 (-)				
Mechanics			9 (+)	8 (+)		1 (-)	1 (-)			
Construction			1 (+)	2 (+)	2 (-)	3 (-)	5 (-)			
Electronics			1 (+)	1 (+)		2 (-)				
Vehicle Op.					1 (-)					
Unlikely Resp.										

Table E-2
Content of All-Significant Occupational Scales Developed in the CV Sample

AVOICE Scale	11B Male	13B Male	63B Male	63B Combined-Sex	71L Male (Full)	71L Female	71L Combined-Sex	91A Male	91A Female (Full)	91A Combined-Sex
Combat	8 (+)	3 (+)	3 (+)	5 (+)	8 (-)	2 (-)	10 (-)	5 (-)	1 (-)	2 (-)
Rugged Indiv.	11(10+,1)		11 (+)	11 (+)	12 (-)	4 (-)	12 (-)	2 (+)	3 (-)	2 (-)
Firearms	6 (+)		5 (+)	5 (+)	6 (-)	5 (-)	7 (-)	1 (-)		6 (-)
Drafting			3 (-)	3 (-)			3 (-)			
Audiographics		3 (+)	4 (-)	4 (-)		1 (-)	1 (-)			
Aesthetics	1 (-)		6 (-)	5 (-)		1 (+)	4 (+)	5 (+)	2 (+)	5 (+)
Medical Serv.	1 (-)		9 (-)	11 (-)	3 (+)	3 (-)	4 (+)	12 (+)	11 (+)	12 (+)
Lead/Guidance	1 (+)	3 (+)	7 (-)	7 (-)	5 (+)		3 (+)	9 (+)		6 (+)
Science/Chem.	2 (+)	2 (+)	3 (-)	3 (-)	1 (+)	1 (-)	5 (-)	2 (+)	2 (+)	3 (2+,1-)
Computers		2 (+)	3 (-)	3 (-)	4 (+)		3 (+)			
Math		•			2 (+)		1 (-)			
Elect. Commun.			4 (-)	3 (-)	1 (+)	2 (-)	1 (-)		2 (-)	
Cler./Admin.	9 (-)	14 (+)	13 (-)	13 (-)	13 (+)	7 (+)	12 (+)	3 (2+,1-)	2 (-)	1 (+)
Warehouse/Ship		2 (+)			1 (-)	2 (-)	2 (-)	1 (-)		
Food-Employee	1 (-)				1 (-)				1 (-)	
Food-Prof.	8 (-)	2 (+)	4 (-)	4 (-)			1 (+)		3 (-)	
Fire Prot.		1 (+)	1 (-)		1 (-)	1 (-)	2 (-)	1 (+)		
Law Enf.	7 (+)	3 (+)	5 (-)	5 (-)	7 (-)	5 (-)	6 (-)	3 (-)		3 (-)
Mechanics			10 (+)	10 (+)	9 (-)	7 (-)	10 (-)	8 (-)		10 (-)
Construction	1 (+)	3 (+)	13 (+)	12 (+)	14 (-)	12 (-)	13 (-)	9 (-)		9 (-)
Electronics		1 (+)	8 (+)	8 (+)	5 (-)	11 (-)	12 (-)	5 (-)		8 (-)
Vehicle Op.			3 (+)	3 (+)	3 (-)	2 (-)	3 (-)	3 (-)		3 (-)
Unlikely Resp.										

Appendix F

Least-Squares Regression Weights for Predictor Composites in Chapter 8

Core Technical Proficiency Criterion

Least-Squares Regression Composites by MOS (ASVAB only):

MOS	Fold-Back	N	Standardized Regression Weights AR	AS	CS	EI	GS	MC	MK	NO	VE
11B	.751	235	.09	.15	.15	.20	.10	.20	.12	.09	-.15
13B	.429	551	.16	.07	.07	-.04	.09	.13	.11	-.04	-.06
19K	.467	445	-.04	.14	.07	.15	.02	.13	.15	-.01	-.05
31C	.633	172	.23	.37	-.15	-.05	.15	-.11	.08	.16	.03
63B	.607	406	.13	.35	.04	.01	-.04	.09	.06	.07	.04
71L	.820	251	.19	-.00	.16	.02	.04	-.02	.29	.00	.26
88M	.541	221	.17	.16	.07	.15	-.03	.22	-.00	.09	-.20
91A	.672	535	.01	.07	.20	.14	-.01	.12	.20	.02	.10
95B	.749	270	.12	.06	.05	.12	.06	.11	.01	-.01	.36

Least-Squares Regression Composites by MOS (ASVAB + cognitive):

MOS	Fold-Back	N	Standardized Regression Weights AR	AS	CS	EI	GS	MC	MK	NO	VE	SPATL	MTPLC	NMSPC	PCACC	PCSPD	PSYCM	STMEM	BASPD	BASAC
11B	.789	235	-.01	.10	.12	.19	.09	.09	.07	.05	-.14	.25	.14	.14	-.04	-.02	.02	-.04	-.12	.03
13B	.453	551	.13	.05	.04	-.04	.10	.05	.09	-.03	-.04	.19	.05	-.04	-.00	-.09	.04	.03	.00	-.02
19K	.484	445	-.06	.13	.04	.16	.02	.08	.12	.01	-.05	.21	-.01	-.04	-.04	-.02	-.07	.03	.01	-.03
31C	.678	172	.14	.38	-.17	-.03	.18	-.11	.01	.13	-.02	.24	-.08	.11	-.07	-.12	-.16	.08	.13	-.04
63B	.642	406	.09	.32	-.03	.02	-.05	-.03	.01	.06	.06	.26	.06	-.03	.01	-.02	.01	.10	.00	.03
71L	.830	251	.19	-.01	.14	.02	.05	-.06	.28	.03	.26	.10	.08	-.06	.03	-.05	-.03	.03	-.02	.05
88M	.572	221	.14	.15	.02	.15	-.04	.14	-.03	.10	-.20	.12	.02	-.02	.11	.00	.05	.09	-.07	.02
91A	.694	535	-.03	.06	.15	.14	-.01	.04	.16	.03	.09	.24	-.05	-.04	.01	-.02	-.03	.11	-.01	-.00
95B	.787	270	.05	.04	-.02	.13	.05	-.01	-.06	-.04	.36	.27	.07	.03	-.01	-.01	-.04	.13	.02	.05

Least-Squares Regression Composites by MOS (ASVAB + non-cognitive):

MOS	Fold-Back	N	Standardized Regression Weights AR	AS	CS	EI	GS	MC	MK	NO	VE	TCACH	TCADJ	TCOND	TCONT	TCOOP	TCDPN	TCLED	TCADM	TCAUD	TCFSR	TCMCH	TCPSR	TCRGD	TCSOC	TCTCH
11B	.766	235	.09	.15	.15	.19	.08	.20	.11	.08	-.14	-.04	-.01	.03	.06	.01	.07	-.11	-.14	-.04	.02	.01	.02	.01	.06	.09
13B	.476	551	.17	.04	.06	-.04	.10	.12	.12	-.03	-.06	-.05	.06	-.01	.05	-.02	.11	-.09	-.02	-.03	-.12	.12	.07	-.06	.13	-.03
19K	.514	445	-.05	.08	.07	.14	.04	.15	.16	-.00	-.00	.14	-.07	-.06	.00	-.01	.07	-.06	.06	-.14	.09	.06	-.13	.13	.01	-.03
31C	.700	172	.20	.32	-.18	-.08	.20	-.10	.09	.16	.06	.08	.03	-.15	-.12	.01	.10	.08	.20	-.12	-.22	.01	-.04	.11	-.04	.06
63B	.634	406	.13	.30	.04	-.02	-.03	.08	.06	.06	.09	.01	.07	-.08	-.05	-.06	.09	.05	-.12	-.09	-.03	.15	.00	-.03	.08	.05
71L	.844	251	.19	.05	.16	.02	.04	.00	.27	.00	.25	-.01	-.04	.01	.02	-.02	-.00	.02	.02	.10	.13	-.05	-.12	.03	-.01	.03
88M	.609	221	.15	.05	.07	.11	.02	.21	.01	.11	-.10	.02	-.01	-.10	-.05	.13	.05	.01	.07	-.12	-.17	.19	-.01	.04	-.12	.07
91A	.689	535	.02	.07	.19	.15	-.02	.13	.22	.02	.08	.13	-.01	-.04	-.08	-.05	.08	-.06	.01	-.01	.07	-.03	.04	.05	.06	-.10
95B	.769	270	.11	.08	.05	.14	.04	.11	.02	-.01	.31	.11	-.16	.00	.04	.01	-.11	-.02	.03	.03	-.02	-.13	.07	.07	.03	-.01

Overall Performance Criterion

Least-Squares Regression Composites by MOS (ASVAB only):

MOS	Fold-Back	N	Standardized Regression Weights AR	AS	CS	EI	GS	MC	MK	NO	VE
11B	.638	235	.03	.20	.13	.16	.06	.06	.29	.01	-.15
13B	.491	551	.12	.19	.06	-.00	.02	.15	.11	.01	-.05
19K	.562	445	.07	.18	.04	-.04	-.05	.21	.23	.02	.01
31C	.591	172	.10	.17	.00	-.02	-.11	.17	.18	.07	.15
63B	.656	406	.20	.24	.11	-.02	.04	.22	.03	-.09	-.10
71L	.717	251	.10	.05	.07	.03	.03	.05	.33	.04	.15
88M	.593	221	.14	.15	.12	.05	-.05	.24	.17	.05	-.15
91A	.653	535	-.02	.13	.17	.15	-.07	.18	.26	.05	-.02
95B	.781	270	.06	.22	.13	.07	.08	.08	.15	-.00	.19

Least-Squares Regression Composites by MOS (ASVAB + cognitive):

MOS	Fold-Back	N	Standardized Regression Weights AR	AS	CS	EI	GS	MC	MK	NO	VE	SPATL	MTPLC	NMSPC	PCACC	PCSPD	PSYCM	STMEM	BASPD	BASAC
11B	.677	235	-.00	.15	.11	.15	.06	-.05	.28	.02	-.11	.19	.13	-.00	-.09	-.09	.10	-.01	-.12	.03
13B	.514	551	.09	.18	.02	.00	.02	.07	.07	.01	-.03	.17	.08	-.02	-.01	-.05	-.01	.07	.04	-.02
19K	.595	445	.03	.17	-.01	-.03	-.05	.12	.17	.03	.00	.31	.02	-.01	.02	.03	-.12	.00	.00	-.06
31C	.630	172	.06	.16	-.05	-.01	-.11	.06	.14	.08	.16	.30	-.07	-.07	-.02	-.08	.02	.03	.13	.02
63B	.676	406	.19	.24	.06	-.02	.04	.15	.01	.09	-.08	.10	.03	-.06	.02	-.04	.04	.14	.03	.00
71L	.738	251	.09	.03	.04	.03	.03	-.02	.31	.04	.17	.07	.09	-.01	-.03	-.03	.05	.10	-.07	.08
88M	.640	221	.15	.14	.06	.04	-.06	.14	.14	.08	-.13	.17	.01	-.12	.02	.05	.01	.20	-.10	-.04
91A	.675	535	-.01	.14	.12	.15	-.07	.15	.25	.07	-.03	.07	.02	-.11	.08	-.01	-.05	.10	.05	.06
95B	.812	270	.04	.21	.07	.07	.06	-.03	.10	-.04	.20	.11	.08	.02	.04	.09	.03	.12	.00	.00

Least-Squares Regression Composites by MOS (ASVAB + non-cognitive):

MOS	Fold-Back	N	Standardized Regression Weights AR	AS	CS	EI	GS	MC	MK	NO	VE	TCACH	TCADJ	TCOND	TCONT	TCOOP	TCOPN	TCLED	TCADM	TCAUD	TCFSR	TCMCH	TCPSR	TCRGD	TCSOC	TCTCH
11B	.670	235	.05	.22	.11	.15	.05	.05	.27	.01	-.16	-.01	-.06	.07	.02	-.01	.16	-.06	-.14	.02	-.01	-.00	.00	-.01	.09	.08
13B	.529	551	.13	.17	.05	.00	.02	.16	.11	.02	-.04	.02	-.05	-.01	-.03	.01	.19	-.07	-.02	-.09	-.01	.04	.07	-.03	.07	-.05
19K	.598	445	.07	.20	.02	-.04	-.05	.22	.21	.02	.00	.07	-.03	.01	-.01	-.01	.13	-.07	-.01	-.08	.06	-.02	-.13	.08	.14	.03
31C	.690	172	.11	.12	-.04	-.01	-.08	.16	.21	.08	.11	.09	-.06	-.07	-.03	.07	.15	-.00	.15	-.14	-.29	.05	.02	.03	.13	-.15
63B	.684	406	.19	.26	.09	-.04	.04	.21	.02	.08	-.10	.09	.08	.02	-.08	.02	.06	-.03	-.07	-.04	-.03	-.03	.02	-.02	.06	.06
71L	.751	251	.10	.09	.06	.03	.03	.06	.30	.03	.16	.11	-.11	.06	-.09	.02	.02	-.00	.08	.09	.07	-.13	-.07	.11	-.08	.09
88M	.681	221	.16	.09	.12	.02	.02	.25	.15	.07	-.02	-.19	-.01	.08	-.12	.12	.28	.09	.12	-.07	-.14	.24	-.01	-.12	-.11	-.03
91A	.688	535	-.01	.10	.15	.16	-.07	.17	.29	.05	-.03	.15	-.03	.03	-.06	-.01	.12	-.11	.04	.00	-.04	-.00	.08	.05	.05	-.11
95B	.803	270	.05	.19	.13	.07	.08	.07	.15	-.01	.19	.07	-.17	.04	.14	-.04	-.01	.01	-.03	-.05	-.01	-.01	.02	.06	-.01	.05